UNIVERSITÉ DU QUÉBEC À MONTRÉAL

METHODS FOR THE FUNCTIONAL ANALYSIS OF SMALL RNAS

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

FOR THE DOCTORATE IN COMPUTER SCIENCE

BY

CHAO-JUNG WU

MARCH 2026

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES D'ANALYSE FONCTIONNELLE DES PETITS ARNS

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE

PAR

CHAO-JUNG WU

MARS 2026

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

ACKNOWLEDGMENTS

I am grateful to Professor Abdoulaye Baniré Diallo, my thesis supervisor, who taught me to approach science with clarity, to define problems precisely, to prioritize wisely, and to practice the art of delegation by entrusting tasks and sharing knowledge. That balance is something I will carry with me.

I also wish to thank Professor Xiang-Jiao Yang at McGill, my former supervisor, and Professor Alexandre Blondin Massé, whose course I found especially engaging. Their recommendation letters helped me secure important fellowships that supported my graduate studies.

I am especially grateful to Dr. Alix Boc and Professor Nadia Tahiri, whose teaching and guidance helped me quickly gain confidence in programming and use it effectively in my work.

I further thank Dr. Zahra Agharbaoui, whose expertise in plant science helped me recognize how biological knowledge can inform computational thinking. Her insights enriched the way I approached analysis and interpretation.

I am also deeply grateful to Professor Jean Danyluk, who introduced me to a new research direction and gave me the opportunity to observe how biologists and bioinformaticians exchange ideas and collaborate in practice.

I want to express my warmest thanks to Dr. Amine M. Remita, who invited me to join a team project early in my studies. His mentorship taught me to explore established tools and apply them to advance my work, to work toward shared goals, and to grow into my role as a researcher.

I am also thankful to Farzaneh Rahmdani, whose dedication and steady support helped me navigate this project through its many challenges. Her positive spirit and collaborative mindset made a real difference throughout our work together.

I am sincerely thankful to Dr. Hui-Wen Liu, a former colleague from Academia Sinica, who helped me identify a practical direction for my research and facilitated the opportunity to bring it to life through an internship at a pharmaceutical company.

I thank Drs. Wei Zhang, Baohong Zhang, Thomas Carlile, Xiaohui Zhao, Yizhou Zhou, and Cullen Mason for their support in making my internship possible. In a setting where staff transitions happen frequently and such opportunities can easily be deprioritized, their steady involvement ensured that the internship continued meaningfully.

I would also like to thank the members of my thesis jury, Professors Roman Sarrazin-Gendron, Nadia Tahiri, and Petko Valtchev, for their time, insightful comments, and rigorous evaluation of my work.

I gratefully acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies.

I want to thank Dr. Hayda Almeida and Armand Bandiang Massoua, whose thoughtful feedback and guidance supported my learning process and helped refine my ideas. I am also grateful to Golrokh Kiani, whose calm presence and sense of direction helped me stay focused during times of uncertainty.

I would also like to thank Dr. Dylan Lebatteux and Vahid Naghashi, my cohortmates throughout this Ph.D. journey. We took courses together, encouraged each other during difficult times, and shared many long discussions as we worked through the challenges of graduate school. I truly value the time we spent learning together, supporting each other, and growing side by side.

I am grateful to the many current and former members of the Diallo Lab whose support enriched my time in the laboratory. I especially thank Dr. Ahmed Halioui and Karen Reynard for generously answering my questions and offering helpful advice.

My heartfelt thanks go to my husband, whose steady support carried me through the most demanding periods of this work. I am grateful to my three children, who gave me the motivation to persevere and the joy that made everything worthwhile. I thank the Cheng Family Scholarship, awarded through my uncle, and hope to live up to the values of those who came before me. I am also thankful to the many other family members and friends for their kindness, patience, and continued support throughout this long endeavor.

Finally, I give thanks to Má-tsóo, Thôo-tē-kong, Póo-seng Tāi-tè, Koan-sè-im Phô-sat, Confucius, and Bûn-chhiong Tè-kun, who have accompanied me throughout my life. I received scholarships from Nanyao Temple, and I am grateful to Má-tsóo for her continued protection, even if I often only realized it in hindsight.

CONTENTS

LIST OF FIGURES xiv

LIST OF TABLES xvii

ACRONYMES xix

GLOSSARIES xxii

RÉSUMÉ xxv

ABSTRACT xxvi

INTRODUCTION 1

0.1 Real-World Applications of Small RNAs in Agriculture and Medicine 1

0.2 Computational Challenges for RNA Analysis 2

0.3 Thesis Overview 4

0.4 Contributions 6

CHAPTER 1 BACKGROUND 7

1.1 RNA Biomarkers and Therapeutic Applications 7

1.2 Definitions of Small RNA and Other Biological Concepts 8

1.2.1 Small RNA 8

1.2.2 NGS Techniques 9

1.2.3 Nucleotides and the Flow of Genetic Information 10

1.2.4 Base Pairing and Molecular Interaction 11

1.2.5 Textual Representation of RNA Sequences 12

1.2.6 In Silico, In Vivo, and In Vitro 13

1.3 RNA Feature Representation 13

1.4 Modeling Features for RNA Classification Tasks 16

1.5 Bioinformatics Concepts and Preprocessing 16
1.5.1 Benchmark Datasets 16
1.5.2 Dataset Definition 18
1.5.3 Feature Transformation 18
1.5.4 Distance Functions 19
1.6 Statistical Modeling 20
1.6.1 Statistical Distributions 20
1.6.2 Differential Expression Analysis 21
1.6.3 Multiple Testing 22
1.6.4 Hypergeometric Enrichment Analysis 22
1.7 Machine Learning 23
1.7.1 Supervised Learning 24
1.7.2 Classification Evaluation Metrics 26
1.7.3 Selecting Supervised Models for Bioinformatics 27
1.7.4 Unsupervised Learning 28
1.7.5 Selecting Unsupervised Methods for Bioinformatics 29
1.8 Roadmap to Small RNA Biomarker Discovery 29
1.8.1 Step 1: Sample Preparation 31
1.8.2 Step 2: Quantitative Data Acquisition 32
1.8.3 Step 3: Classification and Candidate Identification 33
1.8.4 Step 4: Candidate Biomarker Selection (Differential Profiling) 34
1.8.5 Step 5: Functional Assessment of Small RNA Biomarkers 35
1.8.6 Step 6: Experimental Validation 35

1.8.7 Step 7: Therapeutic and Diagnostic Feasibility 36
1.8.8 Representative Study on miRNA and siRNA Biomarker Discovery 37
1.9 Milestones in Plant miRNA and siRNA Identification and Prediction 39
1.10 Evolving Criteria for Plant miRNA Prediction in the Big Data Era 41
1.11 Computational Strategies for Plant miRNA Identification 44
1.11.1 miRDeep-P 45
1.11.2 ShortStack 46
1.11.3 miR-PREFeR 47
1.11.4 miRdup 48
1.12 Current Requirements for Bioinformatics Tools 49
1.12.1 Scalable Computing 50
1.12.2 Containerization and Software Packaging 52
1.12.3 Workflow Automation 53
1.12.4 Open-Source Licensing 54
CHAPTER 2 RESEARCH PROBLEMS 55
2.1 Motivation 55
2.1.1 Complexity of sRNA-seq Composition 55
2.1.2 Big Data Are Getting Bigger 60
2.1.3 Disentangling the RNA Network: Studying miRNAs and siRNAs Side by Side 61
2.1.4 Missing Piece in Small RNA Bioinformatics 63
2.1.5 Optimizing sRNA Delivery: Transferable Lessons from rAAV to Plant Viral Vectors 64
2.2 Hypothesis 65
2.3 Objectives of This Thesis 66

CHAPTER 3 LITERATURE REVIEW 69

3.1 Computational Complexity of miRDeep-P as a Representative Pipeline for miRNA Prediction 69

3.1.1 Algorithmic Complexity Breakdown 70

3.1.2 Worst-Case Complexity Analysis 71

3.1.3 Memory Complexity Analysis 71

3.1.4 Optimization Strategies 71

3.2 Applying MapReduce to Accelerate Data Analysis 75

3.3 Optimizing with Apache Spark 77

3.4 Joint miRNA-siRNA Analysis in Wheat 80

3.4.1 Functional and Evolutionary Overlap Between miRNAs and siRNAs 80

3.4.2 Wheat as a Model for iRNA Evolution and Stress Responses 82

3.4.3 Limitations of Existing Joint Small RNA Analyses 84

3.5 siRNA Biogenesis Hypothesis 86

3.5.1 Dicer Processing for RNAi Activation 86

3.5.2 Molecular Origins of Dicer Substrates 88

3.6 Assessing Phasing Potential 94

3.6.1 Example Calculation of Phase Scores 95

3.6.2 Critical Evaluation for Howell, Chen, and Guo Scores 96

3.7 Computational Strategy for Gene Therapy Viral Vector Integrity 101

3.7.1 AAV-GPseq: High-Resolution Profiling of AAV-Genome Population 102

3.8 Validation Strategies in Computational Small RNA Research 103

CHAPTER 4 MIRLIBSPARK: A SCALABLE NGS PLANT MICRORNA PREDICTION PIPELINE FOR MULTI-LIBRARY FUNCTIONAL ANNOTATION 106

4.1 Abstract 107

4.2 Introduction 107

4.3 MirLibSpark architecture and execution 108

4.3.1 MirLibSpark overview 108

4.3.2 MiRNA prediction module (M2) 110

4.3.3 Additional annotation modules 111

4.3.4 Execution 112

4.4 Preparations for benchmarking 113

4.4.1 Construction of simulated data sets 113

4.4.2 sRNA-seq libraries 114

4.4.3 Implementation of a sequential version 114

4.4.4 Implementation of a miRDeep-P wrapper 114

4.4.5 miRNA clustering 114

4.5 Benchmarking and discussions 115

4.5.1 Improved execution time 115

4.5.2 Comparison of performance with other predictors 115

4.5.3 Scalability 117

4.5.4 Case study 118

4.6 Conclusion 119

4.7 Implications 120

CHAPTER 5 PREDICTION AND EVOLUTIONARY DYNAMICS OF WHEAT MIRNAS AND SIRNAS DURING FIELD COLD ACCLIMATION 122

5.1 Abstract 123

5.2 Introduction 124
5.2.1 Justification for Simultaneous Prediction of miRNAs and siRNAs 124
5.2.2 Categorizing Interfering RNAs 125
5.2.3 Current Studies on miRNAs and siRNAs in Plants 125
5.2.4 Wheat as a Model for iRNA Evolution and Stress Responses 126
5.2.5 Computational Challenges in Small RNA Prediction in Wheat 127
5.2.6 Major Contributions 128
5.3 Materials and methods 129
5.3.1 Small RNA preparation from wheat field samples 129
5.3.2 Bioinformatics approach for the identification of miRNAs and siRNAs in wheat cold response 130
5.3.3 Identification of miRNA and siRNA candidates from sequenced small RNAs 130
5.3.4 Sequence analyses for precursor and mature iRNAs 133
5.3.5 Expression statistical analyses of iRNAs 133
5.3.6 Assigning regulated iRNAs in genetic backgrounds 134
5.3.7 Differential regulation analysis based on iRNA expression patterns during field cold acclimation 134
5.3.8 Functional annotations of XLOC database 135
5.3.9 Functional enrichments of iRNAs through target genes 136
5.4 Results 136
5.4.1 Integrated miRNA-siRNA Prediction 136
5.4.2 Discovery of Dual-Identity Small RNAs 138
5.4.3 Novel iRNA Discovery 138
5.4.4 Identification of Wheat siRNA-Producing Genes 139

5.4.5 Evolutionary Landscape of Wheat iRNAs 141
5.4.6 Identification of cold response iRNAs 142
5.4.7 Functions of Cold-Response iRNAs in Acclimation 143
5.4.8 iRNAs in *Vrn-A1*-Mediated Vernalization 146
5.5 Discussion 147
5.6 Figures and Tables 150
5.7 Implications 166
CHAPTER 6 LEVERAGING MACHINE LEARNING FOR IMPROVED SIRNA PREDICTION IN ARABIDOPSIS 168
6.1 Abstract 169
6.2 Introduction 169
6.3 Materials and methods 174
6.3.1 Datasets 174
6.3.2 Algorithms for determining phasing boundaries 177
6.3.3 Feature vectors and training 180
6.3.4 siRNA indication score 182
6.4 Results 184
6.4.1 Evaluation of individual predictive features 185
6.4.2 Training and evaluation of siRNA classifiers 188
6.4.3 Effector siRNA localization in precursor 194
6.4.4 The siWalk program 195
6.5 Discussion 197
6.6 Implications 198

CHAPTER 7 ARCTICFOX: A QUANTITATIVE PROFILING FRAMEWORK FOR RAAV GENOME POPULATION 200

7.1 Abstract 201

7.2 Introduction 202

7.3 Materials and Methods 204

7.3.1 ArcticFox overview 204

7.3.2 Datasets 205

7.3.3 Design of reference genomes 205

7.3.4 Correcting size biases of read abundance 207

7.3.5 Functional quantification 208

7.4 Results 208

7.4.1 Size bias related to sequencing 208

7.4.2 Multiple manufacturing batches 210

7.5 Conclusion 211

7.6 Implications 212

CONCLUSION 213

APPENDIX A MIRLIBSPARK 217

A.1 Definitive version 218

A.2 MirLibSpark manual 224

A.3 Example html output 230

APPENDIX B WHEAT PROJECT 231

B.1 Manuscript 232

B.2 Supplementary figures 277

B.3 Supplementary tables (legends only) 283
B.4 S17: TAS contigs 292
B.5 S18: precursor index 301
B.6 S19: iRNA annotations of 339 iRNAs 322
APPENDIX C SIWALK 349
C.1 Manuscript 350
C.2 SiWalk manual 372
APPENDIX D ARCTICFOX 378
D.1 RECOMB abstract 379
D.2 ISMB abstract and poster 380
BIBLIOGRAPHY 386

LIST OF FIGURES

Figure 1.1 Classification of small RNAs collected from NGS sRNA-seq 9

Figure 1.2 DNA, nucleotides, and base pairing 11

Figure 1.3 RNA secondary structure 13

Figure 1.4 Functional annotation of sRNA 15

Figure 1.5 Small RNA biomarker discovery involves several steps 30

Figure 2.1 Size distribution of small RNAs in plants 56

Figure 2.2 RNA interference pathways 58

Figure 2.3 Nomenclature of siRNA classes 59

Figure 2.4 Small RNA alignments on representative miRNA and siRNA precursor models 59

Figure 3.1 Execution time of miRNA prediction tools 73

Figure 3.2 MapReduce-based word count for DNA sequences 77

Figure 3.3 Model of TE-derived evolution of miRNA and siRNA precursors 82

Figure 3.4 Hybridization and divergence events leading to bread wheat 83

Figure 3.5 Dicer-mediated sequential cleavage of dsRNA 87

Figure 3.6 Dicer processing is essential for RNAi activation 88

Figure 3.7 Distinct endogenous sources of dsRNA substrates for Dicer processing 91

Figure 3.8 Small RNA alignment patterns across siRNA precursor and non-precursor models 92

Figure 3.9 RDR2- and RDR6-mediated siRNA biogenesis 93

Figure 3.10 Illustration of an example phased siRNA locus 95

Figure 3.11 Score distributions of phase scoring functions 97

Figure 3.12 Alignments of heterogeneous populations to the rAAV vector reference 102

Figure 4.1 MirLibSpark overview 110

Figure 4.2 MirLibSpark miRNA prediction module. 112

Figure 4.3 Comparative performance of miRNA annotation software from plants..... 117

Figure 4.4 Venn diagram characterizing the predictions of GSE44622 by Jeong et al and mirLibSpark 119

Figure 5.1 Bioinformatics analysis workflow of wheat miRNA and siRNA prediction and characterization 150

Figure 5.2 Classification of predicted iRNAs..... 151

Figure 5.3 Chromosomal distribution of miRNAs and siRNAs in the wheat genome..... 152

Figure 5.4 Cross-database comparison of predicted miRNAs and siRNAs 153

Figure 5.5 Temporal differential expression of miRNAs and siRNAs across four wheat cultivars in 2010 and 2013 154

Figure 5.6 Winter-spring paired temporal expression profiles of cold response iRNAs 155

Figure 5.7 Functional enrichment of target genes regulated by cold-response iRNAs 157

Figure 5.8 Gene expression patterns of cold response mature sequences and their kinase targets ... 158

Figure 5.9 Gene expression patterns of cold response mature sequences and their transcription factor targets 159

Figure 6.1 Phasing patterns of small RNAs aligned to genomic sequences that produce dsRNA precursors 171

Figure 6.2 Illustration of algorithms for identifying phasing event boundaries 178

Figure 6.3 Attributes ranked by importance and their correlation with precursor prediction. 185

Figure 6.4 Relationship between the number of features and classifier performance..... 187

Figure 6.5 Classifier accuracy compared with random guessing 191

Figure 6.6 ROC curves of classifiers predicting from sRNA-seq libraries 191

Figure 6.7 Comparison of Precision and MCC among classifiers 192

Figure 6.8 Comparison of contigs predicted by different methods 194

Figure 6.9 Effector siRNA localization in an example precursor 195

Figure 6.10 Offset distribution of the 1st effector candidates 196

Figure 6.11 Workflow of the siWalk algorithm 197

Figure 7.1 rAAV profiling by long-read sequencing 203

Figure 7.2 LRS pipeline for assessing functional impurities in rAAV 205

Figure 7.3 Design of reference genomes 207

Figure 7.4 Size bias related to sequencing 209

Figure 7.5 Correcting size biases of read abundance 210

Figure 7.6 Comparing multiple datasets with ArcticFox LRS profiling 211

LIST OF TABLES

Table 1.1 Confusion matrix for binary classification 26

Table 1.2 miRBase criteria for high-confidence miRNA annotations 43

Table 1.3 Criteria for plant miRNA annotations 44

Table 2.1 Summary of research gaps and corresponding thesis contributions 68

Table 3.1 Comparison of Howell, Chen, and Guo phasing scores 101

Table 3.2 Validation techniques used across computational components in this thesis. 105

Table 4.1 Performance of plant miRNA annotation software using (A) simulation data (B) real world data 116

Table 4.2 Scalability of mirLibSpark for sRNA-seq data from small and large genomes 118

Table 5.1 Overview of miRNA and siRNA prediction originating from (A) ancestral genomes, (B) coding regions, and (C) transposable elements 160

Table 5.2 Known cold-response miRNA families identified in this study 161

Table 5.3 High-confidence novel cold-response iRNAs 162

Table 5.4 The cold response mature sequences and their kinase targets 163

Table 5.5 The cold response mature sequences and their DNA-binding protein targets 164

Table 5.6 Homology-Based Identification and Classification of TAS Precursors. 165

Table 5.7 Comparison of *Vrn-A1*-Dependent iRNA Functions in Winter- and Spring-Habit Wheat. 165

Table 5.8 Cultivar-specific iRNA families 166

Table 6.1 Example of a reference contig 174

Table 6.2 Literature-supported *Arabidopsis* siRNA-generating contigs 175

Table 6.3 Training dataset. 177

Table 6.4 Hyper-parameters of classifiers optimized for precision using Grid Search. 181

Table 6.5 SiWalk feature summary 182

Table 6.6 Top predictive features for precursor **pred**iction and effector **loca**lization. 186

Table 6.7 Time efficiency of various prediction methods 188

Table 6.8 Performance evaluation of various classifiers with imbalanced examples 189

Table 6.9 Performance between standard and stratified split strategies 193

Table 7.1 LRS datasets of rAAV productions from public and in-house sources 206

ACRONYMS

ArcticFox my rAAV genome population analyzer, see https://github.com/bioinfoUQAM/ArcticFox.

Ath Arabidopsis thaliana, a small flowering plant, model organism.

BLAST Basic Local Alignment Search Tool.

bp Base pair.

CR Cold Response.

Cstv Constitutive.

DE Differential Expression.

diploid having two sets of chromosomes.

DNA Deoxyribonucleic acid, the hereditary molecule that encodes genetic information in living organisms.

dw down.

F1 F-measure.

FDR False Discovery Rate.

FN False Negative.

FP False Positive.

GB Gradient Boosting.

GBA Gradient Boosting with AdaBoost.

GO Gene Ontology.

hexaploid having six sets of chromosomes.

iRNA interfering ribonucleic acid; interfering RNA.

IsoS wheat spring-habit Norstar-background cultivar.

IsoW wheat winter-habit Manitou-background cultivar.

IWGSC International Wheat Genome Sequencing Consortium.

KEGG Kyoto Encyclopedia of Genes and Genomes.

Manitou wheat spring-habit cultivar.

MCC Matthews Correlation Coefficient.

mirLibSpark my miRNA classifier, see https://github.com/bioinfoUQAM/mirLibSpark.

miRNA micro ribonucleic acid.

ML Machine Learning.

NCBI National Center for Biotechnology Information.

NGS Next Generation Sequencing technology.

Norstar wheat winter-habit cultivar.

nt nucleotide.

PHAS Phased siRNA-generating gene or locus.

PHASIS a siRNA predictor.

polyploid having more than two sets of chromosomes.

rAAV recombinant adeno-associated virus (rAAV), a drug delivery system to deliver small RNA-based treatments in humans.

RDD Resilient Distributed Dataset.

RFA Random Forest with AdaBoost.

RISC RNA-Induced Silencing Complex.

RNAi RNA interference or post-transcriptional gene silencing, a conserved biological process that regulates gene expression through small RNAs. RNAi refers to the biological mechanism itself, whereas interfering RNAs (iRNAs) are the RNA molecules that participate in and mediate this process.

ShortStack a siRNA predictor.

siRNA Small interfering ribonucleic acids, including siRNA subclasses, miRNA, piRNA, and mirtron.

siWalk my siRNA classifier, https://github.com/bioinfoUQAM/siWalk.

sRNA small RNA, short RNA molecules, including miRNAs, siRNAs, and other non-coding RNA fragments.

sRNA-seq small RNA sequencing technology, capturing < 50 nts, 10–500 million reads per sample.

SVM Support Vector Machine.

Tae Triticum aestivum, common wheat, hexaploid, 13 gigabytes or 16 gigabases genome.

TAS Trans-Acting siRNA gene.

tasiRNA Trans-acting siRNA.

TE Transposable Element.

TN True Negative.

TP True Positive.

WMP Wheat Micro RNA Portal , see http://www.bioinfo.uqam.ca/recherche/index.php?action=logiciels.

XLOC wheat mRNA transcripts collected from our field samples.

GLOSSARIES

Alignment The process of matching sequencing reads to a reference sequence.

Base pairing The hydrogen-bond interactions between complementary nucleotide bases within or between RNA strands.

Batch variability Differences in composition or quality observed between independently produced batches of viral vectors.

Bias Systematic distortion introduced during sequencing or data processing that affects sequence representation or abundance.

Biogenesis The biological process by which RNA molecules are generated and processed into their functional forms.

Capsid The protein shell of a virus that encloses and protects its genetic payload.

Chimera An artificial sequence formed by unintended joining of fragments from different genomic regions.

Chromosome A highly organized and compacted DNA molecule, packaged with proteins, that carries genes and regulatory elements as part of the genome.

Computational complexity A measure of the computational resources required by an algorithm as a function of input size.

Contamination The presence of unintended sequences originating from external biological sources or technical artifacts.

Data skew An uneven distribution of data that leads to imbalanced computational workloads and reduced performance.

Delivery The transfer of a vector and its payload into target cells or tissues.

Differential expression A change in RNA expression levels between biological conditions or treatments.

Distributed computing A computing approach in which tasks are executed across multiple interconnected machines to handle large data or workloads.

Expression profile The quantitative pattern of RNA expressions measured in a sample.

Feature engineering The process of designing, transforming, and selecting features to improve the performance of computational models.

Feature, attribute A measurable property or characteristic extracted from data and used as input for computational analysis or modeling.

Folding The process by which an RNA molecule adopts a specific secondary structure through base pairing.

Gene expression The process by which genetic information is transcribed into RNA and, when applicable, translated into functional products.

Genome The complete DNA sequence of an organism.

Heuristic A problem-solving strategy that applies practical rules or approximations to efficiently obtain acceptable solutions.

Host genome The genetic material of the producer cell used to generate viral vectors during manufacturing.

Impurity Unintended or unwanted molecular components present in a viral vector preparation.

Indexing The use of short sequence tags or computational identifiers to label, multiplex, or efficiently retrieve sequencing data.

In-memory processing A data-processing strategy that stores intermediate results in memory to reduce disk access and improve computational efficiency.

Integrity The extent to which a sequenced genome or molecule is complete, structurally correct, and free of defects.

Library A collection of sequencing reads obtained from a biological sample and prepared for high-throughput sequencing. A collection of sequencing reads obtained from a biological sample. In a transcriptomic library, sequencing depth refers to the number of reads observed for each RNA transcript.

Noise Random variation in sequencing data that does not reflect true biological signals.

Normalization A computational adjustment applied to sequencing data to enable quantitative comparison between samples.

Packaging The process by which genetic material or RNA is incorporated into a viral capsid during vector production.

Parallelization The execution of multiple computational tasks simultaneously to reduce overall processing time.

Payload The genetic or RNA material packaged within a vector and intended for delivery to target cells.

Phasing A regular pattern of small RNA production along a precursor, occurring at fixed intervals and resulting in periodicity.

Pipeline A structured sequence of computational steps designed to process data from input to output in an automated and reproducible manner.

Quality control Procedures used to assess the composition, consistency, and suitability of viral vectors for therapeutic use.

Reference genome A representative genome sequence used as a coordinate framework for read alignment and genomic analysis.

RNA interference effector A short RNA molecule that guides gene silencing by interacting with complementary target sequences.

RNA precursor A longer RNA molecule that is processed to produce a functional RNA product.

RNA secondary structure The folded configuration of an RNA molecule formed by internal base pairing.

Scalable Able to maintain or improve performance as data size or computational resources increase.

Target gene (of RNAi) A gene whose expression is regulated or silenced by an RNA interference effector.

Therapeutic integrity The extent to which a therapeutic vector retains its intended structure, composition, and functional properties.

Thermodynamic stability A measure of how energetically favorable and stable an RNA structure is under given conditions.

Transcriptome The complete set of RNA molecules expressed in a cell, tissue, or condition.

Truncation The presence of incomplete or shortened sequence fragments relative to the expected full-length sequence.

Vector A biological carrier used to deliver genetic material or RNA molecules into target cells.

Viral genome The nucleic acid sequence carried by a virus or viral vector.

Workflow An organized set of computational tasks and dependencies that define how data and analyses are executed and coordinated.

RÉSUMÉ

Les petits ARN (sARN), notamment les microARN (miARN) et les petits ARN interférents (siARN), sont des régulateurs essentiels de l'expression des gènes et des biomarqueurs moléculaires prometteurs. Chez les plantes, ils régulent le développement, la défense contre les pathogènes et l'adaptation au stress, tandis que leur reproductibilité en fait des candidats idéaux pour la découverte de biomarqueurs. Les progrès du séquençage de nouvelle génération (NGS) ont permis une détection approfondie des petits ARN. Cependant, l'extraction de signaux fonctionnellement significatifs à partir des données de séquençage des petits ARN (sARN-seq) demeure difficile en raison de l'hétérogénéité des profils d'expression, du chevauchement des voies de biogenèse et de la fragmentation des flux d'analyse. Relever ces défis nécessite une approche interdisciplinaire combinant biologie végétale, biochimie, informatique et apprentissage automatique. L'objectif principal de cette thèse est d'améliorer la découverte de biomarqueurs d'ARN en mobilisant des méthodes informatiques avancées. À cette fin, la recherche intègre la prédiction computationnelle, la modélisation statistique de la biogenèse de l'ARN, le développement de logiciels à grande échelle et des stratégies de contrôle qualité afin de soutenir la faisabilité des applications thérapeutiques. Cette thèse présente une suite de méthodologies bioinformatiques permettant d'identifier, de classer et d'interpréter les petits ARN dans des contextes biologiques et thérapeutiques.

Afin de permettre la découverte de biomarqueurs miARN reproductibles à grande échelle, la première contribution de cette thèse présente mirLibSpark, un pipeline distribué pour la prédiction et l'annotation de miARN végétaux utilisant Apache Spark. En intégrant des critères établis par la communauté, ainsi que la prédiction de cibles et l'enrichissement fonctionnel, mirLibSpark permet une analyse à grande échelle sur des données multi-librairies. Tout en offrant un support intégré pour des espèces importantes telles que l'*Arabidopsis*, le blé et le maïs, mirLibSpark est conçu pour s'adapter à tout génome végétal fourni par l'utilisateur. Afin de mettre en évidence les rôles coordonnés des miARN et siARN en conditions naturelles, la deuxième contribution étudie la dynamique des petits ARN lors de l'acclimatation au froid du blé. Cette étude applique un flux de travail unifié, parallélisé par Spark, à des librairies dérivées d'échantillons de terrain et identifie des ARN interférents (ARNi) sensibles au froid, connus et nouveaux, notamment des espèces à double identité présentant des caractéristiques communes aux miARN et siARN. Ces résultats offrent de nouvelles perspectives sur les origines évolutives et la diversification régulatrice des ARNi dans l'adaptation au stress. Pour améliorer la prédiction des précurseurs d'ARNi, la troisième contribution présente siWalk, un cadre d'apprentissage automatique entraîné sur *Arabidopsis*, combinant des caractéristiques de séquence, de structure et d'expression guidées par la biogenèse. Bien qu'un modèle de forêt aléatoire ait obtenu la meilleure précision, un modèle d'amplification de gradient est distribué comme alternative compacte et efficace. siWalk comprend aussi un module permettant de localiser les ARNi effecteurs à partir des séquences précurseurs, facilitant la conception rationnelle de vecteurs. Enfin, à titre d'exploration, ArcticFox a été développé pour évaluer l'intégrité des vecteurs d'ARN. Grâce aux lectures longues NGS, ArcticFox profile l'ADN résiduel et les troncatures génomiques dans les préparations de virus adéno-associés recombinants, et introduit des mesures quantitatives pour le contrôle qualité des thérapies à base d'ARN. Fondées sur l'informatique à grande échelle, la modélisation statistique et l'apprentissage automatique, ces contributions constituent une base fiable et interprétable pour faire progresser la bioinformatique des petits ARN dans la classification à haut débit, les études intégratives et les applications thérapeutiques.

Mots-clés: Bioinformatique des petits ARN, Apprentissage automatique en génomique, Annotation fonctionnelle des ARN, Analyse ARN à grande échelle, Découverte de biomarqueurs ARN, miARN, siARN

ABSTRACT

Small RNAs (sRNAs), including microRNAs (miRNAs) and small interfering RNAs (siRNAs), are essential regulators of gene expression and promising molecular biomarkers. In plants, they regulate development, pathogen defense, and stress adaptation, while their stability and reproducibility make them ideal candidates for biomarker discovery. Advances in next-generation sequencing (NGS) have enabled the comprehensive detection of small RNAs. However, extracting functionally meaningful signals from small RNA sequencing (sRNA-seq) data remains challenging due to expression heterogeneity, overlapping biogenesis pathways, and fragmented analysis workflows. Addressing these challenges requires an interdisciplinary approach combining plant biology, biochemistry, computer science, and machine learning. The primary objective of this thesis is to enhance RNA biomarker discovery through advanced computational methods. To this end, the research integrates computational prediction, statistical modeling of RNA biogenesis, scalable software development, and quality control strategies to support therapeutic feasibility. It presents a suite of bioinformatics methodologies for identifying, classifying, and interpreting small RNAs in both biological and therapeutic contexts.

To enable large-scale and reproducible miRNA biomarker discovery, the first contribution of this thesis introduces mirLibSpark, a distributed pipeline for plant miRNA prediction and annotation using Apache Spark. By incorporating widely accepted structural and expression criteria alongside target prediction and functional enrichment, mirLibSpark enables scalable and biologically credible analysis across multi-library datasets. While it provides built-in support for important species such as *Arabidopsis*, wheat, rice, and maize, mirLibSpark is designed to accommodate any user-provided plant genome. To uncover the coordinated roles of miRNAs and siRNAs under natural conditions, the second contribution investigates small RNA dynamics during cold acclimation in wheat. This study employes a unified Spark-parallelized workflow to sequencing libraries derived from field-collected wheat samples and identifies known and novel cold-responsive interfering RNAs, including dual-identity species with both miRNA-like and siRNA-like features. These findings offer new insights into the evolutionary origins and regulatory diversification of interfering RNAs in genotype-specific stress adaptation. To improve siRNA precursor prediction, the third contribution presents siWalk, a machine learning framework trained on *Arabidopsis* that combines biogenesis-aware sequence, structural, and expression features. While a Random Forest model achieved the highest predictive accuracy, a Gradient Boosting model is distributed as a compact and efficient alternative. siWalk also includes a sorting module that localizes effector siRNAs directly from precursor sequences, facilitating the rational design of RNA interference vectors. Finally, as an exploratory extension, ArcticFox was developed to assess the molecular integrity of RNA delivery vectors. Using long-read NGS data, ArcticFox profiles residual DNA and genome truncations in recombinant adeno-associated virus preparations and introduces quantification metrics to support quality control in RNA-based therapeutics. Built upon scalable computing, statistical modeling, and machine learning, these contributions form a reliable and interpretable foundation for advancing small RNA bioinformatics in high-throughput classification, integrative studies, and therapeutic applications.

Keywords: Small RNA bioinformatics, Machine learning in genomics, Functional RNA annotation, Scalable RNA analysis, RNA biomarker discovery, miRNA, siRNA

INTRODUCTION

0.1 Real-World Applications of Small RNAs in Agriculture and Medicine

Small RNAs play a vital role in how plants respond to stress and defend against pathogens, establishing their importance as biomarkers in agriculture. miR398 in *Arabidopsis thaliana* regulates superoxide dismutase (SOD) genes and is downregulated under oxidative stress, making it a reliable indicator of drought and heavy metal toxicity (Sunkar *et al.*, 2006). In tomato and soybean, the miR482/miR2118 superfamily plays a role in plant immunity by targeting nucleotide-binding leucine-rich repeat (NLR) resistance genes. Suppression of miR482 enhances resistance to *Phytophthora sojae*, making it a useful biomarker for pathogen attack (Shivaprasad *et al.*, 2012). Similarly, miR393 in *Arabidopsis* regulates auxin receptor genes and is upregulated upon *Pseudomonas syringae* infection, preventing bacterial manipulation of auxin signaling (Navarro *et al.*, 2006).

Beyond diagnostics, small RNAs offer therapeutic potential for disease resistance in plants. RNAi-mediated silencing targets *CYP51* genes in *Fusarium graminearum* to control fungal pathogens and enhance wheat resistance to Fusarium Head Blight (Koch *et al.*, 2013). Engineered artificial miRNAs have shown viral resistance for crop protection, successfully targeting *Tobacco Mosaic Virus (TMV)* in *Nicotiana benthamiana*, thereby inhibiting viral replication (Qu *et al.*, 2007).

Small RNAs also serve as critical biomarkers and therapeutic agents in animals, particularly in human medicine. miR-122 is a key biomarker for liver diseases such as hepatitis, fibrosis, and hepatocellular carcinoma, as its circulating levels reflect liver injury (Jopling *et al.*, 2005). Similarly, miR-155 plays a major role in immune response and inflammation. Its high presence in macrophages correlates with autoimmune diseases, such as rheumatoid arthritis and multiple sclerosis (O'Connell *et al.*, 2007). Furthermore, miRNA-based diagnostics such as GASTROclear, a blood test panel of 12 miRNAs for gastric cancer that has received FDA Breakthrough Device Designation, demonstrate the clinical utility of small RNAs in early cancer detection and risk stratification (So *et al.*, 2021).

Small RNAs have led to transformative developments in therapeutic treatments. Patisiran, the first FDA-approved siRNA drug, silences the *TTR* gene to treat hereditary transthyretin-mediated amyloidosis (Adams *et al.*, 2018). Inclisiran, another siRNA-based therapy, targets *PCSK9* gene to lower cholesterol and reduce cardiovascular risk (Ray *et al.*, 2020). Spinraza, an antisense oligonucleotide therapy, is the first FDA-

approved treatment for spinal muscular atrophy (SMA). It functions by modulating the splicing of the *SMN2* gene to restore survival motor neuron (SMN) protein levels, resulting in significant improvements in motor function for SMA patients (Finkel *et al.*, 2017). MRX34, a synthetic mimetic of miR-34, was developed as a cancer treatment due to its tumor-suppressive properties; however, clinical trials were stopped because of adverse effects to the immune system (Beg *et al.*, 2017). These examples demonstrate the utility of small RNAs as precision biomarkers and therapeutic agents, offering valuable applications in agricultural biotechnology and human medicine.

0.2 Computational Challenges for RNA Analysis

The discovery of RNA biomarkers presents significant computational challenges, primarily related to scalability and the need for comprehensive and reliable analysis pipelines. As sequencing technologies continue to advance, the volume of data has increased substantially, driven by the rising number of samples and the increasing depth of sequencing required for comprehensive small RNA profiling. This challenge is particularly pronounced in plant genomes, which are large and complex due to extensive genome duplication and human breeding. Existing tools often struggle to handle such massive datasets efficiently while maintaining high accuracy and reliability. Simply scaling up processing capabilities is not enough; making the analysis pipeline fast, reliable, and insightful requires integrating standard criteria for miRNA prediction, seamless incorporation of downstream analyses, and comprehensive presentation of results. As datasets continue to grow in size, with more datasets to process, more instances per dataset, and more features per instance, developing pipelines that meet the demands of scalability, reliability, and meaningful interpretation is becoming an increasingly critical challenge for accurate and efficient RNA biomarker discovery.

Accurate classification of small RNAs is complicated by the overlapping biogenesis pathways of siRNAs and miRNAs, the primary candidates for RNA biomarkers. Both often share common processing machinery, making it difficult to distinguish siRNAs generated through canonical pathways from those produced via alternative routes. Although miRNAs are relatively well annotated, the lack of a standardized framework for siRNAs hinders accurate prediction. This challenge becomes especially critical when designing siRNA precursors for RNA interference (RNAi)-based therapeutics. Evaluating whether a precursor produces the desired siRNA while minimizing unintended small RNAs that could compete or cause off-target effects requires analysis of sequence composition, structural preferences, and biogenesis signatures—features that can be learned from existing examples and applied to predict vector designs. Accurate classification and

prediction demand computational approaches that generalize across datasets while maintaining high accuracy, which is essential for understanding the functional roles of small RNAs in diverse biological contexts.

miRNAs and siRNAs participate in similar mechanisms to regulate plant stress responses and developmental processes, yet existing approaches remain fragmented, with miRNAs and siRNAs typically studied separately. This fragmented approach overlooks their potential interactions and co-regulatory networks, which are especially important for understanding complex stress responses such as cold acclimation in wheat. The challenge lies in jointly analyzing miRNAs and siRNAs to determine which small RNAs are involved in regulating stress adaptation and understanding their importance. Developing a computational workflow capable of integrating the analysis of both miRNAs and siRNAs is crucial for uncovering their complementary or redundant roles, evolutionary dynamics, and regulatory interactions. Such a workflow would enable the discovery of stress-responsive small RNAs contributing to enhancing resilience under environmental challenges. However, existing tools are not designed to facilitate this unified analysis, partly due to the perception that such integration is either unimportant or unfeasible. Demonstrating the feasibility and value of such an integrated approach would not only advance our understanding of small RNA regulation but also inspire future computational methods that incorporate more RNA classes into comprehensive analytical frameworks.

Developing RNA-based therapeutics presents additional computational challenges, particularly in ensuring genome integrity and therapeutic consistency. Viral vectors are commonly used for delivering small RNAs to target tissues, but their inherent susceptibility to mutations, genome rearrangements, and incomplete packaging complicates the delivery process. Conventional quality control techniques generally rely on indirect measurements of viral integrity, purity, and efficacy, but these methods suffer from low resolution and are inadequate for providing detailed nucleotide-level information about the viral genome. While useful for detecting gross abnormalities, they fail to offer a comprehensive view of the entire viral genome. In contrast, next-generation sequencing (NGS) offers precise, nucleotide-level resolution while simultaneously providing a global view of the viral genome. Nevertheless, protocols for directly reading genomes have yet to be established, and standardized metrics for assessing viral genome integrity are still in development. Additionally, the potential benefits and limitations of using NGS for viral quality control have not been systematically explored. Addressing these gaps requires the development of robust computational models that can accurately interpret NGS data and ensure high-quality RNA delivery, thereby enhancing the safety, consistency, and efficacy of RNA-based therapeutics.

0.3 Thesis Overview

Objective. The *primary objective* of this thesis is to *develop effective new computational methods for discovering biomarkers from small RNAs, with a specific focus on interfering RNAs (miRNAs and siRNAs)*. The goal is achieved by developing scalable pipelines for high-throughput miRNA prediction, integrating miRNA and siRNA analysis to uncover their regulatory roles in plant stress adaptation, applying machine learning for siRNA prediction and annotation, and establishing quantitative profiling frameworks for quality control in RNA-based therapeutics. To address the challenges associated with achieving this primary objective, the thesis presents a series of innovative computational methodologies. Together, these advancements strengthen RNA biomarker discovery, functional annotation, and the development of RNA-based therapeutics. While small RNAs have important applications in both animals and plants, this thesis focuses on plant small RNAs, whose biogenesis differs substantially from that in animals. Valuable insights and developed algorithms can still be adapted between kingdoms when appropriate.

Thesis Structure. This thesis is organized to reflect a progression from conceptual foundations to original methodological contributions. Chapter 1 provides biological and computational background relevant to small RNA research, establishing the necessary context for the work. Chapter 2 articulates the research problems and objectives that guide the thesis. Chapter 3 presents a literature review of current computational methods for small RNA prediction, focusing on methodological gaps addressed by this work. Chapters 4 through 7 detail the four main contributions, each corresponding to a specific computational framework developed to solve a distinct problem in RNA biomarker analysis or RNA therapeutic delivery. Chapter 8 concludes the thesis by summarizing insights across contributions, identifying limitations, and outlining directions for future research.

- Chapter 1: Background
- Chapter 2: Research Problems
- Chapter 3: Literature Review
- Chapter 4-7: Original Methodological Contributions
- Chapter 8: Conclusion

The research methodologies developed in this thesis are deeply interconnected. Chapter 4 describes the first contribution of the thesis, **mirLibSpark**, which introduces an Apache Spark-based pipeline designed

to efficiently process multi-library miRNA sequencing data, significantly improving scalability and reducing processing time. This work demonstrated how a high-performance framework can transform small RNA prediction, laying the groundwork for more advanced analytical tools. The experience gained from mirLibSpark played a pivotal role in shaping the development of the siRNA predictor, **siWalk**, reinforcing the importance of integrating computational efficiency with predictive accuracy. Additionally, mirLibSpark's reusable annotation modules, which incorporate multi-library processing, statistical testing, and functional enrichment analysis, streamlined the annotation of both miRNAs and siRNAs and proved instrumental in advancing the wheat small RNA workflow.

Chapter 5 introduces the second contribution of the thesis, which examines the evolutionary dynamics of wheat small RNAs in cold acclimation, investigating how miRNAs and siRNAs orchestrate stress responses through comparative genomics and transcriptomics. By analyzing their regulatory roles across cultivars and conditions, we gained deeper insight into their functional conservation and adaptive significance. This work also revealed key limitations in existing siRNA prediction methods, highlighting the need for more precise and reliable computational tools. These findings directly influenced the development of **siWalk**, a machine learning-based methodology designed to improve siRNA precursor identification and classification.

Chapter 6 presents the third contribution of the thesis, **siWalk**, which improves upon traditional siRNA prediction methods that primarily rely on phase scoring by integrating structural and sequence-derived features into a machine-learning model. This approach enables more precise classification and reduces false positives. What makes siWalk particularly innovative is its ability to capture complex patterns in siRNA biogenesis that scoring-based methods often miss. This advancement extends siWalk's utility beyond classification by supporting siRNA vector design, enabling precise localization of plausible siRNA production sites within a given precursor.

Chapter 7 presents the fourth contribution of the thesis, rAAV profiling with **ArcticFox**, which addresses the critical need for quality control in RNA-based drug delivery. As viral replication and packaging are inherently error-prone, ArcticFox applies long-read sequencing technologies to assess genome integrity with high resolution. The pipeline revealed biases in sequencing related to molecule selection and read completion, which can compromise the accuracy of evaluation. By incorporating spike-in controls and mathematical corrections, ArcticFox mitigates these biases, improving the consistency and reliability of quality control. This contribution extends the thesis beyond small RNA prediction into practical deployment, ensuring safe

and reliable delivery of therapeutics identified through earlier methods.

0.4 Contributions

The main contributions of this thesis are supported by a series of research articles that have been published, submitted, or are currently in preparation. These are listed below.

- Chapter 4 (MirLibSpark): **Wu CJ**, Remita AM, Diallo AB. MirLibSpark: A scalable NGS plant microRNA prediction pipeline for multi-library functional annotation. *In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp. 669-674. ACM*, **2019**. DOI 10.1145/3307339.3343463 (Article published).

- Chapter 5 (wheat): **Wu CJ**, Rahmdani F, Agharbaoui Z, Badawi M, Danyluk J, Li Q, Fowler DB, Sarhan F, Diallo AB. Prediction and evolutionary dynamics of wheat miRNAs and siRNAs during field cold acclimation. Manuscript summary was submitted in **2025** to the Research Topic *Plant Responses to Environmental Change* in *Frontiers in Plant Science* for scope review. (Article in preparation)

- Chapter 6 (siWalk): **Wu CJ**, Danyluk J, Diallo AB. Méthodes d'analyse fonctionnelle des petits ARNs – SiWalk: Pipeline NGS de prédiction siARN chez les plantes. *Dans la 2e journée de la recherche de la Faculté des sciences, UQÀM*, Poster C5, **2020**. (Poster published).

  Wu CJ, Diallo AB. Leveraging machine learning for improved siRNA prediction in *Arabidopsis*. (Article in preparation; intended for submission to *Bioinformatics, Oxford Academic*, **2025**).

- Chapter 7 (ArcticFox): **Wu CJ**, Liu HW, Carlile T, Lin D, Zhang B, Diallo AB. High-Resolution Characterization of rAAV Genomes for Vector Quality Using Long-Read Sequencing. BioRxiv, **2025**. DOI 10.1101/2025.10.05.680573 (Preprint published).

  Wu CJ, Liu HW, Tereshko L, Lin D, Bergelson S, Zhang B, Diallo AB, Mason C. Assessing functional impurities in rAAV production platforms by long-read sequencing. *In the 31st Conference on Intelligent Systems for Molecular Biology (ISMB) & the 22nd European Conference on Computational Biology (ECCB), Poster B-159, **2023***. DOI 10.13140/RG.2.2.17884.56962 (Poster published).

  Wu CJ, Liu HW, Lin D, Bergelson S, Zhang B, Diallo AB, Zhang W. A quantitative profiling framework for rAAV genome population. *In the 25th Annual International Conference on Research in Computational Molecular Biology, RECOMB, Poster 321, **2021***. (Poster published).

CHAPTER 1
BACKGROUND

This chapter introduces the fundamental concepts that drive the approaches developed in this thesis. As biological data grows in complexity and volume, effective analysis demands interdisciplinary strategies. Our research lies at the intersection of biochemistry, plant biology, computer science, and machine learning — a convergence essential for unlocking the potential of research methods to tackle the core challenges of this work. In bioinformatics, we aim to develop computational tools that transform small RNA data into valuable insights for biomarker discovery and therapeutic applications.

1.1 RNA Biomarkers and Therapeutic Applications

Biomarkers are measurable indicators of biological states, widely used in health state diagnostics and prognostics, treatment monitoring, and drug development. They can take various forms, including physiological signs (e.g., heart rate, body temperature), genetic variations (e.g., single nucleotide polymorphisms, DNA mutations, RNA expression signatures), and genetic products such as proteins (e.g., urinary albumin) and metabolites (e.g., cholesterol). Each type of biomarker provides unique biological insights, with applications spanning medicine, agriculture, and environmental monitoring.

RNA-based biomarkers have gained increasing attention due to their dynamic regulatory roles and responsiveness to environmental and physiological changes. RNAs reflect real-time cellular processes, making them valuable tools for early disease detection, predicting treatment responses, and functionally characterizing biological pathways (Bajan et Hutvagner, 2020; Tribolet *et al.*, 2020). Various RNA classes, including messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs), and small RNAs (sRNAs), contribute to the development of useful biomarkers.

Small RNAs stand out for their ease of manipulation and delivery. Specialized sequencing techniques identify this functional class by detecting RNA transcripts shorter than 50 nucleotides. Compared to larger RNA molecules, small RNAs are easier to detect and can be efficiently introduced into cells via viral vectors, synthetic carriers, or direct transfection, facilitating their use in both monitoring applications and therapeutic interventions. Small RNAs are particularly useful in non-invasive diagnostics. Liquid biopsies can detect circulating miRNAs associated with cancers, cardiovascular diseases, and neurological disorders (Bajan et

Hutvagner, 2020). In plants, small RNAs serve as molecular indicators of stress responses, growth conditions, and environmental adaptation (Khraiwesh *et al.*, 2012). Their stability, robustness across diverse conditions, and reproducibility in experimental and clinical settings, along with their accessibility and reliability, establish them as powerful biomarkers for both clinical and agricultural applications (Bej et Basak, 2014; Bajan et Hutvagner, 2020). **siRNAs** (small-interfering RNAs) and **miRNAs** (microRNAs) are particularly notable for their ability to modulate gene expression post-transcriptionally (Hutvágner et Zamore, 2002). Their roles in developmental regulation, disease progression, animal immune responses, and plant stress responses make them valuable indicators for crop resilience and major human diseases.

RNA-based therapeutics span a wide range of RNA molecules, including siRNAs, miRNAs, antisense oligonucleotides (ASOs), and CRISPR. These molecules have significant potential to regulate gene expression and modulate disease pathways. They function through distinct mechanisms (Bajan et Hutvagner, 2020). siRNAs and miRNAs participate in the RNA interference (RNAi) pathway, binding to complementary mRNA sequences to induce degradation or translational repression, making them useful for silencing disease-associated genes. ASOs are single-stranded oligonucleotides that hybridize with target RNAs to either block translation or induce degradation. CRISPR–Cas9 utilizes a guide RNA to direct Cas9 to specific genomic loci, enabling gene editing at the DNA level. Recent advancements in structural characterization techniques, including NMR spectroscopy and X-ray crystallography, along with computational modeling and rational design, have accelerated the development of RNA-targeted therapies. Enhanced delivery technologies, such as nanoparticles and viral vectors, have further improved RNA-based therapeutics by overcoming physiological barriers and enhancing stability, bioavailability, and specificity. These advancements have enabled several promising small RNA therapies to reach clinical trials and enter the market. Moreover, the success of Moderna's mRNA vaccine highlights the expanding scope of RNA-based technologies, demonstrating their potential for achieving targeted protein expression and modulating immune responses.

1.2 Definitions of Small RNA and Other Biological Concepts

1.2.1 Small RNA

Small RNA (sRNA) refers to RNA molecules captured by the **sRNA-seq** technique, which detects RNA transcripts shorter than 50 nucleotides from a biological sample. These short RNA molecules include:

- MicroRNAs (miRNAs) and small interfering RNAs (siRNAs): Small RNAs involved in RNA interference, guiding the cleavage or inhibition of target RNA through perfect or near-perfect base-pairing.

- Other RNAs: Includes fragments derived from degraded mRNAs, rRNAs, tRNAs, snoRNAs, or other RNA species.

1.2.2 NGS Techniques

Next-Generation Sequencing (NGS) is a high-throughput technology that enables rapid sequencing of RNA (and DNA) molecules. NGS techniques can be broadly divided into two protocols:

- Short-Read Sequencing: RNA molecules are fragmented into shorter segments by enzymatic or mechanical methods prior to sequencing, a process commonly known as RNA-seq. Alternatively, selective size enrichment can be performed to focus on smaller RNA molecules, referred to as **sRNA-seq**. Typical platforms, such as Illumina HiSeq and MiSeq, produce reads generally shorter than 500 base pairs (bp) and are limited by their maximum read length capacity. Some produce reads as short as 50 bp.
- Long-Read Sequencing: Full-length RNA molecules are directly sequenced, allowing the capture of native RNA molecules without prior fragmentation. Typical platforms include Oxford Nanopore and PacBio, which offer virtually unlimited read lengths.

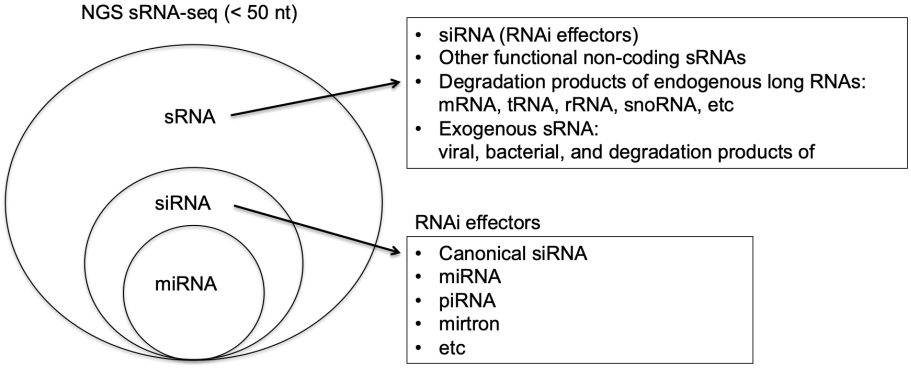


Figure 1.1: Classification of small RNAs collected from NGS sRNA-seq. RNAs expressed in cells are categorized into coding and non-coding RNAs. Coding RNA refers to messenger RNA (mRNA). Non-coding RNAs include ribosomal RNA (rRNA), transfer RNA (tRNA), and small RNAs. Captured short RNAs consist of small interfering RNA (siRNA), small nucleolar RNA (snoRNA), other small non-coding RNAs, and degradation products from mRNA, tRNA, rRNA, snoRNA, and exogenous sRNA. RNAi-effector siRNAs are further classified based on precursor structure, length, and tissue origin into subclasses such as canonical siRNA, microRNA (miRNA), PIWI-interacting RNA (piRNA), and mirtron.

1.2.3 Nucleotides and the Flow of Genetic Information

Life is fundamentally encoded in the language of nucleotides. Nucleotides are the basic building blocks of nucleic acids (DNA and RNA). Each nucleotide consists of a nitrogenous base, a five-carbon sugar (deoxyribose in DNA, ribose in RNA), and one or more phosphate groups. The nitrogenous **bases** include adenine, cytosine, guanine, as well as thymine (in DNA) and uracil (in RNA).

Nucleotides are oligomerized into macromolecules known as DNA and RNA. These nucleic acids have directional sequences defined by their 5' (five prime) and 3' (three prime) ends, corresponding to the orientation of the sugar-phosphate backbone. **DNA** (deoxyribonucleic acid) serves as the primary genetic material and is typically double-stranded in most organisms. **RNA** (ribonucleic acid) functions either as a messenger encoding proteins or as a functional molecule involved in cellular processes.

Coding RNA refers to messenger RNA (mRNA), which is translated into proteins. **Non-coding RNA** (ncRNA) functions without translation and includes microRNA (miRNA), small interfering RNA (siRNA), long non-coding RNA (lncRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA).

Proteins are amino acid chains folded into functional three-dimensional structures. A codon is a three-nucleotide sequence in mRNA that specifies an amino acid. During translation, codons are recognized by complementary anticodons on tRNA molecules.

The central dogma of molecular biology explains how genetic information flows within cells. It describes that genetic information stored in DNA is **transcribed** into RNA, which is subsequently **translated** into proteins. In addition, DNA is replicated to produce exact copies, allowing genetic information to be inherited across generations.

Omics. The **genome** refers to the complete set of DNA in an organism. The **transcriptome** represents the complete set of RNA molecules transcribed from the genome under a specific state, reflecting gene expression patterns. The **proteome** encompasses the complete set of proteins translated from the transcriptome and represents the functional output of gene expression. Together, these molecular layers enable the transmission, interpretation, and execution of genetic instructions within living systems.

1.2.4 Base Pairing and Molecular Interaction

Molecular interactions drive the structure and function of nucleic acids, primarily mediated by hydrogen bond-based pairing, which ensures both specificity and stability. Canonical base pairing follows specific rules: Adenine (A) forms two hydrogen bonds with Thymine (T) in DNA or with Uracil (U) in RNA, while Cytosine (C) forms three hydrogen bonds with Guanine (G) (Figure 1.2). In addition to canonical pairs, wobble base pairing occurs, with the most common example in RNA being Guanine-Uracil pairing.

- DNA-DNA Interaction: Base pairing between complementary DNA strands forms the double-stranded DNA structure, which adopts a secondary structure known as the double helix. This structure is further condensed into one or several chromosomes, allowing efficient storage and organization of genetic material.
- RNA-RNA Interaction: Base pairing between RNA molecules is essential for regulatory mechanisms. For example, microRNA (miRNA) binds to target mRNA through complementary base pairing to influence gene expression. Transfer RNA (tRNA) utilizes its anticodon region to recognize complementary codons on mRNA during protein synthesis, ensuring accurate translation.

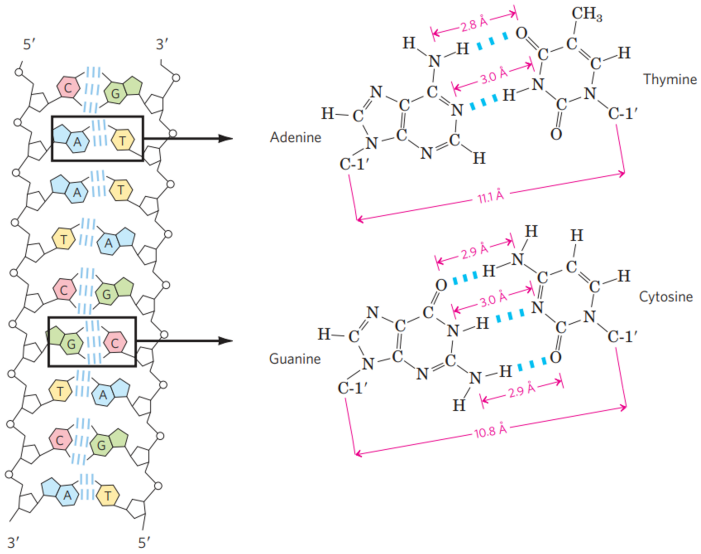


Figure 1.2: DNA, nucleotides, and base pairing. The left panel shows a double-stranded DNA oligomer with complementary base pairing. The right panel illustrates the chemical structures of adenine (A), thymine (T), cytosine (C), and guanine (G). Hydrogen bonds between complementary bases are represented by blue dashed lines, with A-T pairs forming two hydrogen bonds and C-G pairs forming three hydrogen bonds. Adapted from (Nelson et Cox, 2000).

1.2.5 Textual Representation of RNA Sequences

RNA sequences are commonly represented in text format using single-letter nucleotide symbols: A for Adenosine, U for Uracil, C for Cytosine, and G for Guanine. Occasionally, RNA sequences are kept in their original DNA format, with T (Thymine) instead of U, allowing seamless referencing of DNA and RNA sequences without extra conversion. The **primary structure** of RNA is the linear sequence of nucleotides. For example (pre-miRNA of cel-lin-4, miRBase Accession MI0000002. Regions with capital letters are miRNA and miRNA*, respectively):

augcuuccggccugu UCCCUGAGACCUCAAGUGUGA guguacuauugaugcuuc ACACCUGGGCUCUCCGGGUACC aggacgguuugagcagau

RNA molecules fold into **secondary structures** in physiological conditions. One way to describe these structures is through linear paired alignment format, where "-" indicates a skipped nucleotide, and "|" denotes base pairing:

```
--a     ---   g    -uU   U    C    A       u   - u
   ugcuu   ccg ccug   CCC GAGA CUCA GUGUGAg gua c a
   |||||   ||| ||||   ||| |||| |||| ||||||| ||| |
   acgag   ggc ggaC   GGG CUCU GGGU CACAcuu cgu g u
uag     uuu   a    CAU   C    C    C       -   a u
```

Another common representation of RNA secondary structures is the **dot-bracket** notation (Lorenz *et al.*, 2011), where **paired** bases are shown as **parentheses** and **unpaired** bases as **dots**. For the same sequence:

.(((((((.((((..(((.((((.((((.(((((((.((((....).)))))))))).)))).)))).)))...)))))...)))))...

Structural elements commonly observed in RNA molecules include: **Bulges** occur when base pairing is skipped for a few nucleotides. **Loops** form when unpaired nucleotides enclosed by base-paired regions. **Hairpins** are sequences folding back on themselves, creating a double-stranded **stem** and a loop (Figure 1.3).

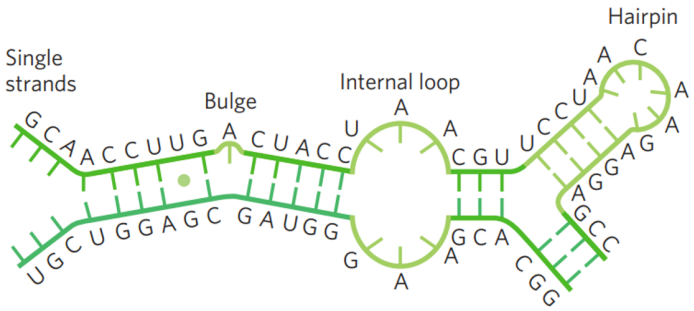


Figure 1.3: RNA secondary structure. Single strands are unstructured regions. Bulges occur when base pairing is skipped for a few nucleotides. Loops are unpaired nucleotides enclosed by base-paired regions. Wobble GU pairing is shown. Hairpins consist of double-stranded stems with a terminal loop. Adapted from (Nelson et Cox, 2000).

1.2.6 In Silico, In Vivo, and In Vitro

In Silico: Refers to experiments conducted through computer simulations or computational models rather than real-world settings. Tools such as miRDeep-P, RNAfold, and miRanda are used to generate *in silico* predictions. ***In Vitro*** and ***In Vivo***: Laboratory-based experiments validate *in silico* predictions. *In vitro* studies are performed in artificial environments like test tubes, petri dishes, or culture systems, while *in vivo* studies occur within living organisms, such as plants, animals, or microorganisms. Techniques such as transgenic models, knockout experiments, and overexpression assays provide empirical evidence supporting computational findings.

1.3 RNA Feature Representation

The representation of RNA features is heterogeneous, derived from techniques such as counting, tokenization, annotation tools, cross-database annotation, and comparison against reference genotypes. Key features are described below.

RNA **sequence features** include properties directly derived from the nucleotide sequence. **Nucleotide composition** describes the proportion of different nucleotide types within an RNA sequence, such as G/C content and specific nucleotide positions critical for processing (e.g., 5' end positions 1 and 2). **Motif frequency** refers to the occurrence of motifs, *k*-mers, or structural patterns within RNA sequences. **Recognition sites**

refer to specific sequences or structures recognized by cellular machinery, such as miRNA recognition sites, restriction sites, and localization signals. **Effector sites** denote predicted miRNA or siRNA locations within precursor sequences. For example, miRdup (Leclercq *et al.*, 2013) uses a random forest classifier trained on validated miRNA duplexes to predict miRNA localization. ***K*-mers** are short sequences of length k representing primary sequences or structural elements. For example:

- 1-mer: A, T, C, G, d (dot), b (bracket).
- 2-mer: CG, AA, db, dd, bb.
- 3-mer: ATC, CGG, ddb, bbd.

Structural features capture properties derived from predicted RNA secondary structure. **Structural elements** commonly observed in RNA include bulges, loops, hairpins, and stems. **Length** can refer to entire RNA molecules or specific structural elements such as bulges and loops. **Minimal Free Energy (MFE)** is the predicted free energy of RNA secondary structures, indicating structural stability, and is calculated by tools like RNAfold (Lorenz *et al.*, 2011).

Functional features describe biological interactions and annotations associated with RNA molecules. **Target prediction** uses computational methods to predict small RNA-target interactions. For instance, miRanda (John *et al.*, 2004) predicts miRNA-mRNA interactions based on sequence complementarity and thermodynamic stability. **Cross-database functional annotation** integrates various databases for improved annotation. This includes sequence validation from miRBase (Kozomara *et al.*, 2019), functional categorization via Gene Ontology (GO) (Consortium *et al.*, 2004), and pathway mapping using Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2017). These annotations contextualize small RNA functions across cellular components, molecular activities, and biological processes, as illustrated in Figure 1.4.

Biogenesis-derived features describe the molecular processes responsible for generating small RNAs. **Dicer processing** refers to the generation of siRNAs by Dicer cleavage, producing siRNAs with characteristic 3' overhangs that are crucial for RISC loading. **siRNA classes** distinguish different subclasses, such as canonical miRNAs, tasiRNAs, nat-siRNAs, hc-siRNAs, and mirtrons, each following distinct biogenesis pathways.

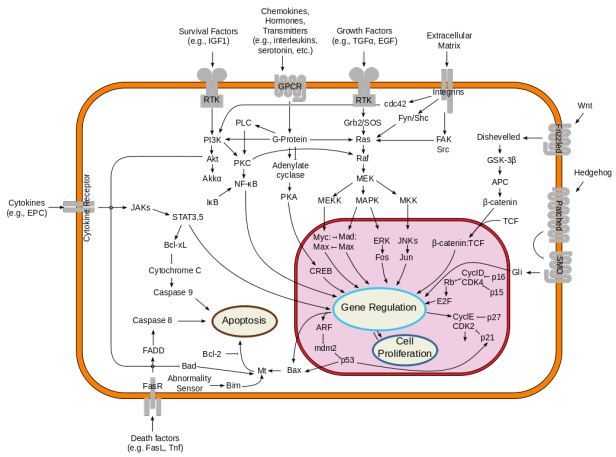


Figure 1.4: Functional annotation of sRNA. The figure illustrates how gene function can be contextualized within cellular localization and regulatory systems. miRNA and siRNA functions are commonly inferred through target prediction followed by annotation using databases such as GO and KEGG, which categorize targets by cellular components (e.g., membrane, nucleus), molecular functions (e.g., kinase activity), and biological processes (e.g., apoptosis). Image credit: CleanPNG.

Statistical features quantify the properties of RNA molecules based on experimental data. **Counting** includes the quantification of RNA molecules, base pairs, phased positions, and miRNA recognition sites. **Abundance** reflects RNA expression levels and is often *normalized* to parts per million (PPM) to account for sequencing depth and experimental variability. **Strand expression ratio** measures the ratio of RNA expression between complementary strands, indicating transcriptional bias or regulation. **Differential expression** compares expression levels across conditions to identify functional RNAs. This process often involves calculating the fold change of expression levels relative to a reference condition, followed by a *log transformation* to stabilize variance and improve interpretability. Statistical testing determines the significance of observed differences, often corrected for multiple testing using methods such as the *False Discovery Rate (FDR)*. **Transformed features** are derived from statistical or mathematical transformations to enhance data interpretability. This includes phasing scores, normalized expression levels, and other derived metrics.

Genotype-based representations compare RNA sequences against a reference genotype to encode similarity and variant information. A binary encoding scheme highlights differences between the sample and reference sequences by comparing genetic markers against a reference genotype and assigning a score of '1' (high similarity) or '0' (low similarity). The similarity is determined by applying a threshold T to the distance

$\boldsymbol{D}$ between the markers, which can be calculated using the Hamming Distance (Equation 1.5). For example, given a reference sequence $x = \{C, T, G, A\}$ and a sample $y = \{C, T, C, A\}$, the resulting genotype vector is $\{1, 1, 0, 1\}$ (assuming $T = 0$ for exact match). Genetic markers can be longer sequences, not just single nucleotides. This encoding supports additional summary statistics such as **coverage** (the fraction of positions represented), **positivity rate** (the proportion of markers matched), **intact rate** (the proportion of full-length matches), and **mass percentage** (the nucleotide-level contribution of a sequence relative to the total). These metrics provide a compact summary of alignment quality and genomic representation.

1.4 Modeling Features for RNA Classification Tasks

Once RNA sequences are represented as quantitative and structural features (Section 1.3), these features must be modeled to extract biologically meaningful patterns and support downstream functional analysis. Modeling provides the computational bridge between raw feature representations and biological interpretation by formalizing how features are filtered, combined, and evaluated under different assumptions and data scales. In practice, this modeling process encompasses multiple complementary strategies, ranging from deterministic preprocessing and rule-based filtering to statistical inference and data-driven learning approaches. Each strategy offers distinct advantages depending on the complexity of the data, the availability of prior biological knowledge, and the analytical objective. Accordingly, this thesis adopts three modeling perspectives: bioinformatics concepts and preprocessing (Section 1.5), statistical modeling (Section 1.6), and machine learning (Section 1.7), which are described in detail in the following sections.

1.5 Bioinformatics Concepts and Preprocessing

1.5.1 Benchmark Datasets

This thesis draws on diverse small RNA and long-read sequencing datasets collected from both public and in-house sources. These datasets serve as benchmarks to evaluate the scalability, predictive performance, and generalizability of the computational methodologies developed for miRNA prediction, siRNA classification, and RNA therapeutic quality control.

sRNA-seq libraries. Arabidopsis: A core set of 29 *Arabidopsis thaliana* sRNA-seq libraries (total 1 Gigabyte, 1.2×10^8 reads) was used for both miRNA and siRNA analyses, including two individual accessions (GSM518432 and GSM738727) (Lei et Sun, 2014) and 27 libraries from the GSE44622 series (Jeong *et al.*,

2013). These libraries represent a range of tissues and conditions in *Arabidopsis*. **Wheat**: To investigate small RNA regulation during cold stress adaptation, 80 sRNA-seq libraries (total 250 Gigabytes, 1.5×10^9 reads) were generated from four *Triticum aestivum* cultivars with contrasting freezing tolerance and vernalization responses. Plants were grown under field conditions in Saskatoon across two years (2010 and 2013) and sampled at five time points between early September and November. Each cultivar contributed 20 libraries, capturing the temporal dynamics of small RNA expression.

miRNA benchmark data. The miRNA prediction task uses positive examples curated from miRBase version 21 (Kozomara *et al.*, 2019), which defines high-confidence miRNA entries. To ensure sequence-level uniqueness, a total of 349 distinct mature *Arabidopsis* miRNA sequences were identified, from which 100 high-confidence entries were selected as the positive set. A corresponding negative set was generated by randomly extracting 1000 genomic sequences of 18–25 nucleotides, excluding known miRNAs and annotated noncoding RNAs based on TAIR10. This setup reflects the natural class imbalance and sequence diversity present in plant genomes. The 29 core *Arabidopsis* sRNA-seq libraries, along with one wheat library, were used to evaluate the scalability and cross-species applicability of the miRNA prediction pipeline.

siRNA benchmark data. For siRNA classification, a reference set of 21-nt phased siRNA loci was curated from four studies: Chen *et al.* (2007), Howell *et al.* (2007), Nakano *et al.* (2020), and Chen *et al.* (2021a). Overlapping loci were unified into 51 standardized genomic reference contigs (5250 nt each). To obtain features associated with the generating loci, segment candidates were extracted from the 29 core *Arabidopsis* sRNA-seq libraries and evaluated for expression and phasing. A segment was labeled positive if it overlapped a reference locus, showed at least two significant phase scores (FDR < 0.05), and contained a dominant effector siRNA expressed in at least two libraries; all other segments were labeled negative. This process yielded 10,683 positive and 174,856 negative training examples.

Long-read sequencing datasets for rAAV analysis. To support RNA therapeutic quality control, 18 long-read sequencing (LRS) datasets (8 Gigabytes) were analyzed, profiling recombinant adeno-associated virus (rAAV) genomes. These included both in-house samples (average 1.7×10^8 reads per library), sequenced using Oxford Nanopore (MinION) and PacBio (Sequel II), and public datasets (6×10^3 to 3×10^4 reads per library) from SRA projects PRJNA383145 (Tai *et al.*, 2018), PRJNA608034 (Tran *et al.*, 2020), and PRJNA794197 (Tran *et al.*, 2022). The datasets span different production systems and transgenes (e.g., EGFP, Cas9, mCherry). Lambda DNA spike-ins were included to assess sequencing bias.

1.5.2 Dataset Definition

Let $S = \{s_1, s_2, \ldots\}$ denote the set of benchmark libraries, which may include 29 *Arabidopsis* sRNA-seq libraries, 80 wheat sRNA-seq libraries, and/or 18 LRS libraries.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of reads obtained from an sRNA-seq library $s \in S$, where each read x_i is a sequence drawn from the alphabet $\{A, T, C, G\}$ and has a length constrained by $15 \leq |x_i| \leq 50$. The total number of reads per library typically satisfies $10^7 \leq n \leq 5 \times 10^8$, where $n = |s|$.

Each x_i is represented as a vector of features, denoted by $\{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, where p is the total number of features. These features, explained in Section 1.3, are generated using specialized software, algorithms, and supporting documentation. They include sequence-derived, structural, functional, statistical, and biogenesis-related properties, with $100 \leq p \leq 250$ in the original feature space. The feature set is heterogeneous, comprising numerical values bounded between 0 and 1, unbounded continuous variables, categorical descriptors, and cardinal features. To enable downstream machine learning analysis, feature transformation is applied to standardize these representations. After transformation, the dimensionality increases, yielding $p' > 1000$.

For LRS libraries, the read length extends to $15 \leq |x_i| \leq 10^5$, with the number of reads typically in the range $10^3 \leq n \leq 10^7$. Feature representation in this context emphasizes sequence similarity between reads and reference genomes, commonly computed using Hamming distance and genotyping statistics, as described in Section 1.3 (genotype representations).

1.5.3 Feature Transformation

The heterogeneity of RNA features requires preprocessing to transform them for compatibility and improve computational modeling. Features derived from biological sequences, structures, functions, statistics, and biogenesis differ in formats and scales, which can introduce biases during model training. Categorical features like nucleotide sequences and structural motifs are converted into numerical representations, while quantitative features such as expression levels and statistical scores are normalized and scaled (Pargent *et al.*, 2022). Different transformation techniques are applied depending on the nature of the features, outlined below:

- **One-Hot Encoding**: This technique transforms categorical sequence features, such as nucleotide com-

positions and *k*-mers, into a binary format suitable for computational analysis. For example, nucleotide sequences (A, T, C, G) are represented as:

$$A = [1, 0, 0, 0], \quad T = [0, 1, 0, 0], \quad C = [0, 0, 1, 0], \quad G = [0, 0, 0, 1]$$

- **Min-Max Scaling**: Transforms numerical features to a fixed range, usually [0, 1], using the formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1.1}$$

- **Z-score Normalization**: Transforms numerical features to have a mean of zero and a standard deviation of one. This scaling is effective when the feature distributions are approximately symmetric.

$$X' = \frac{X - \mu}{\sigma} \tag{1.2}$$

where μ is the mean and σ is the standard deviation of the original feature values.

- **Quantile Transformation**: This technique maps data to a uniform or normal distribution, ensuring robustness to outliers while preserving feature rankings. It is particularly useful when data is highly skewed.

$$X' = F^{-1}(F(X)) \tag{1.3}$$

where F is the cumulative distribution function of the data.

- **Log Transformation**: Log transformation reduces skewness in distributions, particularly when there are large differences in magnitude. It is commonly applied to expression fold change relative to a reference point,

$$X' = \log(X + 1) \tag{1.4}$$

- **Ordinal Encoding**: Ordinal encoding maps categorical variables with intrinsic order to numerical values. Care must be taken to ensure that the rank ordering reflects biological significance. For example, miRNA precursor structure types can be assigned values as follows:

$$\text{Hairpin} = 0, \quad \text{Loop} = 1, \quad \text{Bulge} = 2$$

1.5.4 Distance Functions

The Distance Function, denoted by $\boldsymbol{D}$, quantifies the difference between two observations x_1 and x_2. Each observation consists of measurements over variables $i = 1, 2, \ldots, p$, which are also referred to as attributes. The most commonly used distance functions are defined as follows (Hastie *et al.*, 2009).

Hamming Distance.

$$\boldsymbol{D}_{Hamming}(x_1, x_2) = \sum_{i=1}^{p} \mathbb{I}(x_{1i} \neq x_{2i}) \tag{1.5}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the two elements are different and 0 if they are the same.

Euclidean Distance.

$$\boldsymbol{D}_{Euclidean}(x_1, x_2) = \sqrt{\sum_{i=1}^{p} (x_{1i} - x_{2i})^2} \tag{1.6}$$

The Euclidean distance measures the straight-line distance between two observations in a multi-dimensional space.

Squared Euclidean Distance.

$$\boldsymbol{D}_{Squared\,Euclidean}(x_1, x_2) = \sum_{i=1}^{p} (x_{1i} - x_{2i})^2 \tag{1.7}$$

The Squared Euclidean distance is a variation of the Euclidean distance where the square root operation is omitted for computational efficiency.

1.6 Statistical Modeling

Statistical modeling provides quantitative tools to describe variation in RNA data, detect expression changes between conditions, and evaluate biological significance. This section introduces commonly used probability distributions for modeling RNA-related data, followed by methods for differential expression analysis, multiple testing correction, and enrichment testing to support biological interpretation.

1.6.1 Statistical Distributions

1.6.1.1 Normal Distribution Model

A continuous random variable X follows a Normal distribution with parameters μ (mean) and σ^2 (variance), if its probability density function is defined as (Devore *et al.*, 2012):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x, \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+ \tag{1.8}$$

In this distribution, μ indicates the center of the distribution, while σ determines the spread of values around the mean. The Normal distribution is used to model continuous variables that exhibit a symmetric,

bell-shaped distribution. In bioinformatics, it is often applied to normalized gene expression data, noise of measurement, or other biological quantities assumed to be approximately normally distributed.

1.6.1.2 Gamma Distribution Model

A continuous random variable X has a Gamma distribution with parameters α and β, if its probability density function is defined as (Devore *et al.*, 2012):

$$f(x) = \frac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^{\alpha}\, \Gamma(\alpha)}, \quad x, \alpha, \beta \in \mathbb{R}^{+} \tag{1.9}$$

$$\text{where} \quad \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt$$

The Gamma distribution is commonly applied to model non-negative, right-skewed continuous data.

1.6.2 Differential Expression Analysis

Differential expression analysis identifies small RNAs whose expression levels vary significantly across biological conditions. The process begins with raw read counts obtained from sRNA-seq data, where each count reflects how many sequencing reads are assigned to a given small RNA in each sample. These counts are discrete and often overdispersed, with variability greater than expected under simpler models. To address this, the counts are first normalized to adjust for differences in library size, and then modeled using negative binomial regression (NB), a generalized linear model designed for discrete count data with overdispersion (Love *et al.*, 2014). In this framework, the observed count K_{ij} for small RNA i in sample j is modeled as $K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$, where μ_{ij} is the expected count, and α_i is a dispersion parameter that captures biological variability for RNA i. The expected count is linked to experimental conditions using a log-linear model:

$$\log(\mu_{ij}) = \log(s_j) + \beta_{0i} + \beta_{1i} x_j \tag{1.10}$$

Here, s_j is a scaling factor for sequencing depth in sample j, and x_j encodes the experimental condition (e.g., 0 for control, 1 for treatment). The coefficient β_{1i} reflects the change in expression between conditions on the natural log scale and is typically converted to a $\log_2$ fold change for interpretability.

The statistical significance of the expression change is assessed by testing the null hypothesis:

$$H_0 : \beta_{1i} = \beta_0$$

typically with $\beta_0 = 0$, indicating no differential expression. A test function φ evaluates this hypothesis using a test statistic, producing a p-value. If $p < \alpha$, meaning the p-value is below the significance level, we reject H_0 and conclude that small RNA i is differentially expressed.

1.6.3 Multiple Testing

In high-throughput studies such as sRNA-seq, thousands of hypotheses are tested simultaneously to identify differentially expressed small RNAs. When many statistical tests are performed, some will yield significant results purely by chance. If each test is conducted at a significance level α, then among m tests, the expected number of false positives is:

$$\text{Expected false positives} = \alpha \cdot m \tag{1.11}$$

For example, with $\alpha = 0.05$ and $m = 1000$, one would expect about 50 false positives. This highlights the need for multiple testing correction to avoid interpreting random noise as biological signal. To address this, it is common to control the **False Discovery Rate (FDR)**—the expected proportion of false positives among all declared significant results. A widely used method is the **Benjamini-Hochberg (BH)** procedure (Benjamini et Hochberg, 1995). It works by ranking raw p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, and finding the largest k such that:

$$p_{(k)} \leq \frac{k}{m}\alpha \tag{1.12}$$

All hypotheses with ranks $i \leq k$ are then declared significant. The Benjamini-Hochberg procedure is implemented by default in widely used RNA-seq analysis tools such as `DESeq2` (Love *et al.*, 2014).

1.6.4 Hypergeometric Enrichment Analysis

The **hypergeometric** distribution models the probability of observing exactly k successes in n draws *without replacement* from a finite population of size N that contains K successes. The probability mass function (PMF) is given by (Devore *et al.*, 2012):

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1.13}$$

where N: Total number of elements in the background population; K: Number of successes in the population (e.g., genes annotated to a pathway); n: Number of elements in the selected subset (e.g., differentially expressed genes); and k: Number of observed successes in the subset (e.g., selected genes in the pathway).

In the context of biological **enrichment** analysis, the goal is to determine whether a given category (e.g., a biological pathway or functional gene set) is statistically **overrepresented** in the selected subset compared to what would be expected by random sampling. This is assessed using the cumulative probability of observing at least k successes:

$$P(X \geq k) = P(k) + P(k+1) + \cdots + P(\min(n, K)) \tag{1.14}$$

This right-tail probability serves as the enrichment p-value. A small p-value indicates that the observed overlap is unlikely to have occurred by chance, suggesting significant enrichment of the category in the selected subset. To assess **underrepresentation** (**depletion**), the left-tail cumulative probability is used instead:

$$P(X \leq k) = P(0) + P(1) + \cdots + P(k) \tag{1.15}$$

1.7 Machine Learning

Machine learning refers to algorithms that learn patterns from examples to make predictions or decisions based on new, unseen data. These methods have been applied across various fields, including healthcare, climate modeling, market analysis, and genomic research. In bioinformatics, machine learning is essential for analyzing complex biological data, uncovering patterns, and enhancing predictions that complement traditional statistical approaches. In RNA research, it has proven valuable for discovering relationships within large datasets, predicting biological functions, and identifying biomarkers, providing deeper insights into gene expression, RNA processing, and cellular responses.

The goal of a machine learning model is to find a function f that maps inputs to outputs, defined as:

$$f : X \to Y \tag{1.16}$$

where X is the input space, $x \in X$ is a feature vector; Y is the output space, and $y \in Y$ is a target label or value associated with x. In practice, we work with a finite dataset $S = \{(x_i, y_i)\}_{i=1}^{N}$ sampled from the input and output spaces. Machine learning methods are broadly categorized into supervised and unsupervised learning, depending on the availability of labeled data during training. The following sections describe these categories and the techniques applied in this work.

1.7.1 Supervised Learning

In **supervised learning**, each example in the dataset is labeled, meaning the corresponding output y_i is known for all i. The objective is to construct a model that accurately maps inputs to outputs and generalizes well to unseen data. Supervised learning encompasses two primary tasks: **regression** and **classification**. Regression focuses on predicting continuous numerical values and is often used to model quantitative relationships between biological features. Classification, on the other hand, involves assigning inputs to predefined categories. In bioinformatics, many problems are formulated as binary classification tasks, where the output Y consists of two classes, typically represented as $Y = \{0, 1\}$. Supervised learning algorithms rely on training datasets to learn the mapping between inputs and outputs. The model's performance is then evaluated based on its ability to accurately predict labels or values for previously unseen data. Common methods range from simple models like Linear Regression to more advanced ensemble techniques such as Random Forests and Gradient Boosting.

1.7.1.1 Linear Regression

Linear Regression is a supervised learning algorithm for predicting continuous numerical values by modeling a linear relationship between input features and an output variable. The model assumes that the relationship between the input vector $X = (X_1, X_2, \ldots, X_p)$ and the output Y can be described by a linear function of the form:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon \tag{1.17}$$

where β_0 is the intercept; β_j are the weights corresponding to each feature X_j ; ϵ is the error term, representing the difference between the predicted and actual values (Hastie *et al.*, 2009).

1.7.1.2 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees during training and aggregates their predictions to improve generalization. It achieves robustness by averaging the results of several weak learners. Formally, a Random Forest consists of T decision trees, where each tree f_t is trained on a randomly sampled subset of the data. The final prediction is obtained by combining the predictions of all individual trees. For classification, this is done by majority voting; for regression, by averaging.

$$f(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \tag{1.18}$$

where T is the total number of trees in the forest, and $f_t(x)$ is the prediction made by the t-th tree for input instance x (Hastie *et al.*, 2009).

1.7.1.3 Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds models sequentially, where each new model corrects the errors made by the previous one. While Random Forests build trees independently, Gradient Boosting constructs trees one at a time, with each new tree correcting the errors made by the previous model. The prediction function $f(x)$ at iteration t is given by:

$$f_t(x) = f_{t-1}(x) + \eta h_t(x) \tag{1.19}$$

where $f_{t-1}(x)$ is the prediction from the previous iteration; $h_t(x)$ is the new decision tree trained to approximate the residuals; and η is the learning rate, controlling the contribution of each tree (Hastie *et al.*, 2009). Gradient Boosting is useful for handling imbalanced datasets and improving predictive performance by focusing on hard-to-learn instances.

1.7.1.4 Rule-Based Learning

Rule-based learning involves designing a system that follows predefined rules to make predictions through a hard-coded, supervised process. It can be considered a hybrid approach to machine learning because it is constructed using a supervised process, operates automatically once created, and can integrate machine learning components as rules. This classification system can be represented as a function f that maps inputs X to outputs Y, where X is the input space (e.g., sequence features such as read count and structure); Y is the output space (e.g., Valid miRNA or Invalid miRNA). The rules are manually designed and refined through an iterative process to maximize performance according to evaluation metrics. The rules are typically expressed as conditional statements:

$$\text{If } A \implies B, \text{ where } A \text{ is the condition and } B \text{ is the resulting action.} \tag{1.20}$$

For instance, if the system classifies sequences by read count and structure, a rule can be formulated as:

$$\text{If } \{\text{Count} > \theta\} \text{ and } \{\text{Structure = Valid}\} \implies \text{Valid miRNA} \tag{1.21}$$

In this formulation, the conditional statements are explicitly defined and act as decision boundaries analogous to those learned by machine learning models.

1.7.2 Classification Evaluation Metrics

To evaluate the performance of a binary classification model, we rely on a set of metrics derived from the *confusion matrix*, which summarizes prediction results as follows (Table 1.1):

Table 1.1: Confusion matrix for binary classification

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- True Positive (TP): the number of correctly predicted positive cases.
- False Positive (FP): the number of negative cases incorrectly classified as positive.
- True Negative (TN): the number of correctly predicted negative cases.
- False Negative (FN): the number of positive cases incorrectly classified as negative.

Using the values from the confusion matrix, the following evaluation metrics are defined as what is used in Leclercq *et al*. (2013):

- **Accuracy** measures the overall proportion of correctly classified instances. *Range: 0 (worst) to 1 (perfect)*.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.22}$$

- **Precision** measures the proportion of true positives among all predicted positives. *Range: 0 to 1*.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1.23}$$

- **Recall** (also called Sensitivity or True Positive Rate) measures the proportion of actual positives correctly identified. *Range: 0 to 1*:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{1.24}$$

- **Specificity** (True Negative Rate) quantifies the proportion of actual negatives correctly identified. *Range: 0 to 1*:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{1.25}$$

- **F1 Score** is the harmonic mean of precision and recall. *Range: 0 to 1*:

$$\text{F1 Score} = \frac{2TP}{2TP + FP + FN} \tag{1.26}$$

- **Matthews Correlation Coefficient (MCC)** evaluates performance under class imbalance and considers all confusion matrix components. *Range:* -1 *(inverse prediction),* 0 *(random), to* $+1$ *(perfect)*:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1.27}$$

1.7.3 Selecting Supervised Models for Bioinformatics

For bioinformatics applications, such as the classification of small RNAs, it is common practice to systematically evaluate multiple machine learning algorithms and select the model that yields the best performance. This process involves:

- **Applying Algorithms**: A variety of models are applied depending on the task. For classification, common algorithms include Decision Trees, Support Vector Machines (SVM), Random Forests, Logistic Regression, and ensemble methods. For regression, frequently used models include Linear Regression, Polynomial Regression, and Lasso Regression. Ensemble learning, which combines the predictions of multiple models, is particularly effective for improving robustness and predictive accuracy. Techniques such as Random Forests and Gradient Boosting are especially useful for high-dimensional data, such as small RNA classification, due to their ability to handle many features and reduce overfitting.

- **Evaluation Metrics**: Model performance is assessed using appropriate metrics. For classification, common metrics include accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). For regression, typical metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score (Coefficient of Determination) (Hastie *et al.*, 2009).

- **Cross-Validation**: To ensure generalizability and prevent overfitting, k-fold cross-validation is used. The dataset is partitioned into k subsets; in each iteration, the model is trained on $k - 1$ subsets and validated on the remaining one. The average performance across all k folds provides a robust estimate of the model's accuracy.

- **Hyperparameter Tuning**: The model's performance can be further optimized by tuning hyperparameters using grid search.

- **Final evaluation**: The best-performing model, as determined by the evaluation metrics, is selected for final testing on an independent dataset. Because biological datasets are often large and complex, considerations such as training time and model size are also important. Excessive build time can limit scalability, while large models may be difficult to deploy. Therefore, there is often a trade-off between model complexity, performance, and practical usability.

1.7.4 Unsupervised Learning

In contrast to supervised learning, **unsupervised learning** deals with unlabeled data, where only the inputs x_i are available. Its purpose is to discover patterns or meaningful structures within the data without predefined labels, often by transforming the input space X into a more compact or informative representation C. Unsupervised learning encompasses various algorithmic categories, including clustering, dimensionality reduction, and association rule mining. Clustering is particularly common in bioinformatics, where it is used to identify groups of related genes or samples. Dimensionality reduction methods, such as Principal Component Analysis (PCA), are useful for simplifying high-dimensional datasets and enhancing data interpretation.

1.7.4.1 Hierarchical Clustering

Hierarchical Clustering is an unsupervised learning technique that organizes data points into a hierarchy of nested clusters, represented by a tree-like structure called a *dendrogram*. This structure illustrates how clusters are gradually merged at different levels of similarity, providing a comprehensive view of the data's organization. The most common approach, known as Agglomerative (Bottom-Up) Clustering, begins by treating each data point as an individual cluster. The algorithm then iteratively merges the closest pairs of clusters based on a similarity measure until all data points are grouped into a single cluster (Hastie *et al.*, 2009). Hierarchical clustering is particularly useful for gene expression analysis, where it can reveal relationships among genes or samples without requiring the number of clusters to be specified in advance.

1.7.4.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that simplifies high-dimensional datasets by projecting them onto a new set of orthogonal variables called *principal components*. These components capture the maximum variance in the data, making PCA particularly effective

for visualization, noise reduction, and feature selection. Given a dataset X with n observations and p features, PCA identifies k components $Z = (Z_1, Z_2, \ldots, Z_k)$, where $k \leq p$, such that each component is a linear combination of the original features:

$$Z_j = \sum_{i=1}^{p} w_{ji} X_i \tag{1.28}$$

Where w_{ji} represents the weight of the i-th feature contributing to the j-th principal component. Each principal component represents a new axis capturing a distinct direction of variance in the data. The first component captures the largest variance, the second captures the next largest orthogonal variance, and so on (Hastie *et al.*, 2009). By projecting data onto the first few principal components, PCA allows researchers to simplify complex datasets while preserving most of their variability.

1.7.5 Selecting Unsupervised Methods for Bioinformatics

The choice of unsupervised learning methods in bioinformatics is often guided by established practices in the literature. Specific techniques have proven effective for particular tasks:

- **K-means** clustering is efficient and simple, and therefore is often applied to large-scale clustering problems, especially when the number of clusters is known (Lebatteux *et al.*, 2019; Moeckel *et al.*, 2024).
- **Hierarchical Clustering** is commonly applied to gene expression analysis for identifying co-expressed genes.
- **Principal Component Analysis (PCA)** is frequently used for dimensionality reduction and visualization of high-dimensional data, aiding in the interpretation of complex datasets.

1.8 Roadmap to Small RNA Biomarker Discovery

Biomarkers are measurable indicators of phenotypic states, including normal and pathological biological processes. The discovery of biomarkers is an iterative process that involves refining candidate markers through multiple rounds of computational and experimental validation, incorporating approaches such as RNA sequencing, gene expression analysis, and predictive modeling to identify the most effective set. These biomarkers can exist as single-molecule indicators or as composite signatures derived from multiple RNA

molecules. In some cases, composite signatures demonstrate superior performance in phenotype classification compared to individual markers (Martens-Uzunova *et al.*, 2012; Bej et Basak, 2014; Islam *et al.*, 2017; Lu et Thum, 2019; Bajan et Hutvagner, 2020). Small RNA biomarkers with strong diagnostic power hold significant potential as therapeutic agents. Those identified through robust validation processes may progress toward therapeutic applications, where RNA-based drugs can be designed to target specific transcripts and modulate gene expression. This potential has led to the development of RNA-based therapies for a wide range of conditions, from viral infections to congenital disorders to complex polygenic diseases such as cancer (Bajan et Hutvagner, 2020).

The roadmap to sRNA biomarker discovery provides a structured framework for this process (Sun *et al.*, 2014; Buschmann *et al.*, 2016; Koch *et al.*, 2018; Roy *et al.*, 2018; Yang *et al.*, 2020; Batut *et al.*, 2021; Pola-Sánchez *et al.*, 2024). This roadmap, illustrated in Figure 1.5, consists of seven key steps and integrates computational and experimental approaches to refine candidate biomarkers with diagnostic, prognostic, or therapeutic significance. The following sections explore each step in detail.

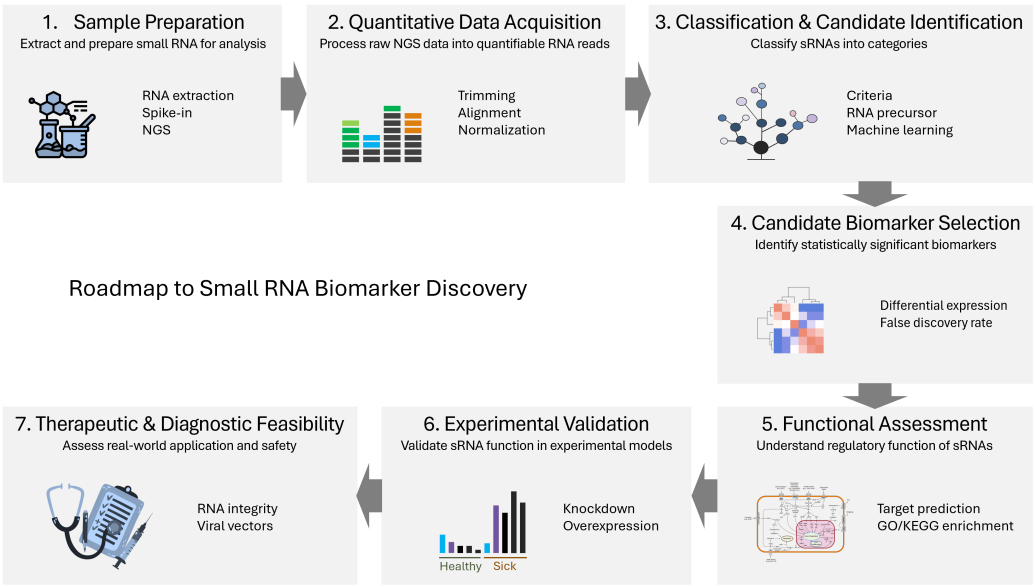


Figure 1.5: Small RNA biomarker discovery involves an iterative process of several steps: (1) Sample Preparation, (2) Quantitative Data Acquisition, (3) Classification and Candidate Identification, (4) Candidate Biomarker Selection, (5) Functional Assessment, (6) Experimental Validation, and (7) Therapeutic and Diagnostic Feasibility. Summarized from Sun *et al.* (2014); Buschmann *et al.* (2016); Koch *et al.* (2018); Roy *et al.* (2018); Yang *et al.* (2020); Batut *et al.* (2021); Pola-Sánchez *et al.* (2024)

Step 1: Sample Preparation. Extract and prepare small RNA from samples for sequencing or targeted detection. Key Techniques: RNA extraction, sequencing (NGS sRNA-seq), spike-in controls.

Step 2: Quantitative Data Acquisition. Process sequencing data to obtain high-quality small RNA reads for analysis. Key Techniques: Trimming, alignment, filtering, normalization.

Step 3: Classification and Candidate Identification. Classify small RNAs into functional groups using computational and machine learning methods. Key Techniques: miRNA, siRNA, feature engineering for sRNA classification, rule-based classification, machine learning.

Step 4: Candidate Biomarker Selection. Identify statistically significant small RNA biomarkers with high reproducibility. Key Techniques: Differential expression, false discovery rate, biomarker extraction for phenotype classification.

Step 5: Functional Assessment. Assess the biological function and regulatory roles of candidate small RNAs. Key Techniques: Target prediction, gene regulatory networks, enrichment analysis.

Step 6: Experimental Validation. Validate small RNA functions in experimental models. Key Techniques: Knockdown, overexpression.

Step 7: Therapeutic and Diagnostic Feasibility. Evaluate biomarkers for clinical and field use. Key Techniques: RNA integrity, delivery systems (viral vectors), safety, efficacy, approval.

1.8.1 Step 1: Sample Preparation

The process begins with **collecting** samples, followed by **RNA extraction** and quality assessment. Next, libraries are prepared for detection or sequencing using various technologies, including sequencing and hybridization-based methods. Next-generation small RNA Sequencing (NGS sRNA-seq) enables the capture of RNA transcripts shorter than 50 nucleotides (nts) in a high-throughput manner, converting them into a structured dataset that records RNA sequences and their abundance. The output consists of formatted text files containing RNA **sequence data**. Each NGS sRNA-seq library is composed of several possible classes of sRNAs as illustrated in Figure 1.1. Additional specialized sequencing techniques include bulk RNA sequencing (RNA-seq), single-cell sequencing (scSeq), whole genome sequencing (WGS), whole exome

sequencing (WES), amplicon sequencing, chromatin immunoprecipitation sequencing (ChIP-seq), metagenomic sequencing, and long-read sequencing (LRS). Other expression-detection technologies offer targeted quantification of preselected small RNA panels. These include microarrays, bead-based detection systems, and quantitative RT-PCR (qRT-PCR), which are useful for focused studies on known small RNA species. Each method has distinct advantages depending on the research objective, sensitivity requirements, and throughput needs. Previous work on the challenges associated with this step has led to various techniques such as Illumina-based small RNA sequencing for library construction and capture (Morin *et al.*, 2008), the use of spike-in controls for normalization and quality control (Locati *et al.*, 2015), and the development of the FASTQ format to encode read sequences with quality scores (Cock *et al.*, 2010). This thesis primarily explores methodologies for biomarker discovery using NGS sRNA-seq data.

Key techniques:

- **Sample Types**: Biofluids (serum, plasma, urine), tissues, cell cultures.
- **RNA Extraction**.
- **Detection**: NGS, qRT-PCR, Microarray, Northern.
- **Demultiplex**: Assigning reads to individual samples based on barcodes.
- **Spike-in**: For facilitating normalization and quality control.
- **Output**: FASTQ (primary format).

1.8.2 Step 2: Quantitative Data Acquisition

To generate high-quality datasets, sequencing reads undergo preprocessing steps that clean and prepare the data for analysis. This begins with **trimming** adapter sequences and low-quality bases to retain high-confidence reads. After trimming, **alignment** is performed against reference genomes or transcriptomes using short-read or long-read aligners. General **filtering** eliminates unwanted sequences based on low complexity or by mapping to specialized databases. Finally, **normalization** is applied to correct for differences in sequencing depth and sample-specific biases, enabling accurate quantification of small RNA expression across large datasets. Previous work on this step generated widely used tools and resources such as `Cutadapt` for adapter trimming (Martin, 2011), `Bowtie` for short-read alignment (Langmead *et al.*, 2009), and the fully annotated wheat reference genome for mapping and analysis (Appels *et al.*, 2018).

Key techniques:

- **Input**: FASTQ from sequencing platforms such as Illumina (MiSeq, HiSeq), Nanopore, and PacBio.
- **Adapter Trimming and Quality Control**: `FastQC`, `Cutadapt`, `Trimmomatic`.
- **Alignment Methods**:
  - **Short Reads**: `Bowtie`, `Bowtie2`, `MAQ`, `HISAT2`.
  - **Long Reads**: `BWA-MEM`, `Minimap2`.
- **Alignment Against Specific Genome or Transcriptome Databases**:
  - **Reference Genome** of research interest: human, wheat, microbiome
  - **Transcriptome or Translated**: SILVA (rRNA), tRNAdb, Rfam (RNA families), EST (expressed sequence tag), UniRef (Protein Reference Clusters)
  - **Source**: Ensembl, RefSeq, GENCODE.
  - **Visualization**: UCSC Genome Browser, Integrative Genomics Viewer (IGV)
- **Bias Correction**: Remove batch effects.
- **Normalization**.

1.8.3 Step 3: Classification and Candidate Identification

Following sequence preprocessing, **classification** assigns small RNAs to functional groups based on sequence features, biogenesis pathways, and regulatory roles. Major small RNA classes such as miRNAs, siRNAs, piRNAs, and tRNA fragments can be annotated through database mapping or computational prediction. However, most existing computational tools primarily target miRNAs (Yang et Li, 2011; Agharbaoui *et al.*, 2015; Leclercq *et al.*, 2017; Zuo *et al.*, 2020), leaving siRNAs relatively underrepresented. Computational prediction methods fall into two main categories: **rule-based**, which apply expert-defined criteria (e.g., motifs, structures), and **machine learning**, which learn features from sequence composition, structural stability, and genomic context.

Key techniques:

- **Direct Mapping**: miRBase, piRNA cluster databases.

- **miRNA Predictors**: `miRDeep-P`, `ShortStack`, `miR-PREFeR`, `miRdup`, `MirLibSpark` (*this thesis*). Keep in mind that miRNA predictors focus largely on precursor assessments.
- **Phased siRNA Predictors**: `PHASIS`, `PhaseTank`, `ShortStack`, `siWalk` (*this thesis*).
- **Other sRNA Predictors**: `RepeatMasker` (repeat-derived sRNAs).
- **Precursor Assessment**: `RNAfold`, `Mfold`, `MIRcheck`.
- **Feature engineering**: Computing features- free energy, melting temperature, hairpin, loop, bulge, length, *k*-mer, GC content, CpG island, codon usage, SNP, indel, haplotype block, genotype, phenotype, coverage, abundance, distribution, repeat, splicing, coding potential, transposable element, isomer, methylation, TF-binding site, miRNA target site, ortholog, phase score, subcellular localization, molecular interaction, pathway.
- **Feature Selection**: Information Gain, Chi-Square Test, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Clustering.
- **Machine Learning Algorithms**.

1.8.4 Step 4: Candidate Biomarker Selection (Differential Profiling)

After classification, small RNA candidates must be **statistically** evaluated for their significance in distinguishing case and control conditions. This step involves differential expression analysis, multiple testing correction, and cross-validation across datasets to ensure reproducibility. To identify biologically relevant small RNAs, **differential expression** analysis compares small RNA abundance across experimental conditions and highlights candidates with significant fold changes (Love *et al.*, 2014). To control for false positives, **false discovery rate (FDR)** adjustment is applied, ensuring statistical confidence in candidate selection (Benjamini et Hochberg, 1995). To further enhance robustness, cross-validation across independent datasets evaluates the **reproducibility** of selected candidates, confirming their consistent differential expression across multiple studies.

Key techniques:

- **Differential Expression Analysis**: such as `DESeq2`.
- **FDR Correction**: Benjamini-Hochberg, Bonferroni correction, *q*-value approach.
- **Reproducibility**: Validation across independent datasets.

1.8.5 Step 5: Functional Assessment of Small RNA Biomarkers

After identifying candidate small RNA biomarkers, biological and functional evaluation is essential to understand their regulatory roles. This step involves target prediction, functional enrichment analysis, gene regulatory network construction, and evolutionary conservation assessment to determine how small RNAs influence gene expression, cellular pathways, and biological processes. **Target prediction** and interaction analysis identify potential mRNA targets through computational approaches, while degradome analysis detects direct cleavage events, providing experimental support for small RNA–mRNA interactions. **Functional enrichment** analysis leverages Gene Ontology (GO) (Consortium *et al.*, 2004) and KEGG pathways (Kanehisa *et al.*, 2017) to reveal enriched molecular functions, biological processes, and cellular components among small RNA-regulated genes. Gene **regulatory network** construction integrates target predictions with co-expression and protein-protein interaction data to uncover broader regulatory modules. **Evolutionary conservation** analysis assesses the functional relevance of candidate small RNAs across species by identifying conserved regulatory elements using sequence comparison, phylogenetic inference, and orthology detection. Previous work on this step has led to widely used tools such as `psRNATarget` (Dai et Zhao, 2011) and `miRanda` (John *et al.*, 2004) for target prediction, and `DIANA-miRPath` for pathway-level regulatory analysis (Vlachos *et al.*, 2015).

Key techniques:

- **miRNA Target Prediction**: `psRNATarget`, `miRanda`, `TargetScan`, `Diana-microT`, `TAPIR`.
- **miRNA-mRNA Interaction (Degradome)**: `CleaveLand`.
- **Annotation Databases**: Gene Ontology, KEGG pathways, Protein-Protein Interaction (PPI) networks.
- **Regulatory Interaction**: `DIANA-miRPath`.
- **Comparative Genomics**: `BLASTN`, `OrthoDB` (ortholog detection), `phyloP` (phylogenetic inference).
- **Gene Set Enrichment**: Over-representation analysis with the hypergeometric test identifies enriched or depleted biological categories in selected gene sets.

1.8.6 Step 6: Experimental Validation

Experimental validation is essential to confirm the regulatory roles and biological relevance of small RNA biomarkers identified through computational and functional analyses. This step employs *in vitro* and *in vivo* approaches to assess direct interactions, functional effects, and phenotypic consequences of small RNA

activity. ***In vitro*** validation focuses on testing direct small RNA–target interactions and their downstream molecular effects using biochemical and cellular assays. ***In vivo*** validation employs relevant biological models, including cell lines, plants, and animals, to investigate the physiological impact of small RNA modulation through gene **knockdown** and **overexpression** studies. These experimental techniques are fundamental in molecular biology and are comprehensively described in *Molecular Cloning* (Green et Sambrook, 2012).

Key techniques:

- **Model Systems**: Plant, animal, and cell line models for *in vitro* and *in vivo* validation.
- **Loss-of-Function**: Gene silencing using CRISPR or RNAi.
- **Gain-of-Function**: Overexpression via plasmid vectors carrying recombinant viral genomes.
- **Gene Expression Quantification**: qRT-PCR.
- **Validation of Target Interaction**: Luciferase reporter assays.

1.8.7 Step 7: Therapeutic and Diagnostic Feasibility

The clinical translation of small RNA biomarkers requires rigorous evaluation of their **safety** and **efficacy**. This step assesses the feasibility of these biomarkers for diagnostic and therapeutic applications, ensuring their robustness in real-world clinical settings. For diagnostic applications, biomarker panels must demonstrate high specificity and sensitivity through validation studies. For therapeutic applications, lipid nanoparticles (Kulkarni *et al.*, 2018), viral vectors, and exosome-based **delivery** systems are actively explored as delivery mechanisms of small RNAs to ensure efficient and targeted delivery while minimizing toxicity and immunogenicity. Regulatory approval requires aligning biomarker validation and therapeutic development with the standards set by relevant regulatory bodies (Health Canada, 2024). A critical challenge is **RNA integrity**. Several factors impact RNA integrity, such as degradation and rearrangement caused by oxidation, hydrolysis, storage conditions, and delivery methods. Researchers employ qRT-PCR (Green et Sambrook, 2012), Liquid Chromatography, and NGS to assess RNA quality, detect degradation patterns, and confirm fragment stability. Ensuring RNA integrity is essential for diagnostic biomarker validation, where degraded RNA can lead to inaccurate quantification, and for therapeutic applications, where compromised RNA stability reduces treatment efficacy.

Key techniques:

- **RNA Integrity**: NGS, Liquid Chromatography, qRT-PCR.
- **RNA Stability**: Chemical modifications.
- **sRNA Delivery Systems**: Lipid nanoparticles, viral vectors for human hosts (AAV, lentivirus, adenovirus), viral vectors for plant hosts (PVX, TMV, CaLCuV), exosome.
- **Immunogenicity**.
- **Regulatory Considerations**: Canadian Food Inspection Agency, Health Canada.

1.8.8 Representative Study on miRNA and siRNA Biomarker Discovery

Understanding the role of small RNAs in gene regulation requires a structured approach that integrates experimental design, data processing, and computational analysis. As outlined in previous sections, the discovery of miRNA and siRNA biomarkers typically follows a multi-step roadmap, including experimental setup, high-throughput sequencing, bioinformatics processing, functional annotation, and biological interpretation. To illustrate how such studies are carried out in practice, we highlight a representative example by Agharbaoui *et al.* (2015), which focused on the discovery of stress-responsive miRNAs in wheat. This study exemplifies the first five steps of small RNA biomarker discovery, from sample preparation to functional assessment, while stopping short of experimental validation and therapeutic feasibility. This limitation likely reflects the high cost, complexity, or methodological gaps associated with those final stages. Nevertheless, the work provides valuable insights into how small RNA biomarker pipelines are implemented in real biological research and underscores the early emphasis on miRNA-only analyses in plant systems, particularly in the context of adapting to abiotic stress under cold conditions.

Agharbaoui et al. (2015): miRNA in Abiotic Stress

The study titled "*An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress*" by Agharbaoui *et al.* (2015) employed a comprehensive methodology to identify miRNAs in hexaploid wheat (*Triticum aestivum*) associated with development and abiotic stress tolerance. The approach integrated experimental design, data processing, bioinformatics analysis, and functional annotation.

Experimental Design and Data Processing. Researchers selected three wheat genotypes with different abiotic stress tolerances. Plants were grown under controlled conditions and exposed to cold, salt, and

aluminum stresses during vegetative and reproductive stages. Tissues were harvested to capture small RNA expression. Total RNA was extracted, and small RNA libraries were constructed and sequenced, generating about 89 million raw reads across ten libraries. The raw sequencing reads underwent preprocessing: adapter removal with Cutadapt, quality filtering to exclude low-quality reads, and length selection for miRNA reads of 18-24 nucleotides. These processed reads were mapped using MAQ to a curated wheat EST database, clustering approximately 1.4 million wheat ESTs into 127,039 UniRef clusters for a well-annotated reference of wheat transcript sequences.

miRNA Prediction and Validation. The miRNA prediction pipeline involved a multi-step bioinformatics approach. A total of 168,834 small RNAs were identified from the mapped reads, from which 337,668 potential precursor miRNAs (pre-miRNAs) were extracted based on genomic context and secondary structure potential. Secondary structure prediction was performed using MiPred and HHMMiR, both assessing the likelihood of a sequence forming a hairpin structure characteristic of miRNA precursors. MiPred identified 17180 potential hairpins, while HHMMiR identified 39,144. MIRcheck was used to further validate pre-miRNAs by evaluating key structural features such as stem length, loop size, and the number of mismatches, ensuring the selection of biologically relevant miRNA candidates. MiRdup was employed to identify the locations of mature miRNAs in the predicted hairpins, using a machine learning approach that integrates sequence composition with structural features. This step reduced the number of candidate hairpins by 81%, refining the list to high-confidence miRNA candidates.

Differential Analysis. Expression profiles of the predicted miRNAs were analyzed across different stress conditions and developmental stages. Differential expression analysis revealed that 182 of 199 miRNAs (91%) were significantly regulated, with a fold change greater than two and an FDR below 0.05. Among these, 165 miRNAs responded to abiotic stresses such as cold, aluminum, and salt, while 99 were associated with developmental processes. Many showed condition-, tissue-, or genotype-specific expression patterns, highlighting their potential regulatory roles in stress adaptation and development.

Functional Annotation. For functional annotation, miRNA target prediction was performed using TAPIR, which facilitated the identification of potential miRNA-mediated regulatory networks in wheat. The predicted target genes were subjected to GO enrichment analysis to identify overrepresented biological processes, molecular functions, and cellular components, providing insights into the functional roles of the miRNAs in stress responses and developmental processes. The expression patterns of the predicted miR-

NAs were analyzed across different stress conditions and developmental stages, revealing differentially expressed miRNAs associated with specific environmental challenges.

This study identified 199 novel and 267 conserved miRNAs associated with abiotic stress responses and developmental stages in wheat. Their findings provided a foundation for understanding wheat miRNAs in stress regulation and development, enabling subsequent research. For example, Hajiahmadi *et al.* (2020) conducted a comprehensive analysis of fatty acid desaturase genes in wheat, Chen *et al.* (2022) performed a genome-wide identification and functional characterization of auxin response factor genes in eggplant, and Du *et al.* (2022) integrated metabolomic and transcriptomic analyses to study flavonoid accumulation in *Cyclocarya paliurus* leaves at different altitudes. All three studies cited Agharbaoui *et al.* for their integrative approach to understanding miRNA involvement in plant gene regulation, metabolic pathways, and stress adaptation. These studies highlight the broader impact of miRNA characterization on advancing crop improvement and stress resilience.

1.9 Milestones in Plant miRNA and siRNA Identification and Prediction

The advancements in miRNA and siRNA research have unfolded over time, beginning with their discoveries in animals and later in plants, followed by the development of computational tools that enabled their systematic identification and analysis. This timeline highlights key milestones in biological discoveries and bioinformatics innovations that have shaped plant miRNA and siRNA research, with important predictors in **bold**. Our laboratory has been at the forefront of computational miRNA prediction and analysis, contributing a series of miRNA classification algorithms and tools. Key contributions from our lab to plant miRNA research are underlined.

Foundational Discoveries (1993–2004)

- The first miRNA, lin-4, was discovered in the nematode *C. elegans* by Lee *et al.* (1993). This was the first evidence of non-coding RNA functioning as a gene regulator.
- Fire and Mello (1998) discovered siRNAs in *C. elegans* and demonstrated RNA interference (RNAi). This discovery earned them the 2006 Nobel Prize.
- Hamilton et Baulcombe (1999) discovered siRNAs in tobacco, demonstrating RNAi in plants.
- Plant miRNAs were first discovered by Reinhart *et al.* (2002).

- Homology-based methods relying on cross-species conservation were introduced by Bartel et Bartel (2003), becoming the standard approach but remaining unable to detect novel miRNAs.
- Jones-Rhoades et Bartel (2004) introduced **MIRcheck**, the first computational tool for plant miRNA identification. Their approach integrated sequence conservation across species with secondary structure prediction. However, the method was not fully automated.

Early Computational Automation (2005–2014)

- Phase scoring methods proposed by Howell *et al.* and Chen *et al.* (2007) laid the foundation for computational detection of phased siRNA biogenesis in plants.
- Tools such as **miRDeep-P** (Yang et Li, 2011), **ShortStack** (Axtell, 2013b), and **miR-PREFeR** (Lei et Sun, 2014) enabled (semi-)automated identification of known and novel plant miRNAs. **miRdup** (Leclercq *et al.*, 2013) introduced a machine learning approach for mature miRNA localization. While ML methods emerged, rule-based models remained dominant.
- Tools for genome-wide phasiRNA detection were developed, including **ShortStack** (Axtell, 2013b), **PhaseTank** (Guo *et al.*, 2014), and **PHASIS** (Kakrana *et al.*, 2017). They enabled *de novo* identification of PHAS loci, with varying support for miRNA trigger prediction.

Advances in Prediction and Application (2015–Present)

- An integrative approach combining sRNA-seq, multiple predictors, and functional annotation was used to identify the hexaploid wheat miRNAome, providing genome-wide insight into stress-responsive miRNAs (Agharbaoui *et al.*, 2015). WMP, a centralized wheat miRNA database, was developed for analysis and annotation (Remita *et al.*, 2016), and MirAncesTar, based on ancestral genome reconstruction, was introduced to improve miRNA target prediction (Leclercq *et al.*, 2017).
- Axtell et Meyers (2018) introduced revised criteria for plant miRNA annotation in the era of big data. **mirLibSpark** (Wu *et al.*, 2019) was introduced as a scalable Spark-based pipeline for multi-library miRNA profiling that incorporates community-defined criteria. `miRador` (Hammond *et al.*, 2023) also implements these criteria.
- Several AI-driven tools for *de novo* plant miRNA prediction have emerged, including `SRICATs` (Zhang *et al.*, 2024) and `AmiR-P3` (Ataei *et al.*, 2024).

- Patisiran (Adams *et al.*, 2018), the first FDA-approved small RNA therapeutic, and the miRNA-based diagnostic test **GASTROclear** (So *et al.*, 2021), which received FDA Breakthrough status, marked the transition of small RNA research into clinical application.

1.10 Evolving Criteria for Plant miRNA Prediction in the Big Data Era

Experiment-driven miRNA annotation achieves high accuracy because of the sharp observational skills of biologists who extract meaningful clues from experimental data. These insights, refined through years of empirical study, have been distilled into golden rules that define miRNA biogenesis and function. As sequencing technologies advance, integrating these foundational rules into computational frameworks helps distinguish genuine miRNAs from background noise in large-scale sequencing datasets. This approach enhances miRNA annotation, making it more precise, reproducible, and adaptable. The establishment of criteria for miRNA annotation has been critical for ensuring the accuracy and consistency of miRNA prediction. The guidelines developed in 2008 provided essential criteria for distinguishing true miRNAs based on their biogenesis, structural features, and expression patterns. As high-throughput sequencing technologies became widely adopted, these criteria were revised in 2018 to address new challenges, such as processing precision, strand bias, and the reproducibility of novel miRNA detection. These evolving criteria, described below, have shaped how miRNAs are annotated and validated, providing a robust framework for computational prediction tools.

In their 2008 paper, *Criteria for Annotation of Plant MicroRNAs*, Meyers *et al.* established comprehensive guidelines for miRNA annotation, emphasizing precise biogenesis and structural features. The defining feature of plant miRNAs is their origin from the precise excision of an approximately 21-nucleotide miRNA/miRNA* duplex from the stem of a single-stranded, stem-loop precursor. This process involves Dicer-like enzymes that recognize and process these hairpin structures.

The mature miRNA and its complementary strand (miRNA*) originate from opposite arms of the hairpin stem, forming a duplex with two-nucleotide 3' overhangs. Extensive base pairing between the miRNA and miRNA* is expected, typically with no more than four mismatches. The duplex should also exhibit minimal asymmetric bulges, usually one or two bases in size. The presence of both the miRNA and miRNA* in sRNA-seq data supports precise Dicer-like processing.

Beyond these primary criteria, ancillary factors further strengthen miRNA annotations. Phylogenetic conser-

vation of the miRNA sequence and the precursor's secondary structure across species suggests functional importance. However, a lack of conservation does not prevent a sequence from being classified as a miRNA. Experimental validation of miRNA-mediated cleavage through degradome sequencing or translational repression assays confirms the functional role of the miRNA, though identifying a target is not mandatory for miRNA annotation. Additional supporting evidence includes miRNA detection through hybridization techniques, further validating its expression.

With advancements in high-throughput sequencing, Axtell and Meyers revisited miRNA annotation guidelines in their 2018 paper, *"Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data"*. The updated criteria emphasized that miRNAs must originate from a single-stranded, stem-loop precursor that undergoes precise Dicer-like processing, producing a miRNA/miRNA* duplex with two-nucleotide 3' overhangs. The miRNA/miRNA* duplex may contain up to five mismatches, with no more than three in asymmetric bulges. A valid miRNA must also exhibit strand bias, meaning that the majority of reads mapped to the precursor correspond to the mature miRNA rather than its complementary strand or other degradation fragments. Mature miRNA length must be within 20 to 24 nucleotides, with exceptionally strong evidence required for 23- or 24-nucleotide miRNAs due to their rarity. The precursor should not exceed 300 nucleotides in length. The mature miRNA must be the dominant small RNA species derived from its precursor to ensure that an annotated miRNA is a true regulatory molecule rather than a degradation product, siRNA-like processing, or unrelated small RNA biogenesis. This dominance reflects specific and consistent biogenesis rather than random degradation. Processing precision is assessed using a 75% threshold, requiring that miRNA and miRNA* expression (including one-nucleotide variants) account for at least 75% of all small RNA reads mapped to the precursor region. This criterion confirms that miRNA biogenesis is driven by specific processing rather than stochastic RNA fragmentation. Additionally, novel miRNA annotations should be detected in at least two sRNA-seq libraries from multiple biological replicates or independent datasets to ensure reproducibility. Annotation of new miRNA loci based on homology is now provisional and requires validation through sRNA-seq. Experimental validation now relies entirely on high-throughput sRNA-seq, making RNA blot analysis optional.

A key update from the 2008 criteria is the removal of experimental validation as a mandatory requirement for miRNA annotation. While target validation remains strong supporting evidence, it is no longer essential due to the reliability of deep sequencing and computational approaches. Similarly, though useful, cross-species conservation is no longer required to distinguish functional miRNAs from background noise.

To maintain high confidence in miRNA annotation, miRBase applies additional validation criteria. A miRNA must be supported by at least 10 sRNA-seq reads aligning perfectly with both the mature miRNA and miRNA*. The predominant small RNA species from each arm of the hairpin precursor must form a duplex with a 0- to 4-nucleotide 3' overhang. At least 50% of reads mapping to each arm must have the same 5' end, ensuring precise Dicer processing. The precursor hairpin must have a folding free energy under 0.2 kcal/mol nucleotide, and at least 60% of nucleotides in the miRNA and miRNA* should be paired within the predicted hairpin. These stringent criteria help prevent the misclassification of endogenous siRNAs or degradation products as miRNAs.

Table 1.2: miRBase criteria for high-confidence miRNA annotations. Adapted from Axtell et Meyers (2018)

	Criteria
1	At least 10 sRNA-seq reads aligned with no mismatches to both the mature miRNA and the miRNA*
2	The most abundant reads from each arm of the hairpin precursor must pair in the miRNA/miRNA* duplex with a 0- to 4-nucleotide overhang at their 3' ends
3	At least 50% of reads mapping to each arm of the hairpin precursor must have the same 5' end
4	The predicted hairpin structure must have a folding free energy of $<$0.2 kcal/mol/nucleotide
5	At least 60% of the bases in the miRNA and miRNA* must be paired in the predicted hairpin structure

Table 1.3: Criteria for plant miRNA annotations. Adapted from Axtell et Meyers (2018)

	2008 Criteria	**2018 Criteria**
1	One or more miRNA/miRNA* duplexes with two-nucleotide 3' overhangs	Exclude secondary stems or large loops in the miRNA/miRNA* duplex; limit precursor length to 300 nucleotides
2	Confirmation of both the mature miRNA and its miRNA*	Disallow confirmation by blot; sRNA-seq only
3	miRNA/miRNA* duplex contains $\leq$4 mismatched bases	Up to five mismatched positions, only three of which are in asymmetric bulges
4	The duplex has at most one asymmetric bulge containing at most two bulged nucleotides	
5	$\geq$75% of reads from exact miRNA or miRNA*	Include one-nucleotide positional variants of miRNA and miRNA* when calculating precision
6	Replication suggested but not required	Required; novel annotations must meet all criteria in at least two sRNA-seq libraries (biological replicates)
7	Homologs, orthologs, and paralogs can be annotated without expression data if all criteria met in one species	Homology-based annotations should be noted as provisional, pending fulfillment of all criteria by sRNA-seq
8	miRNA length not an explicit consideration	No RNAs $<$20 nt or $>$24 nt should be annotated as miRNAs. Annotations of 23- or 24-nt miRNAs require strong evidence.

1.11 Computational Strategies for Plant miRNA Identification

Computational tools for plant miRNA identification have been instrumental in analyzing high-throughput sequencing data to discover novel miRNAs and annotate their regulatory roles. Among them, **miRDeep-P** (Yang et Li, 2011) remains a popular tool for plant miRNA classification, alongside **ShortStack** (Axtell, 2013b) and **miR-PREFeR** (Lei et Sun, 2014), as they integrate genome alignment, structural analysis, and expression criteria to enhance the sensitivity and accuracy of miRNA prediction. These tools process sRNA-seq data to predict miRNA precursors and their corresponding mature miRNAs, making them widely used for benchmarking new miRNA prediction pipelines. In contrast, **miRdup** (Leclercq *et al.*, 2013) serves a different purpose. Rather than identifying miRNAs from sequencing data, miRdup refines the localization of mature miRNAs within a precursor based solely on sequence and structural features. While it enhances the accuracy of mature miRNA localization within a precursor and aids in miRNA validation, it does not perform *de novo* miRNA classification, making it methodologically distinct from the other tools discussed here.

The following sections review these plant miRNA prediction tools and note their relevance to current research directions.

1.11.1 miRDeep-P

The classification of miRNAs relies on a combination of sequence alignment, structural analysis, and expression-based validation. A defining characteristic of miRNAs is their origin from precursor sequences that form stable hairpin structures. These precursors are processed by Dicer enzymes to generate mature miRNAs. The classification process involves identifying candidate precursor sequences, predicting their secondary structures, and confirming whether the resulting small RNAs follow canonical miRNA processing patterns, including the presence of a miRNA-miRNA* duplex.

`miRDeep-P` (Yang et Li, 2011), derived from the animal-based `miRDeep` algorithm, was developed to improve miRNA identification in plants. Plant miRNAs exhibit more variable miRNA and miRNA* positioning within their precursors than their animal counterparts, which follow predictable Drosha processing sites. The `miRDeep-P` pipeline consists of several computational steps, each integrating specialized bioinformatics software.

The workflow begins with preprocessing raw small RNA sequencing data. `Cutadapt` removes sequencing adapters, while `FastQC` filters out low-quality reads. Length selection ensures that only sequences between 18 and 24 nucleotides—characteristic of plant miRNAs—are retained. To improve computational efficiency, redundant reads are collapsed.

After preprocessing, small RNA reads are mapped to the reference genome using `Bowtie`. Sequences matching rRNA, tRNA, snoRNA, and snRNA are filtered out to minimize false positives. Potential miRNA precursors are identified by extracting genomic regions spanning 250 bp around the aligned small RNAs. These candidate precursors undergo secondary structure analysis with `RNAfold`, which predicts their folding patterns. Since hairpin loops are a defining feature of miRNA precursors, only sequences forming stable secondary structures are retained.

Once potential precursors are identified, `miRDeep-P` predicts mature miRNA sequences by selecting the most abundant small RNA read within the hairpin structure. It also detects the miRNA* sequence, the complementary strand generated during Dicer processing. Predicted miRNA-miRNA* pairs are examined to

confirm that they follow expected processing patterns.

A scoring system ranks miRNA candidates based on several factors. The scoring module computes confidence scores by integrating expression levels, hairpin stability derived from `RNAfold` free energy estimates, and sequence conservation using miRBase (Kozomara *et al.*, 2019) as a reference. This step distinguishes genuine miRNAs from degradation fragments or other small RNAs with similar properties. Low-confidence candidates are removed, leaving only high-confidence predictions.

The final result from `miRDeep-P` is an annotated list of predicted miRNAs. These are categorized as known miRNAs, which match existing entries in miRBase, or novel miRNAs, which meet miRNA biogenesis criteria but have not been documented before. The output includes detailed information on precursor sequences, predicted secondary structures, expression levels in reads per million, and confidence scores.

1.11.2 ShortStack

`ShortStack` (Axtell, 2013b) is a specialized tool for analyzing sRNA-seq data, offering efficient annotation of miRNA loci and, in earlier versions, siRNA loci. The pipeline begins with preprocessing, where raw sRNA-seq reads are processed by removing adapters using `Cutadapt`. Identical reads are consolidated into a single entry, reducing redundancy and improving computational speed by minimizing alignment time and memory usage. Reads are then aligned to the reference genome using `Bowtie`, allowing a specified number of mismatches.

A crucial aspect of `ShortStack` is its handling of multi-mapped reads, which align to multiple genomic locations. If a region has strong unique read expression, multi-mapped reads are more likely to originate there than from sites with little or no support. Instead of arbitrarily assigning or discarding them, `ShortStack` calculates weights based on the density of nearby uniquely mapped reads, ensuring probabilistic placement and improving assignment accuracy.

Following alignment, `ShortStack` identifies small RNA clusters by defining coverage islands where small RNA depth exceeds a threshold. These regions are extended and merged to form small RNA clusters. It annotates miRNAs by identifying loci where dominant mature miRNA sequences co-occur with their miRNA* counterparts. Secondary structure prediction is then applied to confirm that these loci meet the canonical hairpin criteria of miRNA precursors.

Earlier versions of ShortStack also included siRNA prediction, detecting regions enriched for siRNAs and scoring their phased production. However, since version 4.0, ShortStack has removed explicit siRNA prediction, shifting its focus to miRNA annotation and clustering. Users seeking siRNA prediction must now rely on alternative tools.

1.11.3 miR-PREFeR

miR-PREFeR (Lei et Sun, 2014) is a computational tool for accurate and efficient plant miRNA prediction using sRNA-seq data. It mitigates high false-positive rates and long processing times by incorporating expression patterns and adhering to established plant miRNA annotation criteria (as defined in 2014).

miR-PREFeR requires SAM files as inputs generated using an external aligner like Bowtie, containing sRNA-seq read alignments to a reference genome. An optional GFF3 file can be provided to exclude specific genomic regions, such as protein-coding genes or repetitive elements. miR-PREFeR scans genomic alignments for regions with significant small RNA accumulation, indicating potential miRNA loci. It evaluates candidate loci based on mature miRNA length, ensuring it falls within 20–24 nucleotides to match known plant miRNA profiles. The presence of the miRNA* sequence is required to confirm precursor duplex detection. Additionally, the tool examines whether the precursor follows canonical Dicer processing patterns, including the presence of a 2-nucleotide 3' overhang, a hallmark of proper miRNA maturation. These patterns help distinguish miRNA-derived reads from degradation products or other small RNA fragments.

Expression pattern analysis plays a critical role in miR-PREFeR's accuracy. When multiple sRNA-seq samples are available, the tool incorporates expression data to improve confidence in miRNA predictions. A candidate consistently detected across various samples is more likely to be a true miRNA. In contrast, sequences appearing sporadically or with inconsistent expression are more likely to be random RNA fragments or degradation products, leading to their exclusion from the final predictions.

Candidate regions undergo RNA secondary structure prediction using RNALfold from the ViennaRNA package, which is optimized for local structure detection. The default precursor length for folding is 100 nucleotides. RNALfold efficiently detects local stem-loop motifs characteristic of miRNA precursors, as opposed to RNAfold, which predicts global structures. RNAfold predicts the global minimum free energy (MFE) structure of an entire RNA sequence, making it ideal for analyzing shorter sequences where overall folding patterns are relevant. In contrast, RNALfold is optimized for local secondary structure prediction

in longer sequences. By applying a sliding window approach, `RNALfold` identifies stable structural motifs, such as the stem-loop structures characteristic of miRNA precursors, through local MFE calculations. This method enhances efficiency and accuracy in detecting specific structural elements within extensive genomic regions.

The final output of `miR-PREFeR` includes an HTML summary of predicted miRNAs, providing details such as the number of identified miRNAs, their mature sequence lengths, associated precursor sequences, and secondary structure predictions. The tool also generates detailed read mapping profiles and additional output files for downstream analysis, facilitating further validation and exploration of candidate miRNAs.

`miR-PREFeR` is designed with user-friendliness, broad compatibility across plant species, and efficient resource utilization in mind. Its ability to process multiple libraries, integrate expression-based evidence, and apply stringent annotation criteria makes it a valuable tool for large-scale plant miRNA research.

Although these tools (`miRDeep-P`, `ShortStack`, and `miR-PREFeR`) have different accuracies and scope, they share the same goal: to predict miRNAs that comply with the most updated criteria from sequencing data. Despite their varying approaches and strengths, they all face a common challenge — **empirical time complexity** (Figure 3.1). While some tools achieve faster runtime than others, they all have room for improvement. Particularly, our experience indicates that none of these tools efficiently handles the analysis of data derived from large genomes, where computational resources and processing time become critical concerns.

1.11.4 miRdup

Leclercq *et al*. (2013) developed `miRdup`, a computational tool that accurately predicts the location of mature miRNAs within precursor sequences (pre-miRNAs). Unlike other tools that directly predict miRNAs from sequencing data, `miRdup` predicts mature miRNAs from provided static RNA sequences, which limits its ability to process large-scale sequencing datasets. In addition to identifying pre-miRNAs, `miRdup` also pinpoints the exact positions of mature miRNAs within pre-miRNAs, facilitating their validation.

To achieve this, `miRdup` employs a random forest classifier, a powerful machine learning technique that constructs an ensemble of decision trees to improve predictive accuracy. The model can be trained on a diverse dataset comprising miRNAs from various species, such as mammals, plants, fish, arthropods, and

nematodes. This allows it to capture lineage-specific characteristics and enhances its applicability across different organisms. Its capacity to self-train on updated miRBase releases ensures adaptability to newly discovered miRNAs, making it a valuable tool for miRNA research.

The classifier employs a selected collection of 100 features, organized into several categories. **Sequence composition** features, such as nucleotide frequencies and k-mer distributions within the miRNA:miRNA* duplex, help distinguish mature miRNAs from other regions. **Secondary structure** features, including the number of base pairs, internal loop sizes, and bulge characteristics, reflect crucial structural attributes important for miRNA maturation and function. **Thermodynamic stability** characteristics include metrics such as the minimum free energy of folding and the stability of the miRNA:miRNA* duplex, as thermodynamic properties affect the formation and stability of functional miRNA structures.

`miRdup` effectively models the complex interplay between sequence and structural elements that define mature miRNA regions within pre-miRNAs. In validation tests, `miRdup` demonstrated high accuracy in predicting mature miRNA locations, marking a significant advancement in miRNA bioinformatics.

The development of `miRdup` serves as a valuable **inspiration** for our approach to **designing siRNA prediction software** using machine learning. The systematic extraction of RNA sequence and structural features, coupled with their integration into effective machine learning models, provides a strong foundation for our methodology. Notably, `miRdup`'s strategic approach to feature selection and model validation offers essential insights for enhancing predictive performance. While `miRdup` predicts mature miRNAs from static sequences, our approach aims to leverage sequencing data directly, addressing a critical limitation where the validation of siRNA precursors depends heavily on expression alignment patterns. Furthermore, our predictor will retain the capability to predict from static sequences—an important feature for designing effective siRNA vectors.

1.12 Current Requirements for Bioinformatics Tools

The rapid expansion of sequencing technologies has led to an unprecedented increase in biological data, necessitating bioinformatics tools that are deployable, scalable, and open. **Deployability** ensures that tools are easy to install and configure, and are portable across different environments through containerization and standardized software packaging. **Scalability** is critical for efficiently processing large datasets, requiring parallel computing and cloud-based solutions. Openness, fueled by community development and **open-**

source licensing, promotes transparency and reproducibility. Bioinformatics tools have evolved to incorporate parallel computing frameworks, containerization and software packaging, workflow automation, and open-source development to meet these modern demands. The following subsections explore these key areas, outlining strategies that enable bioinformatics tools to be effective and accessible in an era of large-scale sequencing analysis.

1.12.1 Scalable Computing

Traditional single-node processing is no longer sufficient for handling large bioinformatics datasets, necessitating scalable computing approaches that integrate additional resources. This scalability is often achieved through distributed computing, where computational tasks are divided across multiple nodes to improve efficiency, enhance fault tolerance, and accelerate processing. In genomic, transcriptomic, and multi-omics research, modern computing frameworks leverage parallel processing, cloud infrastructure, and optimized storage solutions to efficiently manage vast and complex datasets. Below are key computational approaches used to process structured biological data across various applications.

Cloud computing infrastructures support parallel computations, distributed data processing, and storage solutions, enabling large-scale bioinformatics analyses without requiring dedicated hardware. **Cloud platforms** such as AWS, Google Cloud, and Azure provide on-demand computing and storage, which allows researchers to process large-scale genomic data efficiently (Banimfreg, 2023). Compute Canada (Digital Research Alliance of Canada) provides high-performance computing (HPC), large storage, and cloud computing resources for Canadian researchers. Its infrastructure enables researchers to run thousands of parallel jobs simultaneously and grants extensive storage capacity across computing clusters such as Cedar, Graham, and Narval. Compute Canada also features a wide variety of pre-installed bioinformatics software, and offers access to popular tools such as BLAST and Scikit-learn. However, due to high demand, job submissions often encounter queue times, which affects the turnaround speed for computational tasks.

Optimized **storage solutions** play a crucial role in managing the vast and complex datasets generated in bioinformatics, ensuring efficient data retrieval, organization, and long-term accessibility. High-performance storage architectures, such as HDFS (Hadoop Distributed File System) (Borthakur *et al.*, 2008), enable rapid data access and scalable storage capacity, making them essential for high-throughput genomic analyses. Object storage solutions like Amazon S3 and Google Cloud Storage provide cost-effective and scalable storage for unstructured sequencing data, facilitating seamless integration with cloud computing environments.

Additionally, database management systems such as MongoDB optimize the storage and retrieval of large-scale variant annotations and metadata, enhancing query efficiency (Bradshaw *et al.*, 2019).

MapReduce processes data by dividing it into chunks, applying a transformation function independently to each, and then aggregating the results as key/value pairs. It distributes workloads across a large number of nodes in a cluster, enabling parallel processing and significantly speeding up computation for large-scale tasks. While the programming model is straightforward—requiring only the *Map* and *Reduce* abstractions—implementing complex tasks, especially those involving iterative processing, can be challenging. One example of its application is SeqPig, which enables scalable processing of sequencing reads within the Hadoop ecosystem (Schumacher *et al.*, 2014).

Apache Spark is an open-source, distributed computing framework designed for fast, scalable, and fault-tolerant data processing (Zaharia *et al.*, 2010). It extends the disk-based MapReduce model by introducing Resilient Distributed Datasets (RDDs), which enable in-memory computations to reduce latency and improve efficiency. Spark greatly outperforms disk-based frameworks by reducing data movement and duplication, making it ideal for iterative and interactive workloads. However, for one-pass batch jobs, disk-based frameworks like Hadoop MapReduce can still be competitive. Spark provides high-level programming interfaces in Java, Scala, and Python, making it accessible to many data scientists. Its efficiency has made it a popular choice for big data analytics. However, advanced optimizations—such as tuning partitions and managing memory—can be challenging for beginners. Additionally, while in-memory computation speeds up processing, it can lead to high memory consumption. For smaller-scale tasks, the overhead of Spark's distributed execution model may not be justified. Beyond performance, its dynamic Directed Acyclic Graph (DAG) scheduling and RDD chaining structure workflows into interdependent tasks, facilitating seamless task automation and efficient execution at scale. Furthermore, its fault-tolerant distributed processing enhances reliability, while the lazy execution model optimizes computation by reducing redundant data processing. We will describe Spark further in Chapter 3.

MPI (Message Passing Interface) is a standardized communication protocol for parallel computing that enables multiple processes running on distributed memory systems to exchange data efficiently across cluster nodes through explicit message passing. Unlike high-level abstraction frameworks, MPI provides low-level control over parallel processes, allowing for highly optimized performance. However, the need to explicitly manage communication and synchronization between processes also increases coding complexity. MPI is

widely used for computationally intensive tasks such as phylogenetic tree reconstruction and probabilistic model estimation. An example is MrBayes, which performs Bayesian phylogenetic inference (Ronquist *et al.*, 2012). However, MPI is less suitable for sequencing data annotation and classification, which involve complex data dependencies and benefit from more flexible parallelization strategies.

OpenMP is a shared-memory parallelization framework that enables multi-threading by distributing tasks across multiple CPU cores within a single node. It is highly efficient for computationally intensive tasks but is constrained by the memory available on a single machine. An example is BWA-MEM, which utilizes multi-threading to accelerate sequence alignment (Li, 2013).

GPU (Graphics Processing Unit) computing has become a key approach for accelerating deep learning applications in bioinformatics. By leveraging massive parallelism, GPUs excel at computationally demanding tasks such as protein structure prediction, as demonstrated by AlphaFold (Jumper *et al.*, 2021). However, GPUs have limited memory capacity and are inefficient at handling irregular memory access patterns, making them less suited for genomics workflows like RNA-seq and GWAS, which involve complex data dependencies and high memory demands.

1.12.2 Containerization and Software Packaging

Modern bioinformatics tools must be deployable across diverse computing environments. Containerization and software packaging address these challenges by encapsulating software dependencies, reducing installation complexity, and enabling cross-platform compatibility.

Containerization technologies such as Docker provide lightweight, portable environments that package bioinformatics tools and their dependencies. Traditional software installations are prone to version conflicts and system-specific issues, while containers provide consistent execution across different operating systems and infrastructures. **Docker**, for example, is widely used for development and cloud-based applications, allowing users to build and share containerized bioinformatics workflows (Merkel *et al.*, 2014). In high-performance computing (HPC) environments, **Apptainer** (formerly Singularity) offers a secure and user-level alternative that enables containerized execution without requiring administrative privileges. By encapsulating tools in self-contained environments, containerization facilitates workflow reproducibility, minimizes software conflicts, and simplifies deployment in large-scale computational pipelines.

Software packaging systems streamline the installation and management of bioinformatics software. Conda and **Bioconda** (Grüning *et al.*, 2018), an extension of Conda, offer a vast repository of pre-configured bioinformatics software, such as `FastQC` and `Cutadapt`. They simplify dependency management for computational workflows and allow users to install tools and dependencies in isolated environments. **pip install** distributes Python-based bioinformatics software, including pandas and scikit-learn. Many bioinformatics tools are hosted on **GitHub**, which serves as a central repository for community-driven software development, providing access to source code, version control, and issue tracking. These packaging approaches reduce the technical barriers associated with software installation.

1.12.3 Workflow Automation

A typical bioinformatics workflow consists of multiple sequential steps, such as quality control, sequence alignment, variant calling, and functional annotation. Manually executing these steps is inefficient and error-prone, making workflow automation essential for modern computational biology. Since bioinformatics research involves a vast number of specialized tools, it is advantageous to select the most suitable software for each step rather than relying on a single, all-in-one solution. For example, a variant-calling pipeline may integrate `FastQC` for quality assessment, Trimmomatic for adapter trimming, `HISAT2` for sequence alignment, and `GATK` for variant calling (DePristo *et al.*, 2011). Workflow automation ensures that these tools are executed in the correct order, reducing human error while maintaining pipeline reproducibility.

Bioinformatics workflows must be modular and adaptable since no single pipeline fits every project. The capacity to quickly build, modify, and execute workflows is crucial for contemporary bioinformatics research across different species, sequencing technologies, and dataset sizes. To achieve this, workflow management systems facilitate the construction and execution of bioinformatics pipelines by defining tasks in a structured and scalable manner while enforcing standardized execution procedures. For example, **Nextflow** (Di Tommaso *et al.*, 2017) enables efficient execution of workflows in cloud environments using a dataflow programming model, where computational tasks are organized as a Directed Acyclic Graph (DAG). This approach ensures scalability and reproducibility, making Nextflow well-suited for RNA-seq and metagenomics analyses.

Cromwell (Van der Auwera et O'Connor, 2020) executes and manages complex genomic pipelines across cloud environments. It tracks execution, detects failures, and efficiently resumes workflows using call-caching. Cromwell primarily runs pipelines written in the Workflow Description Language (WDL), ensuring

portability across platforms. An example is the `GATK` variant-calling workflow (McKenna *et al.*, 2010), which is implemented in WDL and executed using Cromwell.

Snakemake provides a rule-based approach, simplifying pipeline design while ensuring dependency management and parallel execution (Köster et Rahmann, 2012). It constructs a directed acyclic graph (DAG) to track dependencies and optimize execution. It has been used in bioinformatics workflows, such as for ATAC-seq analysis.

1.12.4 Open-Source Licensing

The success of bioinformatics software depends not only on its functionality but also on its accessibility and maintainability. Open-source licensing ensures that bioinformatics tools remain transparent, reproducible, and widely accessible, fostering collaborative development and continuous innovation. It promotes community engagement, where active user contributions drive feature enhancements, bug fixes, and improved interoperability with other software.

Many bioinformatics tools are released under open-source licenses, which generally fall into two categories. **Permissive licenses** (e.g., MIT, BSD, Apache 2.0) allow unrestricted use and modification, making them suitable for academic and industry applications that require integration into proprietary pipelines. **Copyleft licenses** (e.g., GPL, or AGPL if the software is used on a web platform) allow users to modify and distribute a work as long as modifications remain free and under the same terms, ensuring that companies cannot profit from it or claim the modifications as their own. The longevity of bioinformatics tools relies on active community participation, including platforms like **GitHub** for version control, bug reporting, and feature requests. Additionally, discussion forums (e.g., Stack Overflow) provide spaces where users share troubleshooting tips, best practices, and workflow optimizations, fostering collaboration and continuous improvement.

CHAPTER 2
RESEARCH PROBLEMS

The previous chapter established the biological foundations of small RNA biogenesis and outlined the computational principles underlying their identification from high-throughput sequencing data. These principles define structural and expression-based criteria for miRNA and siRNA analysis. However, their application to large and heterogeneous datasets exposes practical constraints. The scale of modern sRNA-seq libraries, the redundancy of plant genomes, and the complexity of biogenesis signatures introduce substantial computational and analytical challenges. This chapter translates these foundational observations into a structured research framework, articulating the key limitations and research questions that guide the methodological developments presented in the remainder of the thesis.

2.1 Motivation

2.1.1 Complexity of sRNA-seq Composition

NGS sRNA-seq technologies capture a diverse range of RNA molecules shorter than 50 nts (Figure 1.1). Small RNAs belong to the broader category of non-coding RNAs, distinct from protein-coding mRNAs, and originate from multiple pathways. They include RNAi effectors such as siRNAs, miRNAs, and piRNAs (PIWI-interacting RNAs), as well as small nucleolar RNAs (snoRNAs) and degradation products from mRNAs, tRNAs (transfer RNAs), and rRNAs (ribosomal RNAs). Identifying functionally relevant small RNAs within sequencing data is challenging due to the presence of a vast background of non-functional or transient RNA fragments. This complexity is further compounded by overlapping sequence characteristics and alignment patterns, making annotation highly dependent on contextual features. Among the various small RNA types captured in sRNA-seq, miRNAs and siRNAs receive particular attention due to their well-characterized roles as RNAi effectors in gene silencing regulation. Furthermore, their size distributions coincide with the most abundant sRNA-seq size class, making them the dominant and most extensively studied small RNA species in sequencing datasets (see Figure 2.1).

All small RNAs that participate in RNA interference, such as miRNAs, siRNAs, and piRNAs, function as RNAi effectors and can be grouped under the umbrella of *interfering RNAs*, often abbreviated as *siRNAs* or *iRNAs*. Note that while “siRNA” can be used broadly to refer to all interfering RNAs, it is also commonly used more

narrowly to denote small interfering RNAs other than miRNAs. RNA interference pathways, illustrated in Figure 2.2, depict the biogenesis of miRNAs and siRNAs (Hutvágner et Zamore, 2002). miRNA precursors are single-stranded RNA molecules transcribed from the genome, forming characteristic hairpin structures typically shorter than 300 nts. These precursors are processed by Dicer into miRNA/miRNA* duplexes, generally ranging from 21 to 24 nts in length. One strand of the duplex is preferentially loaded into the RNA-induced silencing complex (RISC), where it guides the complex to target complementary mRNA sequences, leading to transcript degradation or translational repression. In contrast, siRNAs originate from long double-stranded RNA precursors that are processed by Dicer into short duplexes. Like miRNAs, siRNAs are incorporated into RISC to mediate gene silencing, but their precursors differ in structure and regulatory context. The key distinction between miRNAs and siRNAs lies in their precursor origin and processing. miRNAs derive from structured hairpins, whereas siRNAs are produced from perfectly base-paired double-stranded RNA substrates. These subtle differences in biogenesis pathways contribute to the complexity of small RNA classification in sequencing data, and necessitate the integration of metadata and cross-references to distinguish functionally relevant small RNAs from background noise.

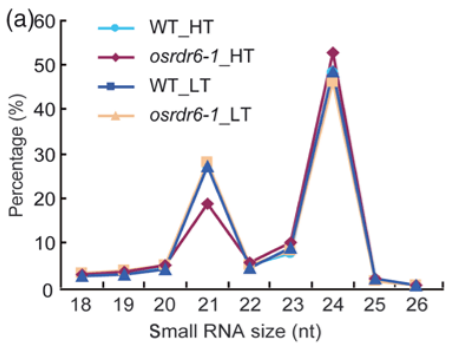


Figure 2.1: Size distribution of small RNAs in plants, represented by rice data. The most abundant small RNA size class is 24 nt, followed by 21 nt, corresponding to the predominant sizes of siRNAs and miRNAs, respectively. Adapted from Song *et al.* (2012).

Biologists classify small RNAs using rule-based methods, relying on expert knowledge of RNA biogenesis, precursor structures, sequence features, and alignment patterns. miRNAs are typically identified based on their structured precursors and Dicer-dependent processing into a single dominant mature strand, whereas siRNAs are recognized by their phased processing patterns from long double-stranded precursors. piRNAs are distinguished by their association with PIWI proteins and their tissue-specific expression profiles. Sev-

eral computational tools implement these predefined classification rules. For example, miRDeep-P (Yang et Li, 2011) predicts novel miRNAs in plants by leveraging miRNA biogenesis criteria, while BBCancer (Zuo *et al.*, 2020) provides an online platform for interrogating small RNA expression patterns by querying known databases.

siRNAs can be further classified into multiple subclasses, each defined by its genomic origin and functional role. These include trans-acting siRNAs (tasiRNAs), which regulate mRNAs at different loci; natural antisense siRNAs (natsiRNAs), which arise from overlapping sense-antisense transcripts; phased siRNAs (phasiRNAs), which are generated in periodic 21-nt or 24-nt intervals by sequential Dicer processing; and heterochromatic siRNAs (hc-siRNAs), which play roles in chromatin modification and epigenetic regulation. Figure 2.3 summarizes the nomenclature of interfering RNA classes based on their precursor structure, biogenesis mechanisms, and functional specificity.

Despite ongoing classification efforts, many small RNAs remain buried within sequencing data, as current classification frameworks fail to provide sufficient coverage for all RNA types. This limitation is particularly pronounced for siRNAs, which lack a standardized reference dataset and a widely adopted computational framework for annotation. While miRNAs have numerous prediction tools for various research scopes (Lukasik *et al.*, 2016), siRNAs are often overlooked in genome-wide studies due to the absence of curated datasets and the complexities associated with their classification. Additionally, regularly spaced alignment patterns, a hallmark of phased siRNAs, can resemble either miRNA or siRNA biogenesis pathways, making their annotation ambiguous (Figure 2.4). Although improvements in RNA expression profiling have been made (Agharbaoui *et al.*, 2015; Leclercq *et al.*, 2017), most bioinformatics pipelines remain optimized for miRNAs and a few well-characterized RNA classes, leaving the broader siRNA category largely unexplored.

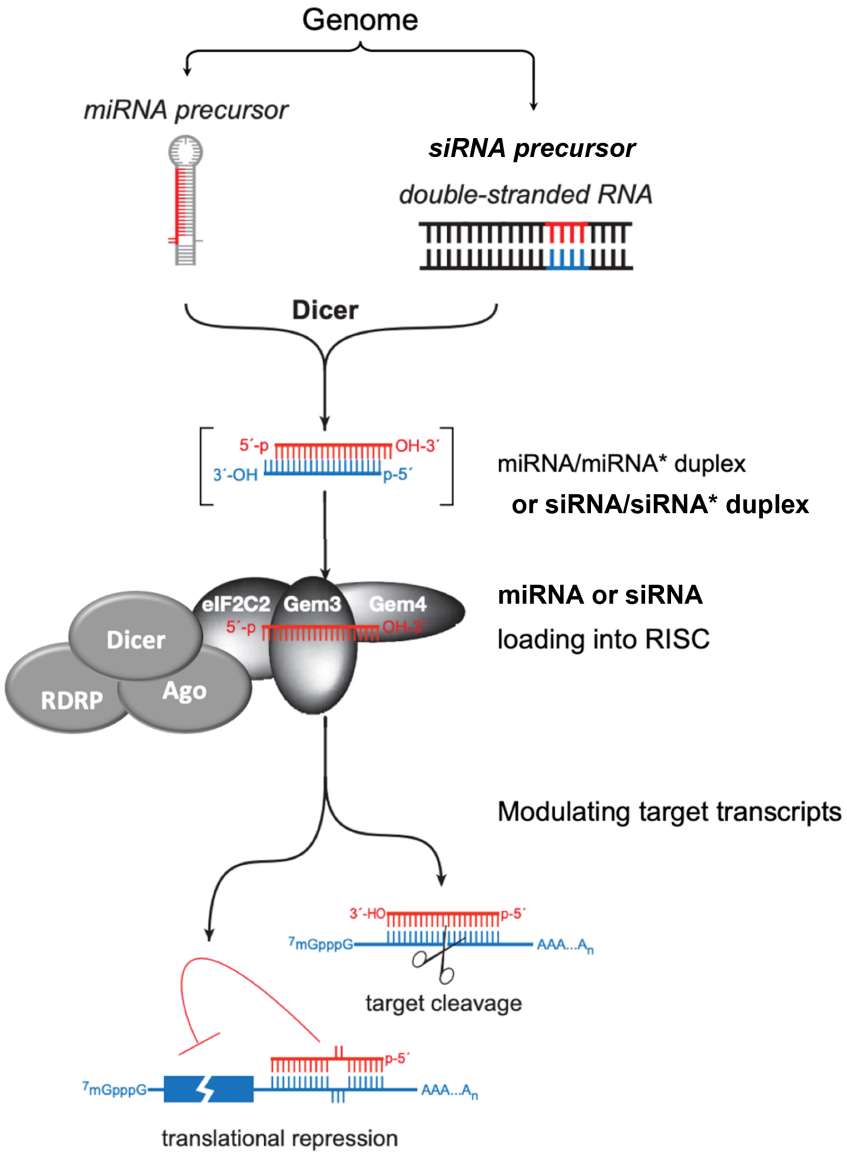


Figure 2.2: Schematic of RNA interference and miRNA/siRNA biogenesis. Dicer processes the precursor's double-stranded region into short RNA duplexes, from which one strand integrates into the RNA-induced silencing complex (RISC). The mature miRNA or siRNA then guides RISC to target mRNA, resulting in cleavage or translational repression. Adapted from Hutvágner et Zamore (2002).

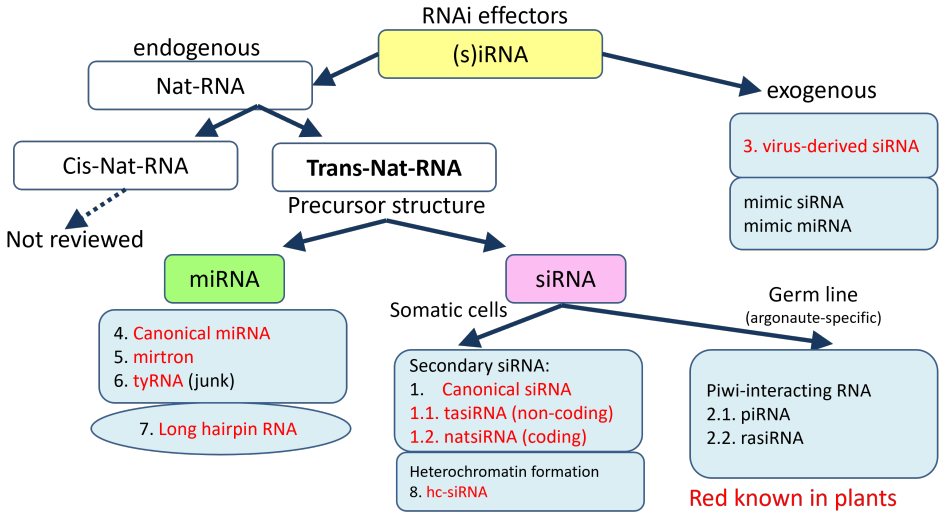


Figure 2.3: Nomenclature of siRNA classes. RNA interference (RNAi) effectors are classified based on their origin (endogenous or exogenous, coding or non-coding), mode of action (cis or trans), precursor structures, and tissue specificity. Subclasses highlighted in red are known to exist in plants. Summarized from Howell *et al.* (2007); Farazi *et al.* (2008); Billi *et al.* (2018); Kakrana *et al.* (2018); Wu *et al.* (2020).

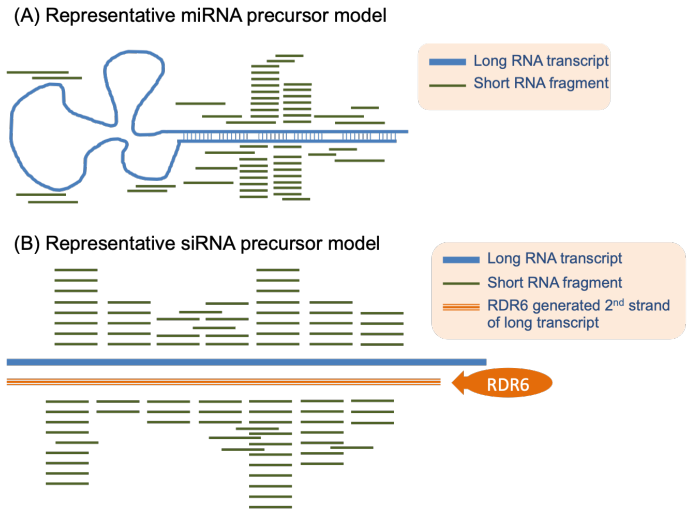


Figure 2.4: Small RNA alignments on representative miRNA and siRNA precursor models. miRNA precursors form hairpin structures, generating double-stranded regions through self-folding, while siRNA precursors arise from complementary transcripts. miRNAs undergo one or two cycles of Dicer processing, whereas siRNAs are processed periodically into multiple duplexes. The resulting small RNAs include the mature effector miRNA or siRNA, the co-product miRNA*, non-functional sibling siRNAs, and degradation fragments of the precursor. Despite distinct origins, their alignment patterns are highly similar, making classification challenging. In some pathways, siRNA second strands are synthesized by RNA-dependent RNA polymerase 6 (RDR6) (Howell *et al.*, 2007).

2.1.2 Big Data Are Getting Bigger

The rapid expansion of NGS and multi-omics studies has led to an unprecedented surge in biological data, necessitating more efficient computational frameworks for RNA analysis. Traditional bioinformatics pipelines, originally designed for smaller datasets, struggle to keep pace with the increasing volume and complexity of sequencing data. Existing single-machine solutions often encounter memory constraints and extended processing times, creating bottlenecks in miRNA prediction, annotation, and biomarker assessment. While algorithmic improvements in time and memory complexity have been beneficial, they are no longer sufficient to meet the growing demands of high-throughput sequencing.

Recent advances in miRNA annotation criteria emphasize computational classification over experimental validation, reinforcing the need for robust and scalable prediction tools (Axtell et Meyers, 2018). As datasets continue to grow, leveraging distributed computing frameworks has become essential to ensure efficiency and scalability. A distributed miRNA prediction system can allocate computational tasks across multiple nodes, optimizing performance without requiring expensive infrastructure. With the rise of cloud-based, pay-per-use services, even small research groups can now perform large-scale computational analyses that were previously limited to well-funded institutions.

Beyond efficiency, scalability accelerates scientific discovery. Parallelized computations dramatically reduce processing time, enabling rapid hypothesis testing and iterative refinement of miRNA predictions. What previously took weeks or months can now be completed in hours, facilitating new insights into gene regulation, biomarker discovery, and functional annotation across diverse species. This capability is particularly critical in precision medicine and agricultural biotechnology, where analyzing massive datasets is key to identifying biomarkers, predicting gene function, and designing targeted interventions.

For complex genomes such as wheat (*Triticum aestivum*), with its $\sim$13-gigabase genome and large sRNA-seq datasets (10-500 million reads per sample), computational efficiency is paramount. The classification of small RNAs involves multiple steps, including genome alignment, precursor annotation, structural prediction, differential expression analysis, and cross-referencing with known functional databases. Each step generates large intermediate datasets, requiring optimized memory usage, parallel data handling, and high-throughput processing to ensure timely and accurate results.

This challenge leads to the **research problem**: How can we enhance computational efficiency for process-

ing sequencing datasets that are high in volume, originate from large genomes, and encompass multiple libraries while ensuring accurate small RNA classification, particularly in species with complex, repetitive, and expansive genomes like wheat? Given these challenges, this thesis seeks to develop a scalable distributed miRNA predictor that integrates high-performance computing techniques to handle massive sequencing data efficiently. By implementing a distributed architecture, the proposed system will support multi-library integration, cross-sample normalization, and functional classification while reducing computational bottlenecks. This approach will enable researchers to analyze extensive sequencing datasets in a fraction of the time.

2.1.3 Disentangling the RNA Network: Studying miRNAs and siRNAs Side by Side

Wheat is not only a globally significant staple crop but also an excellent model for studying small RNA evolution due to its large and polyploid genome. Its hexaploid nature, with three homologous subgenomes, provides an unparalleled system for investigating how miRNAs and siRNAs evolve, function, and interact within a complex genetic background. The different subgenomes may retain, diversify, or lose small RNA regulatory pathways, offering insights into redundancy, functional divergence, and adaptive evolution. This makes wheat an exceptional model for studying small RNA evolution, extending beyond what has been learned from simpler diploid species like Arabidopsis.

Cold stress poses a significant agricultural challenge in Canada and other temperate regions, where wheat must withstand harsh winter conditions. Investigating the role of small RNAs in wheat's cold stress response not only has direct agricultural applications but also offers a broader understanding of how plants cope with environmental stresses. Cold-responsive miRNAs and siRNAs regulate gene expression by targeting transcription factors, metabolic pathways, and stress-related genes, orchestrating responses that enhance survival under low temperatures. Many of these small RNAs are also involved in other stress responses, such as drought, pathogen resistance, and oxidative stress, indicating that understanding cold stress responses could lead to broader applications in improving crop resilience to various environmental challenges.

Despite their importance, siRNAs are often overlooked in computational studies due to a lack of dedicated classification tools. As opposed to miRNAs, which are well-characterized and supported by numerous bioinformatics resources, siRNAs are frequently misclassified or remain unannotated. This gap limits the understanding of their regulatory roles and evolutionary significance. Existing siRNA prediction methods, such as phase score-based approaches, lay the groundwork for the systematic identification of siRNA. Therefore, it

is possible to incorporate these methods into a scalable computational framework alongside miRNA prediction, enabling a more comprehensive analysis of small RNA-mediated regulation. This thesis explores such an integration, facilitating large-scale sRNA-seq analysis while addressing a critical bottleneck in siRNA annotation and characterization.

Studying miRNAs and siRNAs simultaneously in wheat is particularly valuable because it enables the examination of their distinct but interconnected regulatory functions. While miRNAs typically regulate endogenous gene expression through translational repression, siRNAs function more broadly in defense, transposon silencing, and heterochromatin regulation. These pathways, however, are not entirely separate; an unexplored network of miRNA–siRNA coregulation likely plays a significant role in shaping gene expression outcomes. Investigating their interactions within a complex genome like wheat allows us to untangle this hidden regulatory network and uncover new layers of post-transcriptional gene regulation. Beyond revealing their cooperative regulatory dynamics, joint analysis of miRNAs and siRNAs also provides an opportunity to identify small RNAs that mark or influence phenotypic differences, particularly in adaptive traits such as cold tolerance. In such contexts, where gene regulation is multilayered, discovering small RNAs with biomarker potential is key to understanding and predicting phenotypic variation.

Advances in big data and sequencing technologies now make this research feasible on a previously unattainable scale. The ability to analyze large sRNA-seq datasets with high computational efficiency allows us to track small RNA expression changes across different environmental conditions, tissues, and developmental stages. Machine learning approaches further enhance our ability to classify and predict small RNA functions, providing a robust framework for extracting biologically meaningful insights from complex datasets.

This thesis aims to address the following **research problem**: How do small RNAs regulate plant stress responses, and what insights can their evolutionary patterns provide about plant adaptation to environmental challenges? We are driven by both practical and fundamental motivations. On a practical level, identifying cold-responsive miRNAs and siRNAs in wheat could directly contribute to breeding programs to develop more resilient crop varieties. Climate change is increasing the frequency and severity of extreme weather events, making it more urgent than ever to enhance crop tolerance to environmental stressors. Understanding small RNA-mediated regulation in wheat could inform strategies for engineering stress-tolerant plants, improve food security, and reduce agricultural losses. On a fundamental level, this research addresses key questions about the evolution and function of small RNAs, revealing how plants use RNA interference to

regulate gene expression and adapt to environmental conditions. It further investigates how small RNAs reflect phenotypic variation, offering a window into the regulatory architecture of complex traits and supporting their potential as biomarkers for plant adaptation.

2.1.4 Missing Piece in Small RNA Bioinformatics

As research progresses, updated versions of miRNA predictors continue to emerge, incorporating various biogenesis assumptions and implementation approaches. Currently, there are over 50 miRNA prediction tools (Lukasik *et al.*, 2016), but only three dedicated siRNA predictors (Kakrana *et al.*, 2017). Compared to miRNAs, which benefit from well-defined annotation frameworks and numerous prediction tools, siRNAs remain significantly more challenging to classify due to the absence of dedicated resources and standardized computational methods.

Currently, siRNAs are organized into several subclasses, including tasiRNAs, natsiRNAs, phasiRNAs, and hc-siRNAs (Figure 2.3). However, many of these subclasses remain poorly characterized, with limited functional validation beyond a few model species. Moreover, siRNAs often share sequence and alignment characteristics with miRNAs, making their computational identification highly ambiguous (Figure 2.4). These similarities can lead to misclassification, high false-positive rates, or the exclusion of functionally important siRNAs from genome-wide analyses.

Our research on small RNAs in wheat highlighted the indispensable role of siRNAs in adapting to cold stress, emphasizing their regulatory importance in plant stress responses (See wheat project in Chapter 5). However, it also revealed critical weaknesses in current siRNA classification methods. Existing tools lack robustness, often misclassify siRNAs as miRNAs, and offer limited insight into the functional characterization of siRNA subclasses such as tasiRNAs, phasiRNAs, natsiRNAs, and hc-siRNAs. Many of these subclasses are still poorly characterized beyond a few model species, leaving substantial gaps in our understanding of their regulatory functions.

The evolution of siRNAs seems closely tied to plant stress responses, as our findings in wheat indicate that stress-induced siRNAs may show species-specific or lineage-specific regulatory mechanisms. However, without a robust computational framework to classify and track siRNA evolution, it remains difficult to uncover their precise roles in gene regulation. Unlike miRNAs, siRNAs do not have widely adopted reference databases or standardized annotation pipelines, which makes genome-wide siRNA identification and evo-

lutionary analysis particularly challenging.

To address these gaps, there is a need to develop an enhanced siRNA classifier that integrates various sequence and structural features, rather than relying solely on alignment-based heuristics. Machine learning models present a promising approach to siRNA classification by leveraging large-scale sRNA-seq datasets, identifying novel siRNA subclasses, and decreasing dependence on predefined biogenesis rules, which are still limited and insufficient. Moreover, by discovering new, biologically relevant features for siRNA classification, computational models can help experts establish rational, data-informed classification criteria, which create a mutual learning loop between machine learning-based classification and expert-defined rules.

Developing a dedicated siRNA classification tool could have significant implications across multiple fields. In precision medicine, it may help identify siRNA-based therapeutic targets for genetic disorders, viral infections, and cancer. In agricultural biotechnology, it can support the creation of siRNA-driven strategies to improve crop resilience and disease resistance. In viral research, it could enhance the study of host-virus interactions by refining the classification of antiviral siRNAs. More broadly, it might advance fundamental RNA biology by revealing new regulatory roles of siRNAs in gene expression networks.

This leads to the next **research question**: How can we develop a computational model that accurately distinguishes siRNAs from other small RNAs while addressing their biogenesis complexity? The urgency of advancing siRNA bioinformatics cannot be overstated. As small RNA research continues to evolve, the absence of dedicated siRNA classification tools remains a bottleneck in the field. By developing a novel computational framework, we can bridge this gap and facilitate groundbreaking discoveries in gene regulation, plant biology, and therapeutic research.

2.1.5 Optimizing sRNA Delivery: Transferable Lessons from rAAV to Plant Viral Vectors

Delivering small RNA biomarkers into plants is a powerful strategy for regulating gene expression, enhancing stress tolerance, and improving crop resilience. Viral vector-based delivery, where plant viruses serve as carriers for small RNA precursors, enables precise gene silencing or pathway modulation using engineered miRNAs, shRNAs (short hairpin RNAs, synthetic), or siRNAs. However, much like viral vectors in human gene therapy, plant viruses exhibit inherently error-prone genome packaging processes, often leading to truncated, rearranged, or degraded RNA molecules. These errors compromise RNA functionality, highlighting the need for rigorous quality control measures to optimize vector design and ensure RNA integrity.

Valuable insights can be drawn from recombinant adeno-associated virus (rAAV)-based drug delivery systems to deliver small RNA-based treatments. However, the therapeutic efficacy of these approaches is highly dependent on the accuracy and stability of the packaged RNA, which is vulnerable to fragmentation, rearrangement, and unintended modifications. To mitigate these issues, NGS technologies—particularly long-read sequencing platforms such as PacBio and Nanopore—could be applied to evaluate RNA integrity, sequence fidelity, and structural stability within rAAV vectors. These analyses provide critical insights into how viral genome design influences RNA packaging efficiency and stability, leading to optimized vector architectures that enhance therapeutic delivery.

The same principles can be applied to plant viral vectors, which face similar challenges in maintaining stable and functional RNA expression. Viruses such as *Potato virus X* (PVX), *Tobacco mosaic virus* (TMV), and *Cabbage leaf curl virus* (CaLCuV) are increasingly used for small RNA delivery in functional studies and crop improvement. However, their efficiency is influenced by vector backbone composition, regulatory elements, and genome packaging fidelity. Adapting rAAV-inspired NGS-based quality control strategies to plant viral systems can help ensure that engineered small RNAs remain intact and functional throughout the delivery process.

This leads to the fourth **research question** in this thesis: How can we quantitatively assess the integrity of RNA therapeutics? Can high-throughput sequencing and computational pipelines be used to assess and enhance the quality of small RNA-based therapeutics in viral delivery systems? The ability to evaluate RNA integrity and optimize vector backbones is essential for precision agricultural biotechnology and improving crop resilience against environmental stressors.

2.2 Hypothesis

The **main hypothesis** of this thesis is that small RNA biomarker discovery and real-world application can be supported by computational methods tailored to the biological characteristics of small RNAs and the scale of modern sequencing data. This thesis focuses on plant interfering RNA biomarkers (including both miRNAs and siRNAs), addressing key computational challenges in their classification and functional interpretation. The thesis is built on four supporting hypotheses:

- **First**, integrating scalable computing frameworks significantly enhances the efficiency and accuracy of small RNA classification, especially when analyzing large, heterogeneous sequencing datasets from

complex genomes.

- **Second**, combining comparative genomics, transcriptomics, and the joint analysis of miRNAs and siRNAs from the same dataset provides a more comprehensive understanding of cold-responsive small RNA networks in wheat.
- **Third**, machine learning–based models that incorporate biogenesis-informed sequence and structural features can improve siRNA classification performance beyond traditional phase-score–based approaches.
- **Finally**, next-generation sequencing and advanced bioinformatics pipelines can be leveraged to establish robust quality control methods for evaluating RNA genome integrity in viral delivery systems such as rAAV.

2.3 Objectives of This Thesis

The **Global Objective** of this thesis is to *propose computational methodologies that advance the identification and evaluation of interfering RNA biomarkers (miRNAs and siRNAs)*.
These methodologies are designed to be biologically informed and scalable, addressing key challenges associated with small RNA biogenesis, data complexity, and real-world application. To achieve this objective, four sub-objectives have been formulated. Each specific objective is addressed through the development of a dedicated computational methodology, further detailed in the corresponding chapters of this thesis. In brief, the following specific objectives are pursued:

Specific Objective 1: *Develop a scalable and integrated pipeline for high-throughput miRNA prediction and annotation.* This objective addresses computational bottlenecks in large-scale small RNA analysis by introducing *mirLibSpark*, a distributed Spark-based pipeline that enables parallelized miRNA prediction, validation, and functional annotation. The pipeline is designed to efficiently process multi-library datasets across large and complex plant genomes while adhering to gold-standard miRNA annotation criteria. In addition to miRNA classification, it supports downstream tasks such as target gene prediction and differential expression profiling, forming a foundation for scalable small RNA transcriptomic analysis.

Specific Objective 2: *Develop a computational framework to identify small RNAs associated with phenotypic variation in wheat.* To uncover small RNAs associated with phenotypic variation in wheat, it is essential to move beyond fragmented analyses and jointly consider multiple RNA subtypes. This objective introduces a

unified computational workflow that performs integrated classification, annotation, and differential expression of both miRNAs and siRNAs from the same dataset. By combining comparative genomics, evolutionary conservation, and network-based functional prediction, the framework reveals co-regulatory mechanisms and evolutionary signatures underlying cold-responsive small RNAs in field-grown wheat. These analyses identify phenotype-associated small RNAs with regulatory relevance and biomarker potential.

Specific Objective 3: *Design a biologically informed siRNA prediction model using machine learning to classify siRNA-producing loci with improved robustness and accuracy.* This objective addresses the lack of generalized, biogenesis-aware siRNA prediction methods. The *siWalk* framework integrates structural, sequence, and phasing-related features into a supervised classification model. By benchmarking against traditional tools and validating on *Arabidopsis* datasets, the model aims to extend the capabilities of siRNA prediction by moving beyond conventional methods that rely on fixed phasing thresholds and alignment periodicity.

Specific Objective 4: *Construct a computational framework for assessing the putative effects of small RNA therapeutics delivered via viral vectors.* This objective introduces *ArcticFox*, a computational pipeline that leverages long-read sequencing data to classify and quantify genome structures in rAAV preparations. It detects truncated genomes, residual plasmid DNA, and impurities, while applying statistical corrections to address sequencing biases. The framework outputs standardized integrity metrics that support batch-to-batch comparisons and help ensure the safety and efficacy of RNA-based therapies.

A summary of the major research gaps and corresponding thesis contributions is provided in Table 2.1. Each research gap is mapped to specific steps of the RNA biomarker discovery process (as outlined in Section 1.8), clarifying how the thesis contributions address distinct stages of small RNA classification, characterization, and validation. With these challenges and objectives now established, the next chapter reviews foundational work relevant to this thesis across computational and biological dimensions. It examines the computational complexity of small RNA analysis pipelines, advances in scalable data processing through MapReduce-based techniques, principles of siRNA biogenesis and classification, limitations of current phase scoring strategies, and validation practices in computational small RNA research. Together, these perspectives provide the necessary context for the methodological innovations introduced in the following chapters.

Table 2.1: Summary of research gaps and corresponding thesis contributions

Research Gap	**Biomarker Discovery**	**Challenge**	**Methodology**	**Deliverable**
Big Data Are Getting Bigger	Steps 2–5	How can we efficiently process large and heterogeneous small RNA datasets across multiple plant genomes?	mirLibSpark	A scalable Spark-based pipeline for high-throughput miRNA discovery, validation, and annotation
Studying miRNAs and siRNAs Side by Side	Steps 1–5	How can we jointly analyze miRNAs and siRNAs to uncover co-regulatory roles and plant stress biomarkers?	Integrative framework	An integrative computational framework for joint expression profiling and functional analysis of small RNAs in wheat
Missing siRNA Prediction Bioinformatics	Steps 2–3	How can we improve the robustness and biological relevance of computational siRNA prediction?	siWalk	A predictive model based on siRNA biogenesis and machine learning for identifying phased loci
Optimizing sRNA Delivery	Step 7	How can we quantitatively assess the integrity of RNA therapeutics delivered via viral vectors?	ArcticFox	A long-read sequencing–based strategy for assessing rAAV RNA genome integrity

CHAPTER 3
LITERATURE REVIEW

This chapter reviews the computational, biological, and methodological foundations that inform the work presented in this thesis, emphasizing literature directly related to the core research problems. In particular, we examine computational inefficiencies in existing small RNA analysis pipelines, the absence of frameworks for integrated miRNA and siRNA prediction, limitations in the interpretability of current phase scoring methods, and the lack of high-resolution approaches for assessing RNA therapeutic integrity. Sections 3.1 through 3.3 address the computational inefficiencies of small RNA pipelines and trace the emergence of distributed computing frameworks such as MapReduce and Spark, which directly motivate the design and implementation of the scalable miRNA pipeline introduced in Chapter 4. Section 3.4 examines the functional and evolutionary overlap between miRNAs and siRNAs in wheat, establishing the biological rationale for the joint analysis approach developed in Chapter 5. Sections 3.5 and 3.6 explore siRNA biogenesis and phasing-based scoring methods, providing the theoretical and technical foundation for the machine learning approach for siRNA classification presented in Chapter 6. Section 3.7 presents a computational strategy for assessing viral genome integrity using long-read sequencing, which serves as the foundation for the contribution on RNA therapeutic quality control developed in Chapter 7. Finally, Section 3.8 outlines general validation strategies in computational small RNA research, many of which are incorporated throughout Chapters 4 to 7 to support the robustness, reproducibility, and biological relevance of the proposed methods. Taken together, these sections provide the conceptual and technical grounding for the new methodologies introduced in this thesis.

3.1 Computational Complexity of miRDeep-P as a Representative Pipeline for miRNA Prediction

Understanding the computational complexity and resource usage of a representative miRNA prediction pipeline like miRDeep-P provides valuable insights into the key challenges in large-scale small RNA analysis. By examining the algorithmic complexity, memory requirements, and empirical runtime, we can identify bottlenecks that affect scalability and efficiency. Each stage of the pipeline introduces distinct computational challenges, making it essential to break down their complexity to understand where optimizations can be applied. For clarity, Listing 3.1 summarizes the core workflow of miRDeep-P in pseudocode form, highlighting the main computational steps relevant to the complexity analysis. The computational demands of miRDeep-P depend on several key parameters, each contributing to the overall complexity of miRNA pre-

diction from sequencing data, described as follows:

- **N** (**Number of Input Reads**): 10 million to 500 million reads, with unique reads typically constituting 10% to 50% of the total reads.
- **m** (**Length of sRNA-seq Reads**): 15 to 50 nucleotides (nts).
- **M** (**Size of the Reference Genome**): 0.1 gigabytes (Gb) to 20 Gb. For example, *Arabidopsis thaliana* 0.135 Gb, wheat *Triticum aestivum* (wheat) 13 Gb.
- **L** (**Precursor Sequence Length**): 100 to 10,000 nts. To simplify the parameters, L can also represent the length of **mRNA**.
- **B** (**Size of the Reference Database**): Ranges from 38,000 entries (miRBase) to 100 million entries (NCBI nucleotide database, 1.3 trillion total bases, 50 Gb).

3.1.1 Algorithmic Complexity Breakdown

Preprocessing. Adapter trimming and filtering run in $O(Nm) \in O(N)$. Since m is a small constant, the complexity simplifies to $O(N)$. Reads are collapsed using hash-based indexing in $O(N)$, reducing N before further processing. Low-quality sequences, reads of undesired lengths, and low-abundance small RNA candidates are also filtered out and further reducing N. Finally, preprocessing runs in $O(N)$.

Genome Alignment. Bowtie (Langmead *et al.*, 2009) constructs an index for the reference genome in $O(M)$ time. Aligning N reads using Bowtie requires $O(Nm) \in O(N)$ operations. Thus, the overall complexity of Bowtie is $O(M + N)$. Further reductions in N are achieved by discarding unmapped sequences in $O(N)$. Finally, alignment runs in $O(M + N)$.

Precursor Identification. Extracting 250-bp flanking regions from both sides of each read has a complexity of $O(N)$, resulting in precursor sequences of length $L = m + 500$. RNAfold (Lorenz *et al.*, 2011) predicts secondary structures using a dynamic programming approach with a complexity of $O(NL^3)$. Since L is typically in the hundreds, its cubic complexity is substantial. To reduce computation, sequences without relevant secondary structures are discarded from N. The overall complexity of precursor identification is $O(NL^3)$.

Annotation. Database lookups for known miRNAs can have a complexity of $O(NB)$ for a naïve exact match search, or $O(N \log B)$ when allowing mismatches by utilizing sorted or indexed structures. Here, B rep-

resents the size of the miRNA database. In certain workflows, target prediction is also required. miRanda (John *et al.*, 2004) performs target screening with a complexity of $O(NBmL) \in O(NBL)$. Here, B represents the size of the mRNA database, with each mRNA having length L. Finally, cross-database annotation runs in $O(N \log B)$ or $O(NBL)$.

Output Generation. Output processing runs in $O(N)$ to generate a PDF file or static HTML interface, where N could have been reduced to $N' \approx 1000$. Finally, output runs in $O(1)$.

3.1.2 Worst-Case Complexity Analysis

The worst-case algorithmic complexity of miRDeep-P is determined by summing the highest complexity operations at each stage: Trimming, collapsing, and filtering reads $O(N)$, alignment $O(M + N)$, precursor identification $O(NL^3)$, Annotation via database lookup $O(N \log B)$, and this leads to an overall worst-case complexity of:

$$O(M + NL^3 + N \log B) \quad \text{or}$$
$$O(M + NL^3 + N \log B + NBL) \quad \text{if adding target screening.}$$

3.1.3 Memory Complexity Analysis

The memory complexity of miRDeep-P varies across different stages of the pipeline. Storing input sequences during preprocessing requires $O(N)$. Genome indexing with Bowtie incurs a memory requirement of $O(M)$. RNA secondary structure prediction with RNAfold requires $O(NL^2)$ due to the dynamic programming matrix storing pairwise interaction tables. Cross-database annotation requires $O(B)$. Thus, the total estimated memory consumption is:

$$O(N + M + NL^2 + B)$$

3.1.4 Optimization Strategies

The execution times of four miRNA analysis tools are shown in Figure 3.1. These measurements highlight substantial differences in runtime performance, stemming from variations in algorithmic design, computational efficiency, and implementation strategies. Among the tools, mirLibSpark is the fastest, completing the analysis in just 120 seconds due to its efficient parallelization and optimized indexing. In contrast, miRDeep-

P takes 74,601 seconds to complete the same task. This illustration is not intended as a direct tool-to-tool benchmark, but rather to highlight the computational bottleneck of miRDeep-P in its current form. Even on a relatively small *Arabidopsis* dataset, its runtime is prohibitively long. This limitation is magnified when working with larger genomes or deeper sequencing datasets, making large-scale miRNA analysis impractical without further optimization (Hammond *et al.*, 2023).

Bioinformatics pipelines are typically constructed by integrating multiple specialized *boutique* tools, each optimized for specific tasks such as read alignment, RNA secondary structure prediction, or differential expression analysis. Consequently, developers have limited control over the internal algorithmic complexity of individual boutique components. Instead, optimization efforts focus on efficiently orchestrating these tools and implementing strategic workflow modifications to reduce runtime and memory usage. Several practical approaches can be applied to improve efficiency when designing bioinformatics pipelines:

- Optimizing Input – Efficient preprocessing of input data reduces redundancy and minimizes unnecessary computations. This includes deduplicating reads, filtering out low-abundance sequences, and discarding low-quality data early in the pipeline. Applying filtering at multiple stages prevents the propagation of irrelevant sequences, ensuring that only high-confidence candidates proceed to the next steps. This approach significantly reduces memory usage and processing time in downstream stages.
- Selecting Computationally Efficient Alternatives – When multiple tools perform similar functions, selecting those with superior computational efficiency can significantly reduce runtime.
- Reordering Pipeline Steps to Minimize Bottlenecks – Structuring workflows so that the most computationally intensive steps, such as RNA secondary structure prediction, occur later in the pipeline, only after preliminary filtering. This prevents unnecessary processing of low-quality or redundant sequences, optimizing resource usage and reducing overall runtime.
- Parallel Execution of Independent Processes – Many steps in a bioinformatics pipeline can be parallelized to maximize resource utilization. For example, analyzing multiple samples or independent genomic regions concurrently can substantially reduce total runtime.
- Efficient File Handling and Intermediate Data Storage – Reducing unnecessary file I/O (Input/Output) operations and leveraging memory-efficient formats (e.g., BAM instead of SAM) can prevent slowdowns caused by excessive disk access and improve data processing efficiency.
- Adjusting Software Parameters for Performance Gains – Many bioinformatics tools provide config-

urable parameters that balance sensitivity and speed. Fine-tuning these settings based on dataset size and analysis objectives can reduce computational overhead without compromising biological insights.

- Resource-Aware Scheduling in High-Performance Computing (HPC) Environments – When running large-scale analyses on clusters, optimizing job scheduling, memory allocation, and resource distribution can significantly enhance performance and prevent bottlenecks due to inefficient resource utilization.

Rather than focusing on algorithmic modifications to individual tools, these strategies enhance the overall workflow by improving data flow, reducing redundant computations, and maximizing computational efficiency. For example, mirLibSpark implements several of these optimizations—parallel execution, indexing efficiency, and structured filtering—resulting in dramatically shorter execution times compared to pipelines lacking such improvements (Figure 3.1).

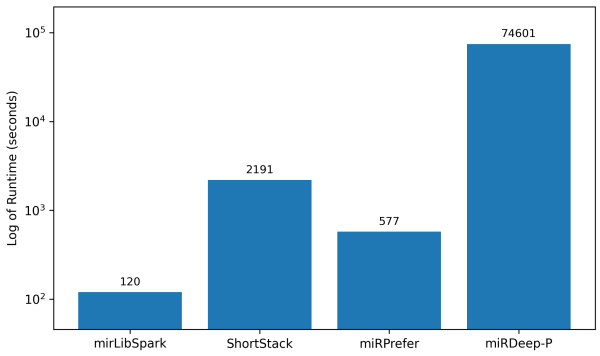


Figure 3.1: Execution time of miRNA prediction tools using an *Arabidopsis* library as input. `mirLibSpark` achieves the shortest runtime, while `miRDeep-P` exhibits significantly higher execution time, highlighting its computational inefficiency. This comparison underscores the impact of computational optimizations, such as parallelization and incremental indexing, on runtime performance. Data retrieved from (Wu *et al.*, 2019).

Listing 3.1: Generic miRNA analysis workflow extending miRDeep-P

```
Input:
  sRNA-seq reads
  genome
  miRNA_database
  mRNA_database
Output:
  Predicted small RNAs and optional functional annotations

1) Preprocessing
   reads_trimmed  <- cutadapt(reads)
   reads_filtered <- length_filter(reads_trimmed)
   unique_reads   <- collapse_identical_reads(reads_filtered)

2) Mapping
   alignments <- bowtie_align(unique_reads, genome)

3) Candidate evaluation (prediction core)
   loci <- derive_candidate_loci(alignments)
   candidates <- []
   for each locus in loci:
     precursor_seq <- extract_flanking_sequence(genome, locus)
     structure     <- RNAfold(precursor_seq)
     if passes_biogenesis_rules(structure, alignments at locus):
       candidates.append(locus)

4) Annotation
   known_matches <- lookup_known_miRNAs(candidates, miRNA_database)

   # The following steps are not part of miRDeep-P,
   # but common in complete miRNA annotation workflows
   targets    <- miRanda_predict_targets(candidates, mRNA_database)
   functions  <- integrate_functional_annotation(known_matches, targets)

5) Output
   report(candidates, metadata, functions)
   # metadata: genomic loci, precursor sequences,
               secondary structures, read frequencies, etc
```

3.2 Applying MapReduce to Accelerate Data Analysis

The computational demands of NGS data have driven the adoption of distributed computing frameworks to improve scalability and performance. Among these, MapReduce has played a foundational role in enabling parallel processing of large-scale bioinformatics pipelines. This section reviews the MapReduce model and its applications to small RNA analysis.

The introduction of **MapReduce** frameworks (Dean *et al.*, 2004) provides a structured way to apply the optimizations discussed previously—particularly parallel execution—to bioinformatics pipelines. Instead of modifying individual tools, MapReduce offers a generalized approach for breaking down complex analytical tasks into smaller, independent units that can be processed simultaneously across distributed nodes. This capability is especially valuable for large-scale sequencing data analysis, where computational efficiency and scalability are critical.

The MapReduce model achieves these optimizations through two fundamental operations: Map and Reduce (Figure 3.2, illustrating the MapReduce architecture). The **Map** operation divides large datasets into smaller chunks that can be processed independently across nodes. The **Reduce** operation aggregates intermediate results produced by the Map phase, consolidating outputs from multiple nodes into a unified dataset. This model enables distributed data processing across various stages of a bioinformatics pipeline, enhancing both speed and efficiency.

Several studies have successfully applied **MapReduce frameworks to bioinformatics tools** such as Hadoop-BAM, Crossbow, Myrna, CloudBurst, SeqPig, and GATK (Zou *et al.*, 2013; McKenna *et al.*, 2010; Kanellos *et al.*, 2014). These bioinformatics applications leverage MapReduce for tasks like sequence alignment, motif identification, and differential expression analysis. Moreover, some studies have applied MapReduce specifically to miRNA-related research. MR-microT, for instance, is a MapReduce framework designed for miRNA target prediction, distributing the computationally intensive task of finding miRNA-mRNA interactions across a Hadoop cluster. The framework handles the massive number of potential miRNA-target pairs by parallelizing sequence comparison tasks (Map phase) and aggregating results (Reduce phase).

A particularly relevant study by Pan *et al.* (2015) introduced a **hierarchical MapReduce framework** specifically designed **for miRNA prediction**. This approach distributes computational tasks across a Hadoop cluster, significantly improving speed and scalability compared to traditional tools like miRDeep. While not directly

comparable to miRDeep, this framework aims to detect miRNA candidates by parallelizing various stages of prediction, including genome alignment, candidate generation, filtering, and result selection. By integrating heavy third-party tools, such as BLAST and RNAfold, within a distributed computing environment, the framework drastically reduces computation time, enabling the processing of millions of sequences within a reasonable timeframe. This distributed architecture effectively utilizes increasingly available computational resources, making it feasible to handle large NGS datasets that would otherwise be processed too slowly.

The hierarchical MapReduce framework offers several **advantages** for miRNA prediction. Its scalability is achieved by distributing tasks across multiple nodes, allowing efficient management of large-scale NGS datasets. The modular layered approach provides flexibility to integrate additional modules or modify existing ones, accommodating evolving research requirements. Moreover, the parallel processing capabilities of MapReduce significantly reduce computational time, accelerating the miRNA prediction process and facilitating timely data analysis. However, despite its strengths, several **limitations** are inherent to the traditional MapReduce architecture. Disk-based processing introduces significant I/O bottlenecks, as intermediate results are frequently written to and read from disk storage, increasing computational overhead. Additionally, the rigid map-shuffle-reduce framework is designed for straightforward, single-pass processing, lacking the flexibility required for iterative computations and complex data dependencies.

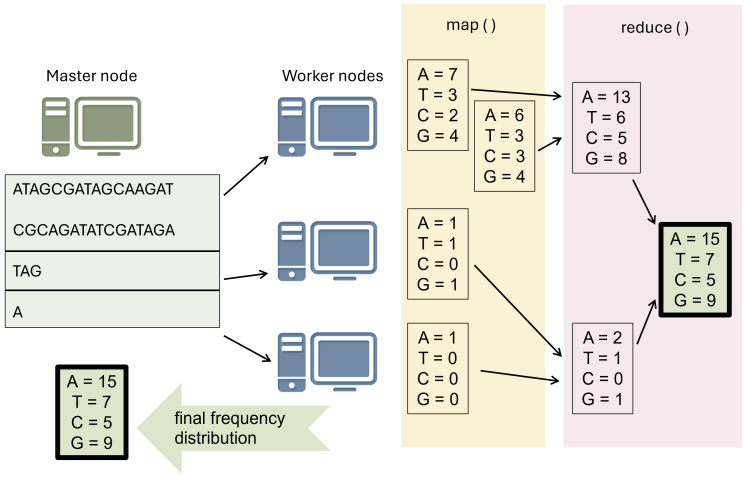


Figure 3.2: MapReduce-based word count for DNA sequences. This illustration demonstrates a MapReduce-based approach for counting nucleotide occurrences (A, T, C, G) in a DNA sequence dataset. The document, representing four DNA sequences as sentences, is split into three chunks for parallel processing by worker nodes. Each worker sequentially processes sentences within its assigned chunk, where subsequent sentences must wait until the preceding ones are fully processed. The mapped results, containing intermediate word counts, are then passed to the reduce phase, where they are grouped by nucleotide and aggregated to produce the final nucleotide frequency distribution, which is returned to the master node.

3.3 Optimizing with Apache Spark

Motivation for Moving Beyond MapReduce. The limitations of traditional MapReduce architectures, particularly disk-based processing and rigid workflow structures, necessitate more efficient frameworks capable of handling the computational demands of modern bioinformatics. **Apache Spark**, introduced by Zaharia *et al.* (2010, 2012), was designed as an alternative to Hadoop MapReduce to overcome high I/O latency caused by frequent disk reads and writes between computation steps. Its fundamental innovation lies in **in-memory** processing and optimized execution models, drastically reducing data processing time and making it an ideal choice for large-scale sequencing applications.

Spark quickly became one of the most active open-source projects after being donated to the Apache Software Foundation in 2013. Its capabilities in large-scale data processing were further demonstrated in 2016

when it achieved record-breaking performance by sorting a 100 terabytes benchmark dataset faster than previous systems (Spark News Archive, 2016). Spark's rapid adoption across industries is largely attributed to its ability to reduce execution times from hours to minutes compared to Hadoop's disk-based MapReduce framework.

Spark's Core Innovations: In-Memory Computation and RDDs. Spark offers three key advantages: **parallel execution**, **fault tolerance**, and **efficient memory management**. Its ability to support **iterative computations**—where multiple map phases are applied before shuffle and reduce—along with real-time streaming and scalable machine learning, makes it particularly well-suited for modern big data and AI-driven analysis. These capabilities enable Spark to process massive sequencing datasets with high performance, scalability, and reliability in sRNA research. Spark's architecture provides essential features for building efficient and scalable sequencing data analysis pipelines. At its core is the **Resilient Distributed Dataset** (**RDD**), a fault-tolerant, distributed collection of objects that enables parallel processing. As opposed to the disk-based storage model of MapReduce, Spark's RDDs reside in memory, significantly improving speed and efficiency. RDDs are partitioned across worker nodes, ensuring computations are both reproducible and recoverable. If a computation fails, Spark recomputes only the lost data partitions, avoiding costly restarts of entire workflows. Spark's ability to cache datasets in RAM using the `persist()` operation eliminates the need for redundant disk I/O operations, which is particularly valuable when multiple processing stages require access to the same dataset.

Transformations, Actions, and Shuffle Operations in Spark. Spark offers two primary methods for initiating RDDs: The `textFile()` function reads data from external storage systems in parallel, such as HDFS, cloud storage, or local files. The `parallelize()` function converts local data structures into RDDs for distributed processing. Once created, RDDs can be processed independently on each worker node. Spark's RDD operations can be broadly classified into **transformations**, **shuffle operations**, and **actions**. **Transformations**, similar to the Map phase of MapReduce, are operations that transform an existing RDD into a new one without immediately executing the computation. Examples of transformations include `map()`, which applies a function to each element of the RDD; `filter()`, which retains only elements satisfying a given condition; and `flatMap()`, which returns multiple outputs for each input element, commonly used for tokenization. **Shuffle operations** involve redistributing data across partitions, which can be costly due to network transfer and disk I/O. Key examples include `reduceByKey()`, which aggregates values locally before shuffling to reduce network traffic; `groupByKey()`, which collects all values for the same key and sends them to a single

partition; and `sortByKey()`, which sorts the dataset by keys across partitions. **Actions**, equivalent to the Reduce phase of MapReduce, are operations that trigger the execution of all preceding transformations to produce a final result or output. Common actions include `collect()`, which retrieves the entire RDD to the master node; `count()`, which returns the number of elements in the RDD; and `saveAsTextFile()`, which writes the RDD's contents to external storage.

Lazy Execution and DAG Optimization. Furthermore, Spark's **lazy execution** model plays a critical role in optimizing complex sequencing data pipelines. Transformations applied to RDDs are not executed until an action is triggered, allowing Spark to construct a Directed Acyclic Graph (DAG) that maps dependencies and determines the optimal order of execution. This approach minimizes unnecessary computations and reduces resource consumption, thereby enhancing efficiency. A typical sequencing data analysis pipeline involves applying transformations such as `map()`, `filter()`, `flatMap()`, `groupByKey()`, and `pipe()`. For instance, `pipe()` can be used to pass sequences individually to external alignment tools like Bowtie, leveraging their optimization for independent alignment processing. Additionally, Spark's dynamic task scheduling enables high-throughput processing, allowing tasks like alignment, annotation, and classification to be performed concurrently.

Handling Data Skew and Integration in Sequencing Pipelines. A major challenge in large-scale sequencing data analysis is **data skew**, where certain partitions contain significantly more data than others, resulting in imbalanced workloads. The most straightforward shuffle operation to mitigate data skew is the `repartition()` function, which redistributes data evenly across partitions. `repartition()` does not require entering a reduce phase, as it simply reshuffles the data without executing a specific aggregation function. However, it introduces additional shuffling overhead, which can impact performance if not used judiciously. Other methods for addressing data skew include `reduceByKey()`, which aggregates values locally before shuffling to minimize data transfer across the network. This approach reduces network traffic and computational overhead, making it generally more efficient than `groupByKey()` for large-scale operations. While `groupByKey()` offers greater flexibility by collecting all values for each key before aggregation, it requires more memory and is less efficient for large datasets. Additionally, the `union()` operation provides an efficient way to integrate datasets from different sources or pipeline stages, proving valuable when merging sequencing data from various analytical steps or multiple libraries.

Compatibility, Machine Learning Support, and Scalability. Spark's compatibility with diverse storage sys-

tems, such as HDFS, cloud storage (Amazon S3, Google Cloud Storage), and local filesystems, enhances its flexibility in handling sequencing data. Its support for multiple data formats, including JSON and CSV, enables seamless data ingestion and preprocessing. Furthermore, Spark's built-in machine learning library (MLlib) and graph processing framework (GraphX) provide powerful tools for advanced analysis. MLlib offers scalable solutions for tasks like classification, clustering, and regression, essential for miRNA annotation and biomarker discovery. GraphX enables network-based analysis, facilitating the study of gene regulatory networks and complex small RNA interactions.

Why Spark is Ideal for Sequencing Data Analysis Pipelines. Apache Spark offers substantial improvements over traditional MapReduce frameworks for sequencing data analysis due to its in-memory computation model, lazy execution, and dynamic task scheduling. By caching datasets in memory and performing local aggregations, Spark significantly reduces I/O bottlenecks, enabling faster processing of large-scale datasets. The lazy execution model allows Spark to optimize complex workflows by building execution plans only when an action is triggered, minimizing unnecessary computations and enhancing efficiency. Furthermore, Spark effectively addresses data skew through efficient shuffle operations and offers broad compatibility with various storage systems, ensuring seamless integration with existing bioinformatics infrastructures. Collectively, these features make Spark a powerful tool for building scalable and efficient sequencing data analysis pipelines, including predicting miRNAs or siRNAs.

3.4 Joint miRNA-siRNA Analysis in Wheat

To understand why joint analysis of miRNAs and siRNAs is essential for studying stress-responsive regulation in wheat, it is first necessary to examine the biological and evolutionary connections between these two major classes of small RNAs. Although they are often analyzed separately, miRNAs and siRNAs share overlapping origins, mechanisms, and regulatory roles that make their combined analysis both conceptually justified and methodologically desirable.

3.4.1 Functional and Evolutionary Overlap Between miRNAs and siRNAs

Among the RNAs detectable in NGS sRNA-seq, microRNAs (miRNAs) and small interfering RNAs (siRNAs) are two major classes of RNA effectors that function through RNA interference (RNAi) (Jones-Rhoades et Bartel, 2004; Doench *et al.*, 2003). These small RNAs are closely linked, both functionally and evolutionarily. They share key components of their biogenesis: both are processed by Dicer-like enzymes into duplexes,

loaded onto Argonaute proteins, and mediate gene regulation through mRNA degradation, translational repression, or chromatin modification. In addition to repression, RNAi pathways can also facilitate transcriptional activation under certain conditions (Catalanotto *et al.*, 2016; Place *et al.*, 2008). The evolutionary origins of miRNAs are thought to lie within the siRNA pathway. siRNAs initially evolved as a defense mechanism against transposable elements (TEs) and viruses, whereas plant miRNAs may have emerged from siRNA-generating loci that acquired structured precursors and more refined regulatory roles (Li *et al.*, 2011b). Notably, several studies have demonstrated that TE-derived loci can generate both siRNAs and miRNAs. For example, Piriyapongsa et Jordan (2008) identified genomic regions derived from miniature inverted-repeat transposable elements (MITEs) in *Arabidopsis* and rice that give rise to canonical miRNAs while simultaneously generating siRNAs, revealing the dual-coding potential of these transposable element loci (Figure 3.3). These loci challenge rigid classification and suggest that RNA biogenesis pathways are not always mutually exclusive. Such findings are further supported by evidence that some miRNA families, including MIR159 and MIR319, may have originated from phased stem-loop precursors with siRNA-like origins (Li *et al.*, 2011a). Functionally, these two classes also blur. While siRNAs are traditionally associated with silencing of TEs and repetitive regions, and miRNAs with post-transcriptional gene regulation, there are exceptions. Certain miRNAs, for instance, can target the 3' untranslated regions (3'-UTRs) and activate gene expression—behaviors typically attributed to siRNAs (Valinezhad Orang *et al.*, 2014). Both classes are also widely implicated in plant responses to environmental stress, where their regulatory roles often converge (Baldrich *et al.*, 2014; Guo *et al.*, 2015; Meijer *et al.*, 2022).

Given their overlapping features in origin, function, and regulatory targets, miRNAs and siRNAs should not be studied in isolation. Instead, there is a growing need for frameworks that support joint analysis. Treating them as separate entities imposes artificial boundaries that overlook shared processing mechanisms, dual-origin loci, and cooperative regulatory roles, especially under stress conditions. These findings underscore the need for computational approaches that enable simultaneous, *de novo* prediction of both miRNAs and siRNAs, allowing for a complete understanding of small RNA-mediated regulation in plants.

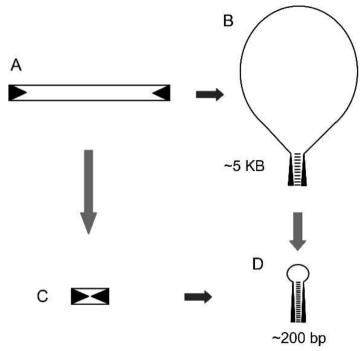


Figure 3.3: Model of TE-derived evolution of miRNA and siRNA precursors. The diagram illustrates how (A) a full-length autonomous DNA-type transposable element with a coding region flanked by terminal inverted repeats (TIRs) can give rise to small RNA precursors. When transcribed, the inverted repeats at both ends may pair to form a double-stranded RNA structure (A $\rightarrow$ B). While typical TIRs are short (10–30 nt), some TE-derived transcripts contain extended or internally complementary regions that support long foldbacks suitable for phased siRNA production. This makes B a prototype siRNA precursor. Internal deletions can yield compact, non-coding elements known as miniature inverted-repeat transposable elements (MITEs) (A $\rightarrow$ C). Through additional structural refinements, both B and C can give rise to *canonical miRNA precursors* (B $\rightarrow$ D and C $\rightarrow$ D), with D representing a typical stem-loop processed by the miRNA biogenesis pathway. This model supports the hypothesis that siRNA and miRNA precursors can arise from a shared pool of TE-derived sequences. Adapted from Piriyapongsa et Jordan (2008).

3.4.2 Wheat as a Model for iRNA Evolution and Stress Responses

As miRNAs and siRNAs play critical roles in plant stress responses (Khraiwesh *et al.*, 2012), their comprehensive identification and functional characterization are essential for understanding the regulatory networks that control environmental adaptation. Wheat (*Triticum aestivum L.*), one of the most widely cultivated crops worldwide, presents a uniquely powerful system for investigating small RNA evolution and function under stress. In particular, the availability of both cold-acclimating and non-acclimating wheat cultivars facilitates experimental comparisons that link small RNA profiles with physiological stress responses (Loukoianov *et al.*, 2005; Li *et al.*, 2018).

Bread wheat is a hexaploid species (2n = 6x = 42, AABBDD) whose genome was formed through a series of ancient hybridization and polyploidization events. According to phylogenomic analyses by Marcussen *et al.* (2014) (Figure 3.4), A and B genomes diverged 6.5 million years ago and later hybridized to form a

tetraploid ancestor (AABB) 0.8 million years ago. This tetraploid species subsequently hybridized with a D-genome donor 0.4 million years ago, giving rise to modern hexaploid wheat (AABBDD). This complex evolutionary history, which combines divergence, hybridization, and genome duplication, makes wheat an ideal model for studying small RNA evolution across homologous and homeologous loci. The presence of three related subgenomes enables comparative analysis of regulatory conservation and divergence, revealing how miRNAs and siRNAs evolve, diversify, or maintain function after genome merger.

Although siRNAs have historically received less attention than miRNAs, emerging research highlights their significant roles in wheat stress adaptation and genome regulation. For example, Yao *et al.* (2010) showed that non-miRNA small RNAs are upregulated under abiotic stress and contribute to mRNA degradation. Tang *et al.* (2012) identified a complex siRNA network involved in cold response. Sun *et al.* (2013) found widespread TE-associated siRNAs in developing grain, suggesting their role in genome stability and developmental regulation. Kenan-Eichler *et al.* (2011) demonstrated that hybridization and polyploidization can lead to siRNA deregulation, affecting gene expression patterns and environmental adaptability.

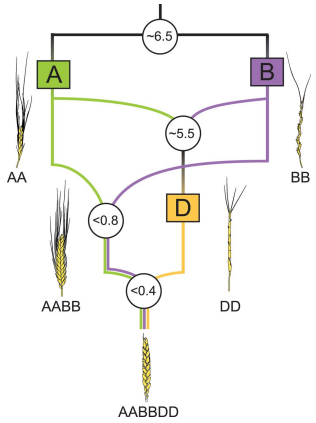


Figure 3.4: Hybridization and divergence events leading to bread wheat. This diagram illustrates the evolutionary history leading to hexaploid wheat (*Triticum aestivum*, AABBDD). The A and B genomes diverged approximately 6.5 million years ago and hybridized around 5.5 million years ago, giving rise to the D genome via homoploid hybrid speciation. A second hybridization between diploid AA and BB species formed tetraploid emmer wheat (AABB), followed by a third hybridization with a diploid DD genome donor to produce modern hexaploid wheat. Adapted from Marcussen *et al.* (2014).

3.4.3 Limitations of Existing Joint Small RNA Analyses

Building on the evolutionary and functional context of small RNAs in plants, we now examine how multiple RNA types coordinate under well-controlled experimental conditions. While the previous sections focused on wheat and the evolutionary convergence of miRNAs and siRNAs, this *Arabidopsis* case study (Tiwari *et al.*, 2020) illustrates how diverse small RNAs can be profiled jointly during cold acclimation. It offers practical insights into experimental design, analytical strategies, and current methodological gaps that motivate the work presented in this thesis.

Tiwari et al. (2020): miRNA, siRNA, and lncRNA in Cold Acclimation

Experimental Design, Sampling Strategy, and Data Processing. The study titled "*Identification of small RNAs during cold acclimation in Arabidopsis thaliana*" by Tiwari *et al*. (2020) investigated small RNAs in *Arabidopsis thaliana* during cold acclimation. Plants grew in long-day conditions (16-hour light/8-hour dark) at 22°C and were transferred to 4°C for cold stress. Samples were collected at 3 hours (early response), 6 hours (intermediate response), and 2 days (acclimated response), with non-treated plants as controls. Each condition had 3 biological replicates, resulting in 18 sequencing libraries. Total RNA was extracted for sRNA-seq, and libraries were sequenced on the Illumina HiSeq 2500 platform to generate 50 bp single-end reads. Data preprocessing involved trimming adapters with Cutadapt and discarding reads shorter than 18 or longer than 30 nucleotides. Read quality was assessed with FastQC.

Classification of Small RNA Types. To classify small RNAs, reads were mapped to the *Arabidopsis thaliana* genome (TAIR10) with Bowtie. miRNAs were identified by miRDeep-P2 and checked against miRBase v22. Natural antisense siRNAs (nat-siRNAs) were detected by aligning small RNAs to overlapping transcripts, while lncRNA-derived sRNAs were annotated using known lncRNA datasets.

Differential Expression and Visualization. DESeq2 was used for differential expression analysis of miRNAs, siRNAs, and lncRNA-derived sRNAs. Expression levels were normalized to reads per million. Small RNAs were selected based on a fold change of at least 2 and an adjusted p-value below 0.05. Hierarchical clustering and heatmaps visualized expression trends, and UpSet plots analyzed overlapping differentially expressed small RNAs across time points.

Functional Annotation and Experimental Validation. Functional annotation included miRNA target predic-

tion via psRNATarget, selecting targets by complementarity scores and energy thresholds. GO enrichment analysis revealed the biological functions of miRNA targets, and KEGG pathway analysis identified key regulatory networks like hormone signaling pathways (ABA, ethylene, auxin), oxidative stress response, and cold-responsive transcription factors. A network analysis was conducted to map miRNA-transcription factor (TF) interactions, with Cytoscape used for visualization. Key miRNA-TF regulatory hubs were identified, providing insights into the regulatory mechanisms of cold acclimation. To validate the findings, a subset of differentially expressed miRNAs was analyzed using qRT-PCR. The expression levels of miRNAs and their predicted target genes were correlated to confirm regulatory relationships. Finally, Tiwari *et al*. (2020) identified 93 differentially expressed miRNAs in *Arabidopsis thaliana* during cold acclimation, including 14 previously implicated in cold response. Although the study did not explicitly label these miRNAs as biomarkers, their condition-specific expression suggests functional involvement in cold stress adaptation.

Relevance to This Thesis. A key strength of this study is its inclusion of multiple small RNA classes—miRNAs, nat-siRNAs, and lncRNA-derived sRNAs, highlighting the importance of analyzing diverse RNA types to better capture the complexity of small RNA–mediated regulation in plants. While the study in Section 1.8.8 introduced a workflow focused exclusively on miRNA analysis across multiple abiotic stresses, the Tiwari et al. study applied a similar workflow to multiple small RNA classes under a single stress condition, demonstrating that consistent analytical pipelines can support multi-class sRNA profiling. At the same time, the study reveals important methodological limitations. While miRNAs were identified *de novo* (from scratch), other small RNAs were detected using prior annotations, limiting the discovery of novel or condition-specific elements—particularly for siRNAs, which remain less characterized. The approach also did not examine potential functional interactions between different small RNA types, such as shared targets or coordinated regulation, nor did it explore how these small RNAs might jointly contribute to stress-responsive regulatory networks. Furthermore, the study did not investigate the evolutionary relationships among the small RNAs identified, missing an opportunity to connect sequence conservation or divergence to functional roles in cold stress adaptation. Finally, it lacked structural or biogenesis-based classification of siRNA precursors, which is important for distinguishing among distinct processing mechanisms and regulatory functions. Addressing these gaps, this thesis develops computational strategies for the simultaneous *de novo* prediction of miRNAs and siRNAs, moving beyond database-dependent workflows. Wheat serves as an ideal model system for this effort, with its polyploid genome, evolutionary depth, and variation in stress tolerance offering a rich context for uncovering regulatory small RNAs. Our approach supports more comprehensive discovery and interpretation of small RNAs involved in plant stress responses.

3.5 siRNA Biogenesis Hypothesis

Understanding siRNA biogenesis is essential for developing methods that accurately predict and annotate siRNAs. The biogenesis pathway determines the size, structure, and functionality of siRNAs, making it crucial to distinguish them from other small RNA classes. Knowledge of siRNA biogenesis enables the establishment of robust prediction frameworks and annotation criteria, enhances the ability to identify true siRNA precursors, and leads to more efficient and accurate siRNA analysis.

3.5.1 Dicer Processing for RNAi Activation

The size of small RNAs is primarily determined by Dicer activity, which varies across tissues and species due to differences in Dicer-like (DCL) protein subtypes and RISC components (Fukunaga *et al.*, 2012). Dicer functions as a molecular ruler, cleaving double-stranded RNA (dsRNA) into small RNA duplexes of 20 to 25 nucleotides (Figure 3.5) (MacRae *et al.*, 2006). However, different DCL paralogs exhibit distinct substrate preferences. For example, DCL1 mainly produces 21-nt miRNAs, while DCL3 generates 24-nt heterochromatic siRNAs (Gasciolli *et al.*, 2005). These variations reflect species- and tissue-specific adaptations, where different RISC components fine-tune small RNA processing and function.

Dicer-mediated cleavage is essential for small RNA activation. Biochemical studies have shown that pre-processed 21-nt duplexes are ineffective RNAi substrates, highlighting the necessity of Dicer-mediated processing for activation. While precursor structure and length can vary, RNAi activation relies on dsRNA length rather than a specific secondary structure. Dicer can efficiently process substrates longer than 27 nt (Kim *et al.*, 2005). Additionally, a preference for a 3' overhang over blunt or 5' overhang structures suggests that Dicer substrate recognition is influenced by the end structure of the dsRNA. These findings reinforce the idea that the primary determinant for RNAi activation is Dicer cleavage rather than the precise precursor conformation (Figure 3.6).

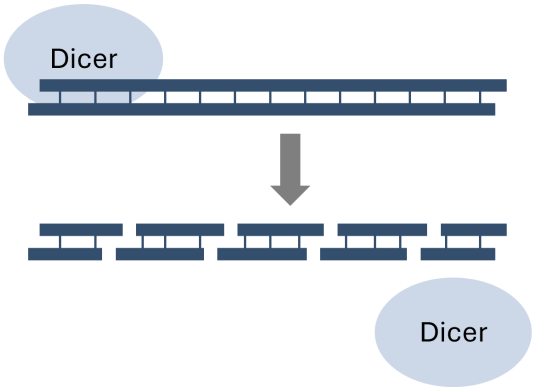


Figure 3.5: Dicer-mediated sequential cleavage of dsRNA. Dicer sequentially cleaves double-stranded RNA (dsRNA), generating small RNA duplexes of varying lengths (20, 21, 22, 23, or 24 nucleotides) depending on the Dicer variant. It processes the dsRNA substrate one cycle at a time until it rides off the substrate, either by reaching the end or encountering structural obstacles (MacRae *et al.*, 2006).

Once Dicer processes the dsRNA, the resulting siRNA duplex is loaded into RISC, where one strand is selected as the active guide RNA while the other is degraded. Following RISC loading, miRNAs and siRNAs become functionally indistinguishable, as both act as sequence-specific RNA guides. The distinction between the two classes primarily arises before RISC incorporation, during precursor processing. Their downstream effects depend on the tissue- or species-specific composition within RISC, which determines whether the small RNA leads to mRNA cleavage, translational repression, or chromatin modifications.

Since miRNAs and siRNAs share the same core targeting mechanism, miRNA target prediction tools can be adapted for siRNA target identification. This mechanistic overlap provides a rationale for extending miRNA-based bioinformatics approaches to siRNA studies.

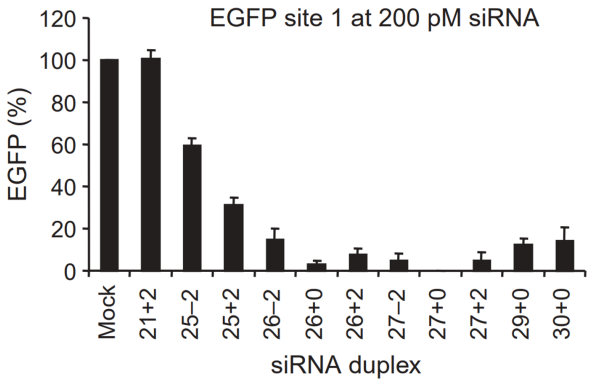


Figure 3.6: Dicer processing is essential for RNAi activation. The experiment measures EGFP fluorescence as a readout of RNAi potency, where lower EGFP levels indicate stronger silencing. Pre-processed 21-nt siRNA duplexes fail to induce potent RNAi, while longer dsRNAs (>27 nt) are significantly more effective, demonstrating that Dicer cleavage is required for activation. siRNAs with a 3' overhang exhibit stronger silencing compared to those with blunt or 5' overhang structures, showing a preference in Dicer substrate recognition. Adapted from Kim *et al.* (2005).

3.5.2 Molecular Origins of Dicer Substrates

Biochemical studies indicate that Dicer processes both single-molecule hairpins and two-molecule dsRNAs (Chakravarthy *et al.*, 2010; Howell *et al.*, 2007; Kakrana *et al.*, 2018). Based on how the double-stranded region forms at the molecular level, siRNA precursors can be classified into three distinct types (Figure 3.7) (Axtell *et al.*, 2006; Howell *et al.*, 2007; Axtell, 2013a). These structural categories reflect different sources of dsRNA substrates for Dicer processing, each arising through distinct molecular mechanisms.

This classification framework helps to contextualize the diverse origins of siRNA precursors and their role in small RNA pathways. Examining siRNA biogenesis from a precursor perspective not only facilitates the distinction between siRNAs and other small RNA classes but also enhances computational approaches for siRNA identification and classification. However, distinguishing between precursor types remains challenging, as different structural categories may share overlapping sequence features or processing signatures, necessitating further refinement of classification criteria.

Type II precursors (intramolecular foldback) arise from a Pol II-transcribed single RNA molecule that folds back onto itself, forming a partially complementary dsRNA region through intramolecular base-pairing. This category includes short precursors, such as canonical miRNA hairpins under 300 nucleotides, and longer transcripts that generate structured regions capable of siRNA production. While short Type II precursors have been extensively characterized in miRNA biogenesis, the role and prevalence of long foldback precursors remain less characterized. Zhai *et al.* (2015) identified such siRNA-generating long hairpins in plants, and our preliminary data suggest their presence in wheat as well. However, distinguishing genuine siRNA precursors from random transcript foldbacks remains a challenge.

Most bioinformatics prediction tools have been optimized for short Type II precursors (Yang et Li, 2011; Axtell, 2013b; Leclercq *et al.*, 2013; Lei et Sun, 2014; Wu *et al.*, 2019), which typically produce only one pair of Dicer-processed RNAs, such as canonical miRNAs and mirtrons (Axtell, 2013b; Axtell et Meyers, 2018; Rorbach *et al.*, 2018). In contrast, computational approaches for detecting long Type II precursors, as well as Type I and Type III precursors, remain underdeveloped. Developing more accurate prediction models for these less-characterized precursor types will be essential for fully understanding the diversity and regulatory roles of siRNA-generating loci.

Type III precursors (hybridization of two molecules) originate from the complementary hybridization of separately transcribed Pol II-derived RNAs, which meet each other within the cellular environment, potentially in the nucleus (Zhang *et al.*, 2012b). These precursors do not form structured secondary folds within a single molecule but instead rely on intermolecular base pairing to generate a dsRNA substrate for Dicer processing. If transcribed from overlapping genes in opposite orientations, the resulting siRNA alignment pattern closely resembles that of Type I precursors. Conversely, if the transcripts originate from non-overlapping genes, the alignment pattern is more similar to Type II precursors but remains restricted to a single genomic strand.

Type I precursors (second strand by synthesis) undergo a specialized biogenesis process that requires the synthesis of a complementary strand through RNA-dependent RNA polymerase (RDRP) activity (Figure 3.9). Consequently, siRNA alignments are evenly distributed on both strands of the genomic locus.

RDR2 catalyzes second-strand RNA synthesis during RNA-directed DNA methylation (RdDM) (Xie et Yu, 2015). In this pathway, Pol IV transcribes single-stranded RNA (ssRNA) from repetitive or transposon-derived

loci, which is then converted into dsRNA by RDR2. The resulting dsRNA is processed by Dicer-like proteins into 24-nt siRNAs, which guide DNA methylation and transcriptional gene silencing. Unlike RDR6-mediated siRNA biogenesis, RDR2-dependent dsRNA synthesis does not rely on miRNA-directed cleavage.

RDR6, in contrast, requires miRNA-directed cleavage to initiate dsRNA synthesis (Howell *et al.*, 2007; Song *et al.*, 2012). The triggering miRNAs include miR390 for tasiRNAs, miR2118 for 21-nt phased siRNAs, and miR2275 for 24-nt phased siRNAs, which initiate precursor cleavage, followed by RDR6-mediated second-strand synthesis. The resulting dsRNA substrates are processed by distinct DCLs, leading to the production of secondary siRNAs with characteristic spatial organization.

The "**two-hit trigger**" model proposes that two miRNA cleavage events, which establish the precursor boundaries, are required to recruit RDRP (Axtell *et al.*, 2006), although some studies suggest that a single cleavage event may suffice. Structural modifications in the RNA-induced silencing complex (RISC), induced by specific miRNAs such as miR390, may facilitate RDR6 recruitment.

Unlike Type II and III precursors, where dsRNA formation occurs naturally through intramolecular folding or transcript hybridization, Type I precursors require enzymatically driven dsRNA synthesis. This enzymatic dependency distinguishes Type I precursors but also complicates their classification, as siRNA precursors without clear triggering events may be misclassified under other precursor categories. A strict reliance on trigger-based classification may, therefore, only capture a subset of siRNA precursors, potentially overlooking those that employ alternative initiation mechanisms.

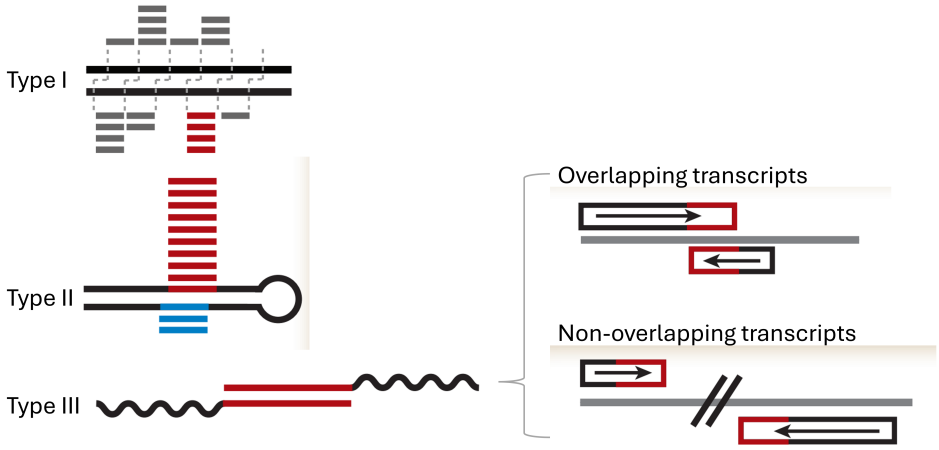


Figure 3.7: Distinct endogenous sources of dsRNA substrates for Dicer processing. The molecular origins of Dicer substrates are classified into three precursor types. Type I – Second Strand by Synthesis arises from RNA-dependent RNA polymerase (RDRP)-mediated synthesis, where a single-stranded RNA is converted into double-stranded RNA following miRNA-directed cleavage and RDRP activity. Type II – Intramolecular Foldback forms through intramolecular base-pairing, where a single RNA transcript folds back onto itself, creating a stem-loop structure. This category includes both short miRNA precursors and longer siRNA-producing hairpins. Type III – Hybridization of Two Molecules originates from complementary base-pairing of two independent transcripts, forming a dsRNA substrate either from overlapping antisense transcripts or separate non-overlapping transcripts that hybridize. These precursor structures provide distinct sources of dsRNA substrates for Dicer processing. Adapted from Axtell (2013a).

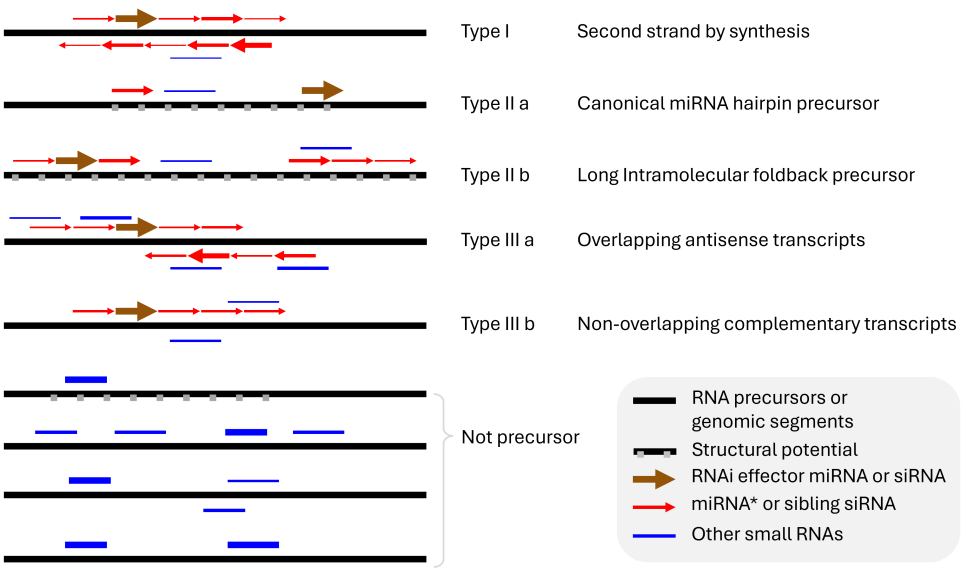


Figure 3.8: Small RNA alignment patterns across siRNA precursor and non-precursor models. **Type I – Second Strand by Synthesis** exhibits bidirectional alignment of phased (regularly spaced) small RNAs on both strands, reflecting RNA-dependent RNA polymerase (RDRP) activity. **Type II – Intramolecular Foldback** is subdivided into Type IIa, representing canonical miRNA precursors, which are shorter than 300 nt and typically produce only a single pair of miRNA and miRNA*; and Type IIb, representing long foldback precursors, where tandem Dicer processing generates multiple phased small RNAs along the precursor transcript. **Type III – Hybridization of Two Molecules** includes Type IIIa, where overlapping antisense transcripts generate bidirectional alignment patterns similar to Type I, and Type IIIb, where non-overlapping but complementary transcripts generate unidirectional phased small RNAs. Non-precursor regions show scattered and less structured small RNA alignment patterns, distinguishing them from true precursors. The legend denotes RNA precursors (black), structural potential (dashed grey), RNAi effector miRNAs or siRNAs (brown), miRNA* or sibling siRNAs (red), and other small RNAs (blue); the thickness of small RNA bars represents expression levels.

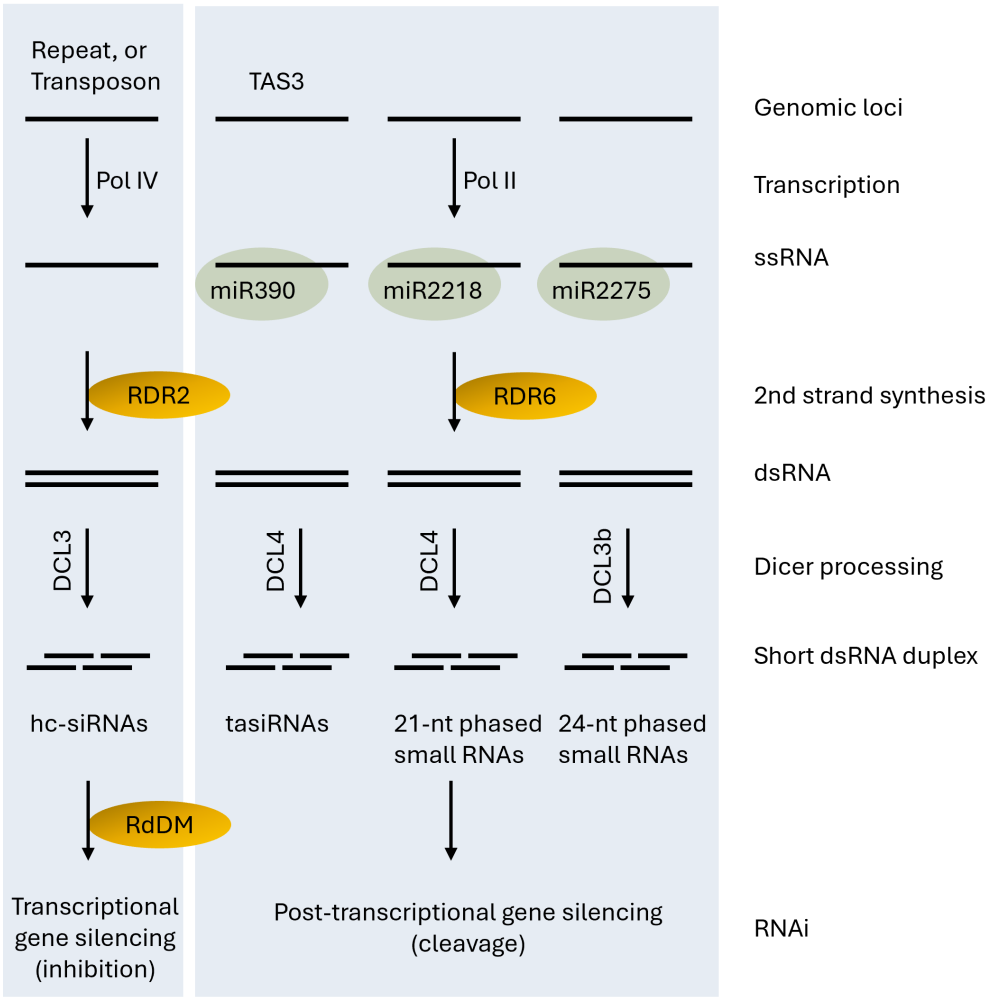


Figure 3.9: RDR2- and RDR6-mediated siRNA biogenesis. The left panel illustrates the RDR2-dependent pathway, which generates 24-nt heterochromatic siRNAs (hc-siRNAs). These hc-siRNAs mediate transcriptional gene silencing (TGS) via the RNA-directed DNA methylation (RdDM) pathway. The right panel depicts the RDR6-dependent secondary siRNA pathway, which is triggered by miRNA-directed cleavage (e.g., miR390 for tasiRNAs, miR2118 for 21-nt phased siRNAs, and miR2275 for 24-nt phased siRNAs). RDR6 synthesizes dsRNA from the cleaved transcripts, which is then processed by Dicer-like enzymes (DCLs) into tasiRNAs and phased small RNAs. These secondary siRNAs function in post-transcriptional gene silencing (PTGS) by directing the cleavage of target mRNAs. Adapted from Song *et al.* (2012); Xie et Yu (2015)

3.6 Assessing Phasing Potential

In the previous section, we established that Dicer cleavage is essential for activating RNAi by generating siRNA duplexes that guide gene silencing. While different types of Dicer substrates exist, accurately distinguishing these substrate subtypes remains challenging. Understanding their origins could enhance computational approaches for siRNA locus prediction, but as a first step, we focus on capturing the fundamental signature of Dicer cleavage itself. Researchers have developed various statistical approaches to detect Dicer-mediated phasing patterns, each designed to capture the periodic alignment of siRNA reads along precursor transcripts. These phase scores measure the regularity of siRNA spacing, reflecting the cyclical nature of Dicer processing.

The principle of phase score computation is to identify RNA alignment patterns produced by Dicer's sequential cleavage. Within a given search window (e.g., 169 nucleotides as used in Chapter 6), **cycles** represent 21-nt intervals, defined by **DicerCall** $= \{18, 19, \ldots, 31\}$, which is 21 in this example. Since Dicer cuts both sense and antisense strands, phase scores combine reads from both strands by aligning their start coordinates while accounting for the 2-nt offset of the antisense strand within the duplex. This process allows statistical methods to detect periodic siRNA signatures, providing a quantitative measure of phasing strength. To compute a phase score, the following parameters are considered:

- **k** the number of phase positions expressed by 21-nt RNAs,
- **n** the number of all positions expressed by 21-nt RNAs,
- **p** the number of reads for 21-nt RNAs at phase positions,
- **u** the number of reads for 21-nt RNAs at all positions, and
- **maxf** the number of reads for the most abundant 21-nt RNAs at phase positions, including those shifted by 1 or 2 positions to account for phase drift.

Existing scoring methods either evaluate phasing strength by assessing the accumulation of phased reads (e.g., Howell (Howell *et al.*, 2007), Guo (Guo *et al.*, 2014)) or determine phasing significance by comparing occupied phase positions to all positions (e.g., Chen p-value (Chen *et al.*, 2007)).

$$\textbf{Howell score:} \quad \text{Howellb}(p,k) = \begin{cases} 0, & \text{if } k < 3, \\ \log\left[(1+p)^{k-2}\right], & \text{if } k \geq 3. \end{cases} \tag{3.1}$$

$$\textbf{Chen score:} \quad \text{pval}(k,n) = \begin{cases} 1, & \text{if } k < 3, \\ \sum_{i=k}^{m} \frac{\binom{20m}{n-i}\cdot\binom{m}{i}}{\binom{21m}{n}}, & \text{if } k \geq 3 \text{ and } n \geq i. \end{cases} \tag{3.2}$$

where $m = (2c-1)$ is a constant if cycle c is given.

$$\textbf{Guo score:} \quad \text{Guo}(u,k,p,maxf) = \begin{cases} 0, & \text{if } k < 3, \\ \left(\frac{1}{u}\right)\cdot k \cdot \ln(p) \cdot maxf, & \text{if } k \geq 3, \end{cases} \tag{3.3}$$

3.6.1 Example Calculation of Phase Scores

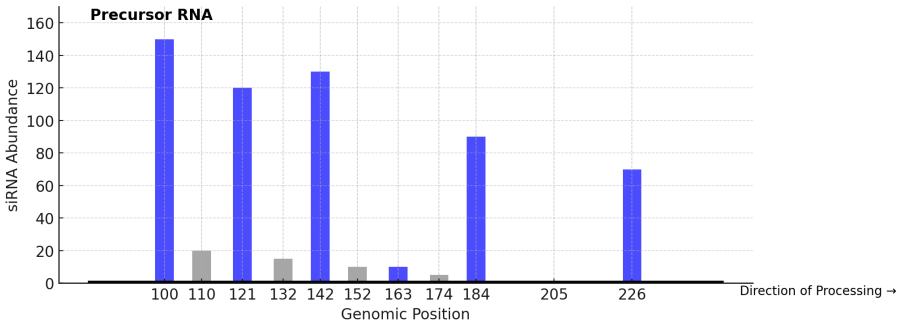


Figure 3.10: Illustration of an example phased siRNA locus showing both phased and non-phased positions. Phased positions (blue) represent siRNAs consistently produced at periodic intervals, while non-phased positions (gray) show siRNAs with sporadic expression. Total phased siRNA abundance: $p = 570$, Total 21-nt siRNA abundance: $u = 620$, Maximum phased siRNA abundance: $maxf = 150$, Number of phased positions: $k = 6$, Number of expressed positions: $n = 10$. The precursor RNA is processed in the indicated direction, with phased siRNAs accumulating at specific positions.

Howell Score. Suppose $p = 570, \quad k = 6$, then:

$$\text{Howell}(p, k) = \log((1 + 570)^{6-2}) = \log(571^4) = \log(10606045641) \approx 25.39$$

Chen Score. Suppose $k = 6, \quad n = 10, \quad m = 17$, then:

$$\text{pval}(k, n) = \sum_{i=k}^{m} \frac{\binom{20m}{n-i} \cdot \binom{m}{i}}{\binom{21m}{n}} \approx 8.45 \times 10^{-7}$$

Since pval $\ll 0.05$, this suggests that the observed siRNA phasing is highly unlikely to have occurred by chance.

Guo Score. Suppose $u = 620, \quad k = 6, \quad p = 570, \quad maxf = 150$, then:

$$\text{Guo}(u, k, p, maxf) = \left(\frac{1}{620}\right) \times 6 \times \ln(570) \times 150 = \left(\frac{1}{620}\right) \times 6 \times 6.3456 \times 150 = 9.21$$

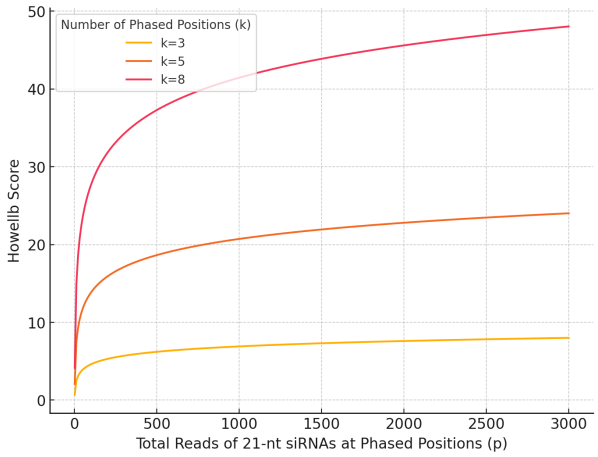
3.6.2 Critical Evaluation for Howell, Chen, and Guo Scores

The Howell, Chen, and Guo scoring methods evaluate phased siRNA loci from different perspectives, each with strengths and weaknesses depending on the biological context and dataset characteristics. To better understand their behavior, we simulate phase scores by applying these methods across different variable conditions. Figure 3.11 illustrates how each scoring method varies with **k**, providing insights into their interpretability and relative scaling. However, these scoring methods have inherent limitations, which can lead to false positives, loss of biologically relevant information, or dataset-specific biases.

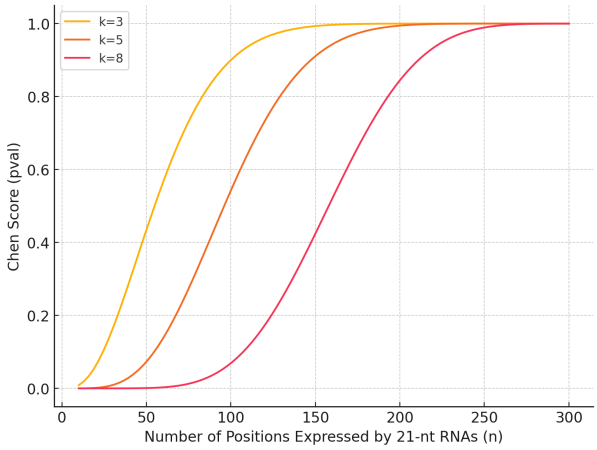
To address these concerns, we consider data pretreatment strategies that assess how filtering may improve score interpretability and relevance, using criteria such as $p/u > a$, $maxf/p > b$, and $maxf > c$, where, for example, $a = 0.8$, $b = 0.3$, and $c = 100$. These constraints are intended to remove weakly phased or low-expression loci, ensuring that only biologically meaningful loci with strong phasing signals are considered. While we do not experimentally evaluate the effects of these filters, this theoretical assessment helps identify potential limitations and refinements for future improvements.

Even with these preprocessing steps, assessing the accuracy and optimal thresholds for each scoring method continues to be challenging. Below, we critically assess each method, and Table 3.1 summarizes the strengths and limitations of Howell, Chen, and Guo scores, highlighting key differences in their sensitivity, statistical

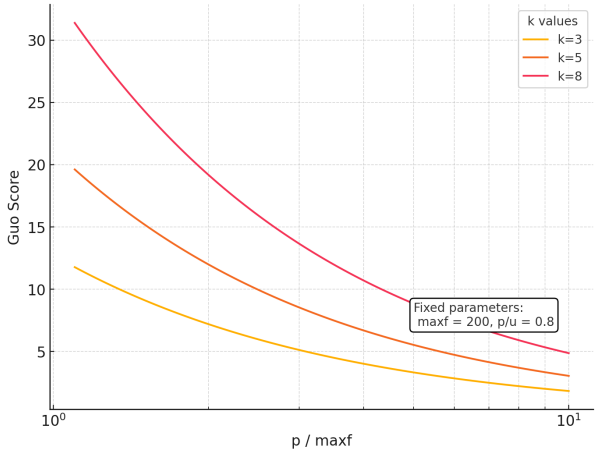
rigor, and interpretability. This overview serves as a foundation for evaluating their performance across different datasets and identifying areas for refinement.



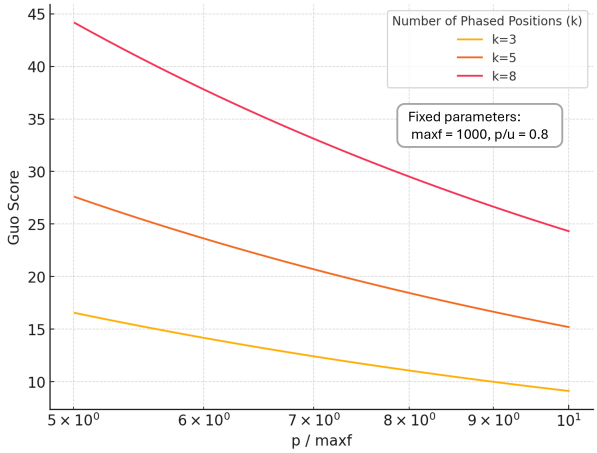
(a) Howell Score Distribution



(b) Chen Score Distribution



(c) Guo Score Distribution (maxf = 200)



(d) Guo Score Distribution (maxf = 1000)

Figure 3.11: Score distributions of phase scoring functions for different k values. Each panel presents a different scoring method: (a) Howell, (b) Chen, (c) Guo with $maxf$ = 200, and (d) Guo with $maxf$ = 1000. **Definitions:** k — number of phased positions, n — number of positions expressed by 21-nt RNAs, p — total phased siRNA abundance, u — total 21-nt RNA abundance, $maxf$ — abundance of the most abundant phased siRNA. In panels (c) and (d), the parameter $maxf$ is varied (200 vs. 1000) to compare its effect on the Guo score, while the ratio p/u is fixed at 0.8 to simplify the illustration by reducing the number of free variables.

3.6.2.1 Howell Score: Strength in Sensitivity, Weakness in Controlling for Background

Howell is highly sensitive to total phased siRNA abundance (**p**) and the number of phased positions (**k**), making strongly phased loci stand out. Higher **k** leads to an exponentially higher Howell score, making it useful for ranking loci based on phasing strength. Additionally, Howell is easy to interpret because a higher number of phased reads directly corresponds to a higher score. However, Howell does not account for background expression. A locus with high expression at both phased and non-phased positions still receives a high Howell score, leading to false positives. It does not penalize loci where phasing strength is diluted by high non-phased siRNA expression. It may offer relative rankings, but establishing a precise and statistically valid cutoff for true phased loci is challenging.

Pretreating the data might mitigate these issues. Applying the **p/u** filter removes loci where phased siRNAs are diluted by high total siRNA expression, reducing false positives. The **maxf/p** filter ensures that the most abundant phased siRNA is sufficiently strong, preventing excessive uniform expression from inflating the score. The **maxf** filter helps exclude weakly expressed loci that might receive a high Howell score despite low biological relevance. With these filters, Howell becomes more reliable, but it still lacks a clear statistical threshold for classification.

3.6.2.2 Chen: Statistical Rigor but Loss of Expression Information

Chen offers a threshold for identifying phased loci with confidence, using a significance cutoff of p-value < 0.05. It accounts for the distribution of phased versus non-phased positions, making it less prone to false positives in datasets with globally high expression. However, Chen completely ignores expression levels. It only evaluates whether phased siRNA positions follow a statistical pattern. A locus with strong phasing but low expression is treated the same as one with strong phasing and high expression, resulting in the loss of biologically relevant information. This makes Chen not useful for ranking loci; it is strictly a binary classifier.

Applying preprocessing filters might improve Chen's effectiveness. The **maxf/p** filter ensures that significant loci contain a dominant phased siRNA. The **maxf** filter removes loci with very weak expression, thus ensuring biological relevance. These filters help retain meaningful phased loci in p-value calculations, but the method still lacks expression strength information.

3.6.2.3 Guo Score: Balancing Background, Expression Strength, and Phasing Structure

The Guo score evaluates phased siRNA loci by integrating four parameters: background noise (**u**), phased positions (**k**), phased expression (**p**), and expression dominance (**maxf**). While Howell follows a simple rule where higher scores indicate stronger phasing and Chen follows the statistical rule where lower p-values indicate stronger confidence, Guo depends on four variables. Guo's complexity makes it less intuitive. It increases with **k**, ensuring higher confidence for loci with more phased positions, while incorporating **u** to penalize background expression and **maxf** to account for expression dominance. However, Guo strongly depends on **maxf**. As seen in Figure 3.11, increasing **maxf** proportionally scales scores. This can inflate scores for loci with a single highly abundant siRNA despite weak overall phasing. Additionally, Guo lacks a statistical threshold, making it difficult to determine an absolute cutoff for strong phasing.

Pretreatment, which improves other scoring methods, might have minimal impact on Guo since it already integrates key factors in its calculation. Rather than enhancing reliability or interpretability, filtering mainly adds computational complexity. Unlike Howell and Chen, where pretreatment mitigates specific biases, Guo's multi-variable nature makes it difficult to assess how filtering affects rankings. Since pretreatment does not directly improve Guo's interpretability, refining **maxf** weighting or applying normalization may be more effective. In summary, Guo balances phasing structure (**k, p**) with expression characteristics (**u, maxf**), making it useful for ranking loci. However, its dependence on **maxf** and lack of a statistical threshold require careful filtering to improve interpretability and prevent overemphasis on single dominant siRNAs.

3.6.2.4 Unresolved Issues in Phasing Score Evaluation and Interpretation

Although widely used, current phasing scores lack standardized thresholds and perform inconsistently across datasets, making them difficult to interpret in large-scale, cross-library analyses. While some tools, like Chen's, provide statistical cutoffs (e.g., p-value < 0.05), others such as Howell and Guo do not, limiting them to relative comparisons. These methods are also sensitive to global expression differences, which makes it difficult to apply a universal scoring threshold. Pretreatment steps such as filtering low-abundance or weakly phased loci can improve score interpretability and ranking stability but do not solve the fundamental problem: they do not establish whether a given score truly indicates a biologically meaningful phased siRNA locus. Our evaluation suggests that continued refinement of phasing thresholds yields limited gains. Rather than optimizing scoring systems in isolation, we propose an alternative strategy that incorporates additional biological features to improve the identification of functionally relevant siRNA loci.

Another limitation of current scoring approaches is their dependence on a narrow definition of phasing. Most metrics focus on 21-nt siRNAs and do not incorporate information about other small RNA sizes, which may also reflect biologically significant processing events. For example, while **u** (total 21-nt siRNA abundance) and **n** (total expressed 21-nt positions) are useful for background correction, they exclude broader biogenesis signals. This omission may compromise score reliability when validating loci with complex or mixed RNA processing signatures. Moreover, existing scoring approaches were primarily developed using *Arabidopsis* or rice datasets, and their applicability to other plant species remains uncertain. Broader cross-species validation would help determine whether these scores generalize or require dataset-specific adjustments.

A major shortcoming of current phasing scores is their inability to distinguish between structurally distinct siRNA precursors. Diverse precursor types—including TAS loci, inverted repeats, and multiple classes of phased precursors arising from second-strand synthesis, intramolecular foldbacks, or intermolecular duplexes—can all yield similar phasing patterns, despite differing in processing mechanisms and regulatory potential. Without considering these structural differences, phasing scores alone offer limited biological insight. Incorporating structural classification alongside scoring could improve interpretation by revealing precursor-specific processing modes. Structural features such as base-pairing patterns, folding energy, sequence composition, and conserved secondary motifs influence Dicer recognition and processing efficiency. Including these attributes in predictive models could refine locus classification and improve the identification of functional siRNA producers.

Finally, phase scoring does not account for downstream interactions that affect regulatory relevance. Some phased siRNAs undergo secondary amplification—often miRNA-triggered—yet these amplification signatures are not captured by existing scoring systems. Furthermore, not all phased siRNAs participate in meaningful gene regulation. A highly phased locus may generate abundant siRNAs, but without functional targets, its biological role remains uncertain. Coupling phase scores with target prediction, interaction analysis, and functional annotation would help distinguish loci with actual regulatory impact from those that merely exhibit technical phasing. To support cross-dataset comparisons, normalization strategies such as quantile normalization, min–max scaling, or Z-score standardization can reduce variability due to sequencing depth or library size. Still, without adaptive thresholding and better functional integration, phasing scores alone are insufficient for robust siRNA locus classification.

Table 3.1: Comparison of Howell, Chen, and Guo phasing scores. Expression-related variables include p (phased abundance), u (total expression), $maxf$ (dominant small RNA), and k (number of phased positions). Pretreatment filters loci using thresholds on p, u, and $maxf$.

Aspect	**Howell**	**Chen**	**Guo**
Score Range	0 to ∞, exponential	0 to 1, sigmoidal	0 to ∞, logarithmic
Expression Features Used	**p**, **k** only	**k** only	**p**, **u**, **maxf**, **k**
Statistical Threshold	No	Yes, $p < 0.05$	No
Interpretability	Simple, easy to rank	Easy, threshold-based	Complex, many variables
False Positives	High if background is strong	Low, may miss weak loci	Lower, but dataset-dependent
Best Use	Ranking by expression	Locus validation/classification	Ranking with noise adjustment
Key Limitation	Ignores background, no cutoff	Ignores expression strength	Difficult to interpret, no cutoff
Effect of Pretreatment	Removes dominant noise	Ensures biological relevance	Reduces noise-related errors

3.7 Computational Strategy for Gene Therapy Viral Vector Integrity

Long-read sequencing has emerged as a powerful tool for assessing RNA integrity in therapeutic applications, particularly in the context of viral vectors used for gene therapy. Its ability to capture full-length sequences provides a comprehensive view of RNA structures, contaminants, and unintended chimeric molecules, thereby enhancing the assessment of therapeutic RNA quality.

RNA-based drugs, such as those delivered via recombinant adeno-associated virus (rAAV) vectors, face significant challenges due to the inherent complexity of viral production processes. The error-prone nature of viral genome packaging can lead to various unintended products, including truncated genomes, host-vector chimeric sequences, and reverse-packaged genomes containing undesired plasmid sequences. Traditional quality control methods, such as gel electrophoresis and qPCR, lack the resolution needed to accurately identify these structural variations and contaminants. Furthermore, the high sensitivity of RNA therapeutics to degradation poses additional challenges in ensuring their integrity and efficacy.

To address these challenges, computational frameworks utilizing long-read sequencing have been developed to provide high-resolution insights into RNA-based therapeutics. One such approach is AAV-GPseq, a bioinformatics pipeline that applies single-molecule, real-time (SMRT) sequencing to assess the integrity of rAAV genomes. By aligning long-read sequencing data to reference sequences, this approach identifies full-

length, truncated, and reverse-packaged genomes, as well as host-vector chimeric sequences (Figure 3.12). Furthermore, normalization techniques using internal controls, such as lambda-phage DNA, enable accurate quantification of these impurities. This comprehensive approach improves the detection of structural variations and contaminants, providing essential metrics for ensuring the safety and efficacy of RNA-based therapeutics.

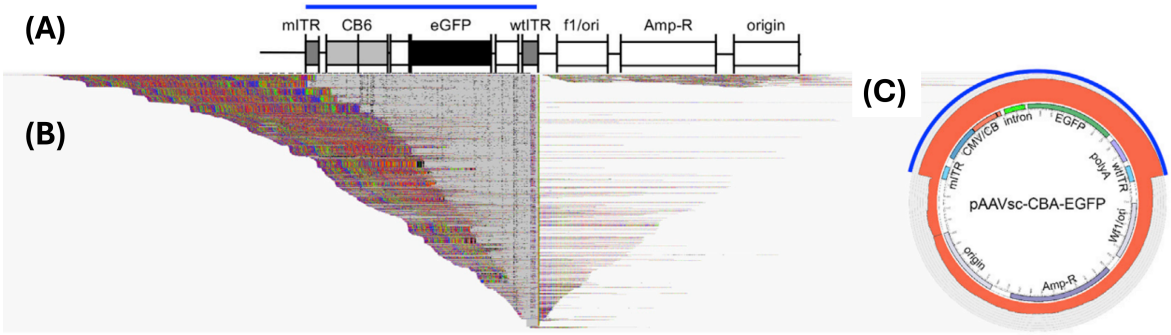


Figure 3.12: Alignments of heterogeneous populations to the rAAV vector reference. **(A)** Linearized representation of the rAAV vector from (C), showing annotated features (m-ITR, CB6 promoter, eGFP, wtITR, Amp-R, origin). **(B)** Long-read alignments to the linearized vector; each line is a single read. High density between ITRs indicates successful rAAV production, while residual expression occurs outside cargo regions. **(C)** Circular vector construct used for therapeutic delivery. Orange indicates expression coverage; the blue line marks the therapeutic region (mock eGFP). Adapted from (Tai *et al.*, 2018).

3.7.1 AAV-GPseq: High-Resolution Profiling of AAV-Genome Population

Tai *et al.* (2018) introduced the AAV-genome population sequencing (AAV-GPseq) bioinformatics pipeline, a methodology utilizing PacBio's **Single-Molecule Real-Time** (**SMRT**, Ardui *et al.*, 2018) sequencing to achieve full vector genome resolution and analyze recombinant adeno-associated virus (rAAV) genome integrity. AAV-GPseq enables direct assessment of packaged vector genomes as intact molecules, revealing previously unknown complexities such as genome heterogeneity and unintended chimeric sequences. The study demonstrated that vector preparations, which appeared homogeneous by traditional gel electrophoresis, contained less than 50% full-length genomes, highlighting the limitations of standard quality control meth-

ods in accurately evaluating vector genome integrity. Additionally, human-vector chimeric sequences were identified, raising concerns about unintended host genome integrations.

The methodology involved a multi-step approach, beginning with the preparation of purified self-complementary AAV (scAAV) genomes, which were ligated to adapters to allow for sequencing from the 5' inverted terminal repeat (ITR) to the 3' ITR. The prepared libraries underwent SMRT sequencing, which captured **long-read** sequences, enabling the analysis of full-length vector genomes and the detection of structural variants within the vector population. The sequencing data were processed using a custom bioinformatics pipeline, where SMRT sequencing reads were aligned to a reference sequence corresponding to the expected rAAV genome. This alignment facilitated the classification of full-length genomes, truncated genomes that lacked portions of the expected sequence, reverse-packaged genomes containing plasmid backbone or helper sequences, and chimeric genomes comprising host-cell genomic sequences fused with vector ITRs.

To quantify the prevalence of impurities, the study calculated the relative abundance of different genome types based on the proportion of aligned reads. Additionally, an enzyme-digested Lambda-phage DNA spike-in was used as an internal control to normalize the data and account for molecular loading biases inherent to SMRT sequencing, ensuring an accurate assessment of truncated genomes. The results showed that in addition to a large fraction of truncated genomes, reverse-packaged genomes containing unwanted plasmid sequences were prevalent. Furthermore, host-cell genomic sequences were found fused to vector ITRs, with between 1.3% and 2.3% of the vector population carrying these unintended chimeric integrations.

The findings of this study underscore the complexity of rAAV vector preparations and the necessity for improved quality control in gene therapy applications. By enabling full-length sequencing and precise classification of rAAV genome variants, AAV-GPseq offers a more accurate approach to assessing vector genome integrity compared to conventional methods. This work highlights the importance of rigorous vector characterization in ensuring the efficacy and safety of gene therapy, particularly in minimizing the risks associated with truncated, reverse-packaged, and chimeric vector genomes.

3.8 Validation Strategies in Computational Small RNA Research

Small RNA analyses require diverse strategies to assess accuracy, robustness, and reproducibility. As discussed in Section 1.7.2, standard metrics such as precision, recall, F1-score, and MCC provide a foundational framework for statistical evaluation. However, these metrics are typically complemented by broader vali-

dation practices that address both methodological soundness and biological relevance.

Cross-validation is widely used to evaluate generalizability, especially when models are trained on limited or imbalanced datasets. In studies involving biological replicates or time-series data, careful partitioning is essential to avoid information leakage and to maintain the independence of test sets. When no established benchmark datasets are available, curated datasets are often constructed from real sequencing data using biologically informed criteria to define positive and negative classes. This approach supports meaningful evaluation consistent with biological knowledge.

Simulated or synthetic data can further support method validation by enabling controlled assessments of sensitivity and specificity. Embedding known patterns within randomized background reads allows researchers to isolate a method's ability to detect true signals in noisy environments. Benchmarking against existing tools is particularly important when introducing new methods, especially if widely adopted alternatives exist. These comparisons help contextualize performance and reveal trade-offs in accuracy, runtime, or scalability.

Evaluating runtime and scalability becomes critical when methods are applied to large datasets or integrated into high-throughput pipelines. Computational efficiency is especially relevant for algorithms involving iterative computation, genome-wide scans, or feature extraction steps that grow with input size. Beyond predictive performance, robust statistical validation—including significance testing and false discovery rate (FDR) control—is essential for minimizing false positives in large-scale analyses.

Functional validation serves as a bridge between computational output and biological interpretation. This may involve gene target prediction, enrichment analysis, or consistency with known regulatory mechanisms and motifs. Reproducibility across libraries, datasets, or batches is equally important, as it underscores the robustness of predictions and helps identify potential technical biases (Auer et Doerge, 2010). While experimental validation, such as degradome sequencing or qRT-PCR, remains solid evidence, evolving community guidelines increasingly recognize the role of bioinformatics validation (Axtell et Meyers, 2018), particularly when computational methods demonstrate statistical rigor and agreement with external annotations or databases.

Table 3.2: Validation techniques used across computational components in this thesis.

Validation Technique	**mirLibSpark**	**Wheat sRNA Analysis**	**siWalk**	**ArcticFox**
Evaluation Metrics	Yes	Not applicable	Yes	Not applicable
Cross-validation	Not used	Stratified by design	Group-stratified	Not used
FDR control	Yes: BH correction	Yes: DESeq2 embedded	Yes: phasing score thresholding	Not used
Simulation	Yes: synthetic libraries	No	No	No
Curated benchmark datasets	No	Real data, not curated	Yes: inferred positives and negatives	Not applicable
Benchmarking against tools	Yes: compared to miRDeep-P, ShortStack, miR-PREFeR	Yes: compared to ShortStack, miRBase, literature, TAS homologs	Yes: compared to ShortStack, PHASIS	No
Runtime benchmarking	Yes	Not benchmarked; pipeline scales by design	Yes	Not emphasized
Functional annotation	Yes: comparative genomics, target prediction, GO/KEGG annotation, enrichment	Yes: target prediction, GO/KEGG annotation, enrichment	Yes: comparative genomics	Motif consistency; experimental validation performed externally, not shown
Experimental validation	Not performed	Not performed	Not performed	Performed externally; data not shown
Reproducibility testing	Yes: across libraries	Yes: across replicates	Yes: across folds	Yes: across batches

CHAPTER 4

MIRLIBSPARK: A SCALABLE NGS PLANT MICRORNA PREDICTION PIPELINE FOR MULTI-LIBRARY FUNCTIONAL ANNOTATION

Handling the exponential growth of sRNA-seq data presents substantial computational challenges, particularly in plants with large and complex genomes. Among small RNA classes, miRNAs are the most well-characterized, with established biogenesis models and standardized prediction criteria. However, tools like miRDeep-P do not scale effectively to multi-library datasets or large plant genomes, limiting their utility in high-throughput biomarker discovery. This chapter introduces `mirLibSpark`, a Spark-based framework that unifies core components of miRNA analysis in large-scale sRNA-seq studies. By supporting parallel computation and adhering to gold-standard annotation criteria, it enables efficient processing of large datasets from diverse plant species including wheat, rice, maize, potato, stiff brome, goatgrass, and *Arabidopsis*. In line with the first specific objective—developing a scalable and integrated pipeline for high-throughput miRNA prediction and annotation— this chapter presents the design, implementation, and evaluation of `mirLibSpark`, a distributed framework for large-scale miRNA analysis. The methodology is structured to:

- Provide a scalable Apache Spark–based pipeline (`mirLibSpark`) for efficient classification of plant miRNAs across multiple sequencing libraries, including large and complex genomes such as wheat.
- Integrate downstream analyses—including differential expression, precursor analysis, target prediction, functional annotation, and functional enrichment—into a unified, automated workflow.

A preliminary version of this work was presented as a poster at the 2018 *Annual International Conference on Research in Computational Molecular Biology* (*RECOMB*) and the *RECOMB-Seq* Satellite Conference. It was later published as a full workshop paper at the 2019 *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (*ACM-BCB*), in the *Parallel and Cloud-based Bioinformatics and Biomedicine* (*ParBio*) track. The implementation and experimental design were conducted by Chao-Jung Wu and Dr. Amine M. Remita. CJW executed experiments and wrote the manuscript under the supervision of Professor Abdoulaye Baniré Diallo. The definitive version is included in Appendix A and is publicly accessible on arXiv at `https://arxiv.org/abs/2501.17998`.

4.1 Abstract

The emergence of Next Generation Sequencing drastically increases the volume of transcriptomic data. Although many standalone algorithms and workflows for novel microRNA (miRNA) prediction have been proposed, few are designed to process large volumes of sequence data from large genomes, and even fewer further annotate functional miRNAs by analyzing multiple libraries. We propose an improved pipeline for a high volume data facility by implementing mirLibSpark based on the Apache Spark framework. This pipeline is the fastest actual method, and provides an accuracy improvement compared to the standard. In this paper, we deliver the first distributed functional miRNA predictor as a standalone and fully automated package. It is an efficient and accurate miRNA predictor with functional insight. Furthermore, it compiles with the gold-standard requirement on plant miRNA predictions.

Keywords: distributed computing, parallelization, Apache Spark, microRNA classification, functional transcriptomic annotation

4.2 Introduction

The emergence of the Next Generation Sequencing (NGS) allows researchers to quickly and cheaply collect billions of expressed small RNAs (sRNAs) sequences according to a given condition so that the volume of small-RNAome data increases dramatically. However, the annotation of sRNAs has never been rapidly and systematically achieved. Parallelization approach is becmoing useful for computation-intensive tasks in biology (Biegert *et al.*, 2006). In the era of big data, CPU speeds and storage capacity are greatly improving. Hard disk input/output (I/O) is also dealt with by Hadoop MapReduce. Furthermore, Resilient Distributed Dataset (RDD) abstraction is performed to imporve Hadoop's disk-based computing (Zaharia *et al.*, 2012). Scalable MapReduce methods are revolutionizing traditional bioinformatics tools in RNA research (McKenna *et al.*, 2010; Nellore *et al.*, 2016; de Castro *et al.*, 2017).

Nowadays, several well-known bioinformatics tools are dedicated to annotating sRNAs as miRNAs in plants. MiRDeep-P (Yang et Li, 2011) employs a probabilistic scoring algorithm to predict novel miRNAs from precursor signatures and annotates directly known miRNA candidates as miRNAs without considering their dynamic expressions within the transcriptome. Currently, the best prediction is performed by ShortStack (Axtell et Meyers, 2018; Axtell, 2013b), and the fastest prediction tool is miR-PREFeR (Lei et Sun, 2014). Both ShortStack and miR-PREFeR incorporate miRNA annotation criteria (Meyers *et al.*, 2008) in their annotation

algorithms and decrease false positive predictions so that they annotate miRNAs with low false positives, high precision and high accuracy (Axtell et Meyers, 2018). Although ShortStack and miR-PREFeR take multiple pre-aligned sRNA-seq libraries as inputs, ShortStack is not scalable and has difficulty to deal with species having larger genomes such as maize that has 2 GB genome (Lei et Sun, 2014). MiR-PREFeR spawns subprocesses to execute a function on distributed input data across multiple processors on a given machine. This approach improved the processing speed for maize sRNA-seq libraries (Lei et Sun, 2014). However, the performance of any parallelization approach has not been assessed in species that have large genomes, such as wheat of 13 GB genome.

In our previous work (Agharbaoui *et al.*, 2015), we established a standardized integrative procedure to annotate miRNAs. The procedure includes four major steps: miRNA prediction, target gene prediction, differential expression analysis and functional enrichment. This data-intensive annotation procedure reduces false positive predictions by taking into account experts criteria, and renders meaningful miRNA annotations due to the incorporation of functional analysis. However, none of the existing tools has worked towards scaling up the entire procedure. To the best of our knowledge, there is only one miRNA predictor employing MapReduce (Pan *et al.*, 2015) to perform a few steps of the procedure. Additionally, MR-microT server (Kanellos *et al.*, 2014) uses MapReduce to annotate the downstream target genes of miRNAs.

In this work, we propose an improved miRNA prediction procedure that adds the miRNA annotation criteria for big data (Axtell et Meyers, 2018), incorporates functional and structural annotations and accelerates the processing time. We present mirLibSpark, a Spark-based miRNA annotation pipeline that executes the entire miRNA annotation procedure and is able to process multiple sRNA-seq libraries with high prediction performance and speed in both personal computer and high performance computing clusters.

4.3 MirLibSpark architecture and execution

4.3.1 MirLibSpark overview

Annotation modules. MirLibSpark contains modules dedicated to integrative procedure of functional miRNA annotation, as shown in Figure 4.1. It includes a module (M1) providing the reference genomes, Bowtie indices (Langmead *et al.*, 2009) and KEGG pathway functional annotations (Kanehisa *et al.*, 2017). Its core module is the miRNA prediction module (M2). Additional modules are attached to help users to better annotate miRNAs (M3: target gene prediction, M6: KEGG pathway annotation of the target genes and M5:

precursor visualization using VARNA (Darty *et al.*, 2009)) and to aggregate multi-library data (M4: differential expression analysis and M7: functional enrichment analysis).

MiRNA prediction module (M2). The mirLibSpark pipeline implements the four major computational steps for miRNA prediction: filtering, alignment, folding and prediction. For each run, mirLibSpark pipeline is initiated with a tunable configuration, then 30 different operations are applied on the multiple RDD created objects to make resilient and distributed data. These operations apply rigorously the miRNA annotation criteria and procedures for plants defined in (Meyers *et al.*, 2008; Axtell et Meyers, 2018). The parameters were already optimized and were designed to be adjustable in the pipeline, because the experts' criteria may evolve in the future.

Additional functional annotation modules (M3-M7). Post-prediction processing among multiple libraries is performed in one run. We compute the statistics of differential expression to select functionally relevant miRNA candidates (Benjamini et Hochberg, 1995; Kal *et al.*, 1999). Their target genes are subsequently analyzed for pathway enrichment (Vlachos *et al.*, 2015; Kanehisa *et al.*, 2017).

Implementation. Based on the computing load, computation steps within miRNA prediction module (shown in Figure 4.2), precursor visualization module and target gene prediction module are chained through Spark RDD operations (Zaharia *et al.*, 2012). In each step, an RDD object is generated to distribute the instances and operations across a cluster of computers. Each instance corresponds to a sRNA sequence and its related data. We used *map*() and *filter*() functions when it is required, where *map*() changes the content of an instance and *filter*() eliminates an instance based on a boolean statement.

We opt for the implementation of adapted functions to the Spark architecture for data and computing distribution instead of using extensively external programs. These programs should require less dependency and be currently maintained and open source to facilitate the pipeline redistribution. All the external programs are either distributed with the project or are automatically installed by mirLibSpark. Therefore, installation procedures for these programs are not required by the user.

Availability. MirLibSpark is available as an open source project in GitHub and as a docker image, along with its manual, in the following link
https://github.com/bioinfoUQAM/mirLibSpark.
We deliver mirLibSpark as a standalone package for two environments, (1) a personal computer or (2) a

remote server with high computational power. For biologists, we recommend to use the docker image of mirLibSpark in their personal computers. For bioinformaticians, the GitHub project can be deployed in any server. Upon installation of mirLibSpark, it is ready to execute experiments for a selected genome in a personal computer or with minimal configuration in a high performance computing server.

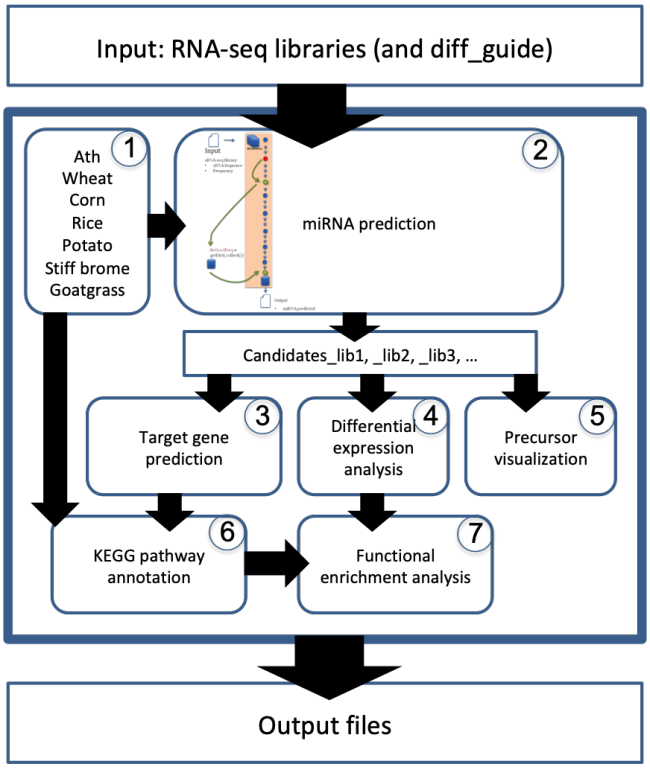


Figure 4.1: MirLibSpark overview. M1: genomic reference annotations. M2: miRNA prediction. M3: target gene prediction. M4: differential expression analysis. M5: precursor visualization. M6: KEGG pathway annotation. M7: functional enrichment analysis. M for module.

4.3.2 MiRNA prediction module (M2)

A simplified illustration of the RDD chaining in the prediction module is shown in Figure 4.2. A $SparkContext()$ is initiated with adjustable parameters of Spark and mirLibSpark. Default values, which are optimized for plants, are used if not otherwise specified by users. In each submission request, the pipeline can handle one to several sRNA-seq libraries. The first RDD object is created by reading the data of each input library. Thereafter, operations are applied on RDDs until the output is obtained and printed to a file. Several filters, including low complexity filter using dustmasker (Morgulis *et al.*, 2006) and various miRNA and precursor length and abundance criteria (Meyers *et al.*, 2008; Axtell et Meyers, 2018), are achieved by the $filter()$

operation. Only aligned sequences using Bowtie within a small number of genomic loci are retained. The RDD generated by alignment is reused as reference for the filter step of the miRNA candidate expression within precursor range. After mapping, only the sRNAs with length between 21 and 24 nucleotides (nt) are retained (Axtell et Meyers, 2018). A set of known non-miRNA sequences according to the annotation of the genome are removed. Pre-miRNA candidates surrounding the sRNA genomic loci are extracted and their secondary structures are computed using RNAfold (Lorenz *et al.*, 2011). The RNA structures and the decisions of miRCheck (Jones-Rhoades et Bartel, 2004; Meyers *et al.*, 2008) are rendered by $map()$ functions with specified rules as parameters. The precursors longer than 300 nt or containing a second loop that is larger than 5 nt are discarded (Axtell et Meyers, 2018). Finally, a dominantly expressed potential miRNA within the pre-miRNA candidate sequence is considered a true miRNA and is subjected to subsequent functional analysis. Most of the parameters are easily adjustable as needed. Since the parameters are optimized for plants and can affect the efficiency and the accuracy of the pipeline, users do not need to change them unless suggested by the experts.

Several incorporated programs, such as Bowtie, already support parallelization, and indeed are not a time-consuming step. Therefore, the acceleration by our implementation is to parallelize and chain every single prediction step in one run. For example, RNAfold is parallelized within mirLibSpark, and giving mirCheck the resulting structure as input is also taken care of by mirLibSpark.

4.3.3 Additional annotation modules

Genomic annotation (M1). To enhance user's experience, mirLibSpark already supports seven plant species of small and large genomes: *Arabidopsis thaliana* (ath), *Triticum aestivum* (wheat), *Zea mays* (corn), *Oryza sativa* (rice), *Solanum tuberosum* (potato), *Brachypodium distachyon* (stiff brome) and *Aegilops tauschii* (Tausch's goatgrass). The project includes a script for user to fetch the genomes and annotations from Ensembl plants (http://plants.ensembl.org/) (Bolser *et al.*, 2017) and to build automatically Bowtie indices for supported plant species. The genome-wide annotations of KEGG pathways are retrieved from KEGG database (Kanehisa *et al.*, 2017).

Target gene prediction (M3). The best 100 target genes are predicted by miRanda (John *et al.*, 2004), parallelized in mirLibSpark. Gene variants are consolidated to a single identifier. Then the best 5 scored identifiers are subjected for functional analysis.

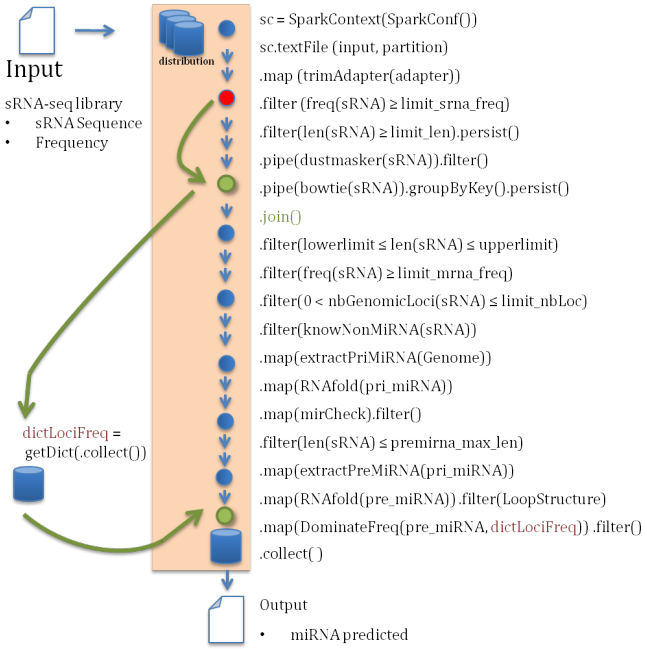


Figure 4.2: MirLibSpark miRNA prediction module.

Differential expression analysis (M4). According to the guide file, comparisons are performed between libraries. Expression fold change for each miRNA candidate is calculated. Then Kal's test (Kal *et al.*, 1999) is performed to examine the difference between samples by calculating a *z*-score and a *p*-value for each miRNA. The *p*-value is corrected by false discovery rate (Benjamini et Hochberg, 1995). Only the miRNAs with fold changes greater than 2 or less than 0.5, and corrected *p*-values smaller than 0.05 are considered differentially expressed.

KEGG pathway annotation (M6) and functional enrichment analysis (M7). Target genes of differentially expressed miRNAs in the libraries are annotated with KEGG pathways by KEGG pathway annotation module (M6). Pathway enrichment is performed by hypergeometric analysis, similar to other miRNA functional analysis software mentioned in (Vlachos *et al.*, 2015). The chance of all the drawing combinations of the genes to obtain the sample frequency of a function based on the background frequency is the *p*-value. Functions with *p*-values smaller than 0.05 are considered enriched.

4.3.4 Execution

Configuration options. Tunable parameters and their default values are accessible in mirLibSpark, including the minimum frequency of sRNA and miRNA, minimum length of sRNA, length range of miRNA, maximum

length of pre-miRNA, maximum number of genomic loci, precursor searching range, extra flank length to the validated precursor for folding visualization, temperature used for RNAfold, and parameter options for miRCheck, miRanda and Spark. Several parameters are related to the choice of species, such as genome, index, target file and pathway annotation. If a supported species is chosen, these related parameters are automatically set accordingly. In the case of custom species, users will provide the files and update the parameters.

Inputs. MirLibSpark takes as input one to several sRNA-seq libraries of the same input format. Four input formats are supported: tsv (readcount), reads, FASTA and FASTQ. A guide file (*diffguide_file*) defining the control-experiment pair is required if functional analysis modules are activated for a multi-library task. As an example, for a task analyzing two libraries (Lib1 and Lib2), if the user wishes to define that Lib1 is the control for Lib2, the guide file has to be provided by the user and formatted as follows.

```
Experiment->Control
Lib2->Lib1
```

Outputs. For any request, it would generate the first 10 of the output files as listed in the manual. For a request processing more than two libraries, if differential expression analysis parameter (DF) is set true, then two more outputs are produced; if KEGG pathway enrichment analysis parameter is set true, with the condition that DF is true, then one more output is produced.

4.4 Preparations for benchmarking

4.4.1 Construction of simulated data sets

Positive data set. The positive data set (PD) is the entire high confidence *Arabidopsis thaliana* miRNAs of miRBase version 21 Kozomara *et al*. (2018). The same miRNA sequences may have different precursors and are registered as different miRNAs. Only unique miRNA sequences were collected for this study. In version 21, there are 349 unique miRNA registries for *Arabidopsis*, and a subset of them are high confidence (100).

Negative data set. The negative data set (ND) was constructed as follows. The universe of non-miRNA is presumably larger than true miRNA. Therefore, the number of ND sequences is 1,000 by applying an arbitrary factor of 10 on the number of PD. The ND sequences were generated by extracting sequences

from *Arabidopsis* genome at random positions with lengths between 18 and 25 nt (inclusive). Sequences found in miRBase v21 *Arabidopsis* miRNAs were excluded in the ND collection. Sequences annotated as CDS, rRNA, snoRNA, snRNA and tRNA in TAIR10 were also excluded in the ND collection.

4.4.2 sRNA-seq libraries

We evaluated the behavior and the performance of mirLibSpark with real sRNA-seq libraries. GSE44622 series (27 libraries in total) (Jeong *et al.*, 2013), GSM518432, GSM738727 (Lei et Sun, 2014) and GSM1087974 *Arabidopsis* sRNA-seq libraries were obtained from NCBI Gene Expression Omnibus (GEO) repository. The *Triticum aestivum* sRNA-seq library, S002B35, was obtained from Li et al (Li *et al.*, 2018).

4.4.3 Implementation of a sequential version

The python script, $sequential.py$, is a prediction pipeline that executes in memory the major prediction steps as in the mirLibSpark but without Spark data management or parallelism.

4.4.4 Implementation of a miRDeep-P wrapper

The miRDeep-P predictor is semi-automated and requires users to execute several scripts and commands one after another. We implemented the python script, $run_mirdeep_p.py$, to execute the prediction of known and novel miRNAs automatically with the same parameters in mirLibSpark.

4.4.5 miRNA clustering

Similar sequences in the predicted miRNA set are clustered as a group based on the bitscore (greater than 20) given by blastn alignment in blastn-short task mode (Altschul *et al.*, 1997). Bitscore indicates a log2-scaled estimation of the search space that is required to find a score as good as or better than this one by chance. When the bitscore is 20, we would have to search about $2^{20} =$ one million sequences to find a sequence similar to our query by chance.

4.5 Benchmarking and discussions

4.5.1 Improved execution time

The performance of several plant miRNA predictors is compared with mirLibSpark (shown in Table 4.1). In the simulated *Arabidopsis* data sets, we incorporated 100 miRBase v21 high confidence miRNAs as positive data set. We extracted 1,000 unannotated random short sequences from the genome of *Arabidopsis* to construct the negative data set. Execution time and prediction performance of mirLibSpark and the other tools were assessed by the processing of the simulated data sets with specified number of cores. For the real sRNA-seq NGS data, GSN518432 was used (Lei et Sun, 2014). MirLibSpark incorporates several miRNA structure and expression criteria (Meyers *et al.*, 2008; Axtell et Meyers, 2018). Therefore, it has better F1 score and accuracy measure than ShortStack on the simulated data sets. Its processing of the *Arabidopsis* sRNA-seq library is 3.6 times faster than that of miR-PREFeR.

4.5.2 Comparison of performance with other predictors

We ran the programs of mirLibSpark, mirDeep-P, ShortStack and mirPrefer to process the libraries of GSM518432 and GSM738727 and obtained the miRNA predictions of each program, respectively. Axtell and Meyer (Axtell et Meyers, 2018) used a statistic method that is different than ours for Table 4.1 for assessing the performance. In order to compare their assessments of plant annotation software with mirLibSpark, we adapted their statistic method. We thus defined the positive set as the miRBase-registered miRNAs (P). The true positive predictions were the predicted miRNAs (TP), while the false negative predictions (FN) were the remaining instances in P. The negative set (N) was defined as the miRBase-registered miRNAs not expressed in the library, and thus the same number across all programs. The performance measures, including false positives, sensitivity, precision and accuracy, are shown in Figure 4.3. We also overlaid the corresponding data points of the assessments by (Axtell et Meyers, 2018) in our figures, labeled as *Lei1* and *Lei2*. Because of the differences of the parameters used in the programs as well as the versions of the genome, RNAfold and so on, our assessments were not the same as theirs.

Table 4.1: Performance of plant miRNA annotation software using (A) simulation data, (B) real world data adapted from (Lei et Sun, 2014)

(A)	mirLibSpark	miR-Deep-P	ShortStack
nbCore	32	1	1
Time (sec)	135	194	389
TP	84	69	22
FP	9	57	0
F1	0.87	0.61	0.36
MCC	0.86	0.57	0.45

(B)*	mirLibSpark	ShortStack	mirPrefer	mirDeepP
nbCore	32	1	32	1
Time (sec)	120	2191	577	74601
TP	39	31	56	47
TN	757514	757575	757247	757469
FP	54	1	304	91
FN	151	159	134	143
F1	0.276	0.279	0.204	0.287
MCC	0.293	0.398	0.214	0.29

nbCore: number of cores allocated. Time: execution time in seconds. TP: true positives. FP: false positives. F1: F-measure. MCC: Matthews correlation coefficient. *: Used the standard method to define positive and negative sets in the library, where negative set were the sequences in the library not registered in miRBase, and thus the FP predictions could be novel miRNAs.

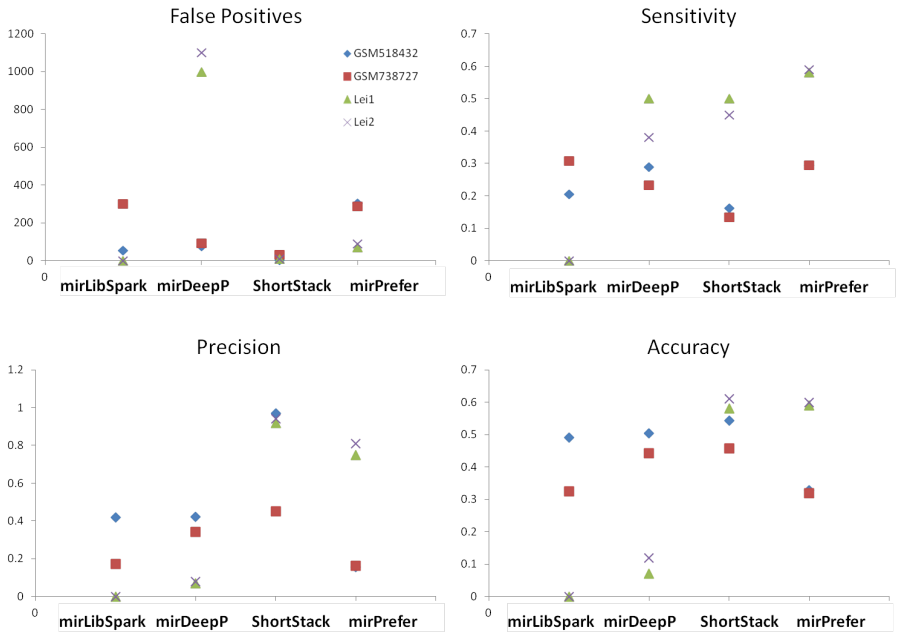


Figure 4.3: Comparative performance of miRNA annotation software from plants.

4.5.3 Scalability

MirLibSpark accelerates according to the number of core processors (nbCore). As the nbCore increases, the time of execution decreases and the demand for memory space remains the same for the same input. In addition, mirLibSpark is capable of processing large genome, such as wheat (see Table 4.2). It reduced the time of processing of a wheat library from more than 30 hours by the non-parallelized program *sequential.py* to 2.2 hours by mirLibSpark with 32 cores.

To assess scalability independently of raw multiprocessing, we performed a single-core comparison across three execution modes (Table 4.2): *sequential.py*, a fully sequential implementation that follows the same logic as mirLibSpark without Spark; *run_mirdeep_p.py*, a single-command execution of miRDeep-P; and mirLibSpark constrained to one CPU. Under single-core execution, *sequential.py* already outperforms *run_mirdeep_p.py* by reducing redundant computation and intermediate disk I/O. When further restricted to one CPU, mirLibSpark achieves additional gains through more efficient workflow organization and task orchestration, indicating that its scalability advantage derives from workflow design rather than increased core usage alone.

Table 4.2: Scalability of mirLibSpark for sRNA-seq data from small and large genomes. The input library for panel (A) is *Arabidopsis* GSM1087974 of 49 MB with genome 117 MB; for panel (B) is *T. aestivum* S002B35 of 2.1 GB with genome 13 GB. *: The program was allocated 30 hours due to limited resources and terminated before reaching the end of the program. The results of time and memory are the average of three executions of the same settings. Done in Graham computing cluster of Compute Canada with Intel E5-2683 V4 CPUs, running at 2.1 GHz and SATA SSD.

(A) ath sRNA-seq lib	nbCore	Time (sec)	MEM (GB)
sequential.py	1	12396	0.8
run_mirdeep_p.py	1	47752	0.7
mirLibSpark	1	2980	14.1
	2	1472	13.2
	3	1071	13.6
	4	837	12.0
	8	501	12.3
	16	270	11.4
	32	215	10.2
(B) wheat sRNA-seq lib	nbCore	Time (sec)	MEM (GB)
sequential.py	1	> 30 hrs *	15.3
mirLibSpark	8	23561	60.3
	16	13476	63.5
	32	8057	62.0

4.5.4 Case study

The miRNA prediction was reproduced for 27 *Arabidopsis* sRNA libraries in GEO Series GSE44622 using mirLibSpark. The computation of differential expressions of the predicted miRNAs and the enriched KEGG pathways of their target genes was guided for twenty experiment-control pairs as defined in a *diffguide_file*. The entire process, including both prediction and functional annotation, took about 2 hours on Compute Canada Graham computing cluster (https://docs.computecanada.ca/wiki/Graham). The miRNA prediction part alone took mirLibSpark 5,962 seconds, or 99 seconds per library on average. After clustering the predicted miRNAs based on sequence similarity determined by blastn, the clusters of mirLibSpark prediction

were compared with those clusters of the original paper and the miRBase-v21 high/low confidence, using Venn digram as shown in Figure 4.4.

Our results show high degrees of overlapping predictions with Jeong et al (Jeong *et al.*, 2013) and the known miRNAs. Based on our simulation results, the non-overlapping clusters predicted by mirLibSpark (175 clusters) are possibly valid novel miRNAs. However, they used the old version of the prediction criteria (Meyers *et al.*, 2008) while we added the updated criteria (Axtell et Meyers, 2018). They removed some predicted miRNAs manually according to additional criteria that were not performed by mirLibSpark. For example, if the sum of abundance in the 27 libraries is less than 400, or the genomic loci of the precursors are in siRNA-like regions, the miRNAs are removed (Jeong *et al.*, 2013). Although we filtered siRNAs using predefined genomic annotations, we did not identify genomic regions that are likely siRNA precursors. Therefore, among our novel miRNAs collections, a fraction of functionally irrelevant predictions might not be excluded based on current criteria.

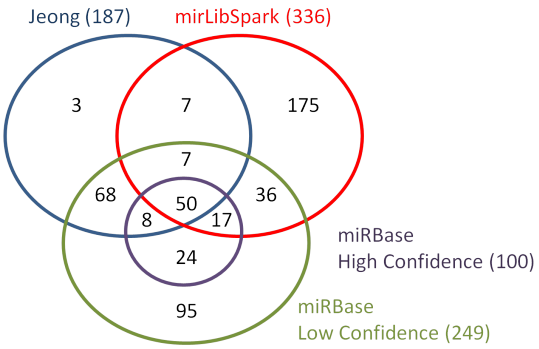


Figure 4.4: The venn diagram characterizing the predictions of GSE44622 by Jeong *et al.* (2013) and mirLibSpark, after a clustering step using blastn to group together similar sequences.

4.6 Conclusion

We facilitate the settings and annotations of several procedures, including the retrieval of the reference genome, building genomic indices, and installing dependency programs and libraries. MirLibSpark takes different input formats. Most of the validation parameters are flexible in order to meet species-specific preferences or criteria updates. This will allow a quick deployment and ready to use environment. The user manuals for Docker and computing cluster environments, respectively, are available in GitHub.

We introduced the simulated data sets to properly assess the performance with statistical measures. We

showed an improved accuracy on the prediction. It is scalable regarding multiple libraries and libraries from large genomes. The case study showed high degrees of overlapping with the prediction by others. One can easily reproduce results among research teams applying the stand-alone, dependency-self-supporting, and fully automated mirLibSpark.

In the future, all the added dependencies will be updated regularly. The miRNAs predicted using mirLibSpark in wheat and other plant libraries will be collected in Wheat MicroRNA Portal (Remita *et al.*, 2016). We will also investigate how to standardize the exclusion procedures in real-world data, such as identifying the miRNA precursors residing in the putative siRNA genomic regions and the low occurrences among libraries.

4.7 Implications

Small RNA sequencing has become indispensable for investigating gene regulation in plants, yet large-scale data analysis remains a critical bottleneck. This chapter presented `mirLibSpark`, a scalable and integrated Spark-based tool that streamlines miRNA prediction and downstream analysis across multiple libraries. The system enables unified expression profiling, classification, differential expression analysis, and biological interpretation in a single, distributed framework.

This methodology fulfills the first specific objective by developing a scalable pipeline for high-throughput miRNA classification and directly supports broader efforts in small RNA biomarker research. By addressing the computational burdens posed by high-throughput sequencing, `mirLibSpark` makes large-scale biomarker screening and cross-condition analysis more feasible. The outcomes of this work include:

- A scalable and memory-efficient Spark-based pipeline for miRNA classification and downstream analysis across multi-library datasets.
- A user-friendly and portable design: easy to install with zero dependencies, runnable on both personal computers and computing clusters, and equipped with clear static visualizations for straightforward result interpretation and sharing.
- A modular computational framework adaptable to broader RNA sequencing applications beyond miRNAs.

While `mirLibSpark` significantly improves scalability and automation for miRNA analysis in plants, it remains focused on canonical biogenesis rules and fixed annotation criteria. It does not detect non-canonical

or atypical miRNA candidates outside the established plant frameworks. Although intentionally tailored for plant miRNA classification, its performance depends on adherence to community-defined criteria, which may evolve over time.

Across the computational frameworks developed in this thesis—including mirLibSpark and the downstream analyses presented in subsequent chapters— performance improvements are achieved through a consistent set of efficiency-oriented design choices rather than reliance on any single optimization technique. These include the use of high-level programming languages for rapid development and flexibility, early pruning of the search space through biologically informed rules, selection of appropriate data structures, reduction of disk I/O by retaining intermediate data in memory, and careful orchestration of parallelism. Scalability is therefore treated as a consequence of workflow organization, data flow management, and reuse of intermediate results, rather than as a simple function of increased hardware resources. These shared assumptions define the computational abstraction layer of the thesis and delineate the conditions under which efficiency gains are expected to generalize across datasets and execution environments.

Similarly, prediction accuracy across the thesis relies on a common strategy that combines strict, domain-informed filtering with post-prediction validation rather than purely data-driven inference. In the case of mirLibSpark, this includes enforcement of gold-standard plant miRNA biogenesis criteria, controlled handling of multi-mapping reads, and multiple layers of post-prediction support such as cross-library consistency, abundance thresholds, differential expression with false discovery rate control, and functional evidence from target analysis. Comparable principles—explicit assumptions about signal structure, expression support, and validation criteria—are also applied in the other projects presented in this thesis. While these assumptions reduce false positives and improve robustness, they necessarily constrain the scope of detectable signals to those that conform to current biological knowledge and annotation standards, thereby defining the conceptual boundaries of the proposed methods.

To demonstrate the methodology's broader utility, `mirLibSpark` was applied in the wheat iRNA study presented in the next chapter. There, it served as the core engine for miRNA classification and was extended to support parallelized siRNA processing, highlighting its adaptability to more complex, multi-class small RNA analyses in real-world conditions.

CHAPTER 5

PREDICTION AND EVOLUTIONARY DYNAMICS OF WHEAT MIRNAS AND SIRNAS DURING FIELD COLD ACCLIMATION

Plants face a wide range of abiotic stresses, and small RNAs play a critical role in mediating adaptive responses. Among major crops, wheat is particularly notable for its ability to survive harsh winter conditions through cold acclimation. While numerous studies have identified stress-responsive miRNAs or siRNAs, few have explored how these two RNA classes operate together during adaptation in field-grown conditions. Despite their overlapping biogenesis pathways and shared roles in gene regulation, miRNAs and siRNAs are often studied in isolation. This fragmented perspective limits our understanding of small RNA–mediated stress adaptation and fails to capture the full regulatory landscape. To address the second specific objective of the thesis, developing a computational framework to identify small RNAs associated with phenotypic variation in wheat, this chapter presents a combined computational and biological analysis that:

- Develop an integrative classification pipeline to jointly identify cold-responsive miRNAs and siRNAs from the same dataset.
- Analyze their expression, evolutionary conservation, and co-regulatory roles across wheat cultivars through differential expression, target prediction, functional annotation, and enrichment analysis.
- Infer small RNA–mediated networks to uncover coordinated functions in stress resilience.

The integrative pipeline incorporates two prediction components: mirLibSpark, a high-performance Spark-based tool developed for plant miRNA prediction and annotation, and an in-house Python script that detects siRNA-generating loci using established phasing scores. Together, these tools enable a unified analysis of small RNA classes from the same input libraries, enhancing the interpretability of co-regulatory patterns and locus-level relationships. The methodology was applied to a time-series field experiment comprising 80 small RNA libraries collected from wheat cultivars with varying *VRN-A1* expression and freezing tolerance. This analysis fills a key gap in understanding how small RNAs jointly regulate stress responses in agriculturally relevant environments.

This work was conducted within a collaborative team combining expertise in computational biology and plant stress physiology. The small RNA and mRNA datasets were generated prior to the author's enrollment at UQAM as part of a large-scale field experiment designed to study plant cold acclimation. While the ex-

perimental design and biological data generation were led by collaborators, the formulation of the research questions, the definition of the computational objectives, and the analytical strategy applied to this dataset were determined by the author. The central contribution of this work consists of the conception, design, and implementation of a unified computational framework for joint miRNA–siRNA discovery and phenotypic analysis. Chao-Jung Wu (CJW) designed and conducted the majority of the in silico experiments, defined the analytical criteria, iteratively refined tools and parameters, and synthesized the results. Given the scale and complexity of the analysis, selected computational tasks were delegated to Farzaneh Rahmdani (FR), a bioinformatics specialist, under the author's direction. Biological interpretation was informed by domain experts: Dr. Zahra Agharbaoui (ZA) contributed expertise on plant cold acclimation, and Professor Jean Danyluk provided insight into siRNA biology. Dr. Mohamed Badawi and Professors Fathey Sarhan and Jean Danyluk contributed biological perspectives. Professor Abdoulaye Baniré Diallo (ABD) supervised the doctoral research and reviewed the computational methodology. The manuscript was written and structured by CJW, with specific sections reviewed or contributed by FR and ZA. This research was carried out under the doctoral supervision of Professor Abdoulaye Baniré Diallo. The manuscript and supplementary figures and tables are included in Appendix B.

5.1 Abstract

MicroRNAs (miRNAs) and small-interfering RNAs (siRNAs) are key regulators of plant stress responses. While siRNAs share biogenesis and silencing mechanisms with miRNAs, their role in stress adaptation remains less explored. Here, we introduce a unified bioinformatics workflow to simultaneously analyze miRNAs and siRNAs involved in wheat cold acclimation. We use *interfering RNAs (iRNAs)* to collectively refer to miRNAs and siRNAs. Using a large sRNA-seq dataset from field-grown wheat cultivars with varying*VRN-A1* expression and freezing tolerance, we identified 339 iRNAs, including 96 novel iRNAs, 33 trans-acting siRNAs, 90 dual-identity iRNAs, and 66 cold response iRNAs (58 miRNAs and 9 siRNAs). To enhance prediction confidence, miRNAs and siRNAs were validated using several bioinformatics approaches, including miRBase, published wheat miRNA studies, ShortStack predictions, siRNA collections, and BLASTN analysis. Evolutionary tracing suggested that wheat miRNAs and dual-identity iRNAs likely originate from ancestral subgenomes, while siRNAs and novel iRNAs emerged more recently. siRNAs primarily derive from coding regions, whereas TE-derived siRNAs may be rapidly purged. The lack of ancestral enrichment in cold-response iRNAs suggests that recent selective pressures shaped their role in cold adaptation. Functional analysis showed that miRNAs and siRNAs jointly regulate cold adaptation by targeting transcription factors, kinases, and hor-

mone signaling pathways, and may contribute to adaptation through *Vrn-A1*-dependent vernalization. Our findings provide insights into the genomic origins and functional diversification of cold-responsive iRNAs, revealing a coordinated regulatory network governing wheat stress adaptation. This study underscores the necessity of a unified approach to small RNA analysis and offers a comprehensive framework for investigating regulatory RNA evolution and function in plants.

Keywords: cold acclimation, wheat small RNA prediction, small RNAs in plant stress response, small RNA evolution, bioinformatics pipeline for small RNAs

5.2 Introduction

5.2.1 Justification for Simultaneous Prediction of miRNAs and siRNAs

Among the RNAs detectable in next-generation small RNA sequencing (NGS sRNA-seq), miRNAs and siRNAs are two major classes of RNA effectors involved in RNA interference (RNAi) (Jones-Rhoades et Bartel, 2004; Doench *et al.*, 2003). miRNAs and siRNAs are functionally and evolutionarily intertwined, sharing biogenesis, mode of action, and regulatory mechanisms. Both are processed by Dicer-like enzymes into small RNA duplexes, loaded onto Argonaute proteins, and guide RNAi to regulate gene expression through mRNA degradation, translational repression, or chromatin modification (Jones-Rhoades et Bartel, 2004; Doench *et al.*, 2003). In addition to gene silencing, RNAi can activate target transcription (Catalanotto *et al.*, 2016; Place *et al.*, 2008). Their evolutionary connection is evident as siRNAs originally served as an ancient defense mechanism against transposable elements (TEs) and viruses, while primitive miRNAs likely evolved from siRNA-hosting intermediates precursors to provide more precise regulatory control and complexity in plants (Li *et al.*, 2011b). Certain TE-related precursors can generate both miRNAs and siRNAs (Piriyapongsa et Jordan, 2008; Li *et al.*, 2011b), and the miRNA families MIR159 and MIR319 may have originated from a phased stem-loop precursor (Li *et al.*, 2011a). Some miRNAs target the 3'-UTR and lead to mRNA upregulation, which was thought to be a functional characteristic of siRNAs Valinezhad Orang *et al.* (2014). These findings blur the boundary between these RNA classes. Furthermore, both small RNA types are involved in plant stress responses, highlighting their overlapping roles in gene regulation Baldrich *et al.* (2014); Guo *et al.* (2015); Meijer *et al.* (2022).

Given these shared features, we adopt the term **interfering RNAs (iRNAs)** as a broad category encompassing both miRNAs and siRNAs. We argue that miRNAs and siRNAs should be analyzed as a unified regulatory sys-

tem to provide deeper insights into small RNA-mediated regulation, rather than treating them as separate entities.

5.2.2 Categorizing Interfering RNAs

Canonical miRNA precursors are single-stranded RNA molecules transcribed from the genome, typically forming hairpin structures and measuring less than 300 nucleotides in length (Axtell et Meyers, 2018). In contrast, siRNAs arise from double-stranded RNAs, which can form through bidirectional transcription, complementary strand pairing, or hairpin structures similar to miRNA precursors (Axtell, 2013a). Various **siRNA subclasses** exist, arising from coding and non-coding transcripts and undergoing distinct processing mechanisms. These include natural antisense siRNAs (natsiRNAs), derived from overlapping transcripts; trans-acting siRNAs (tasiRNAs), derived from TAS genes and targeting mRNAs at different loci (Howell *et al.*, 2007); heterochromatic siRNAs (hc-siRNAs), involved in heterochromatin formation (Verdel *et al.*, 2004); piwi-interacting RNAs (piRNAs), expressed in animal reproductive cells and associated with specific proteins (Farazi *et al.*, 2008); mirtrons, processed from spliced introns (Rorbach *et al.*, 2018); and phased siRNAs (PHAS), which generate regularly spaced small RNAs (Kakrana *et al.*, 2017).

Although miRNAs and siRNA subtypes have diverse origins, they are all processed by Dicer-like proteins, making their identification feasible using bioinformatics approaches (Axtell, 2013a). Understanding these subclasses not only enhances small RNA prediction and annotation but also refines our ability to systematically analyze their regulatory roles in plant stress adaptation.

5.2.3 Current Studies on miRNAs and siRNAs in Plants

Over the past two decades, **miRNAs** have been extensively studied as key post-transcriptional regulators in plant development, stress responses, and genome maintenance (Mallory et Vaucheret, 2006; Wilusz *et al.*, 2009; Lucas et Budak, 2012; Kurtoglu *et al.*, 2013; Akpinar *et al.*, 2015; Agharbaoui *et al.*, 2015; Budak et Akpinar, 2015). Numerous studies have demonstrated their crucial roles in plant responses to abiotic stresses such as drought, heat, cold, and salinity (Liu *et al.*, 2008; Kantar *et al.*, 2011; Tang *et al.*, 2012; Wang *et al.*, 2014; Stief *et al.*, 2014; Fang *et al.*, 2014; Agharbaoui *et al.*, 2015; Ma *et al.*, 2015; Song *et al.*, 2017). These regulatory networks are finely tuned to specific species, tissue types, and environmental conditions (Xie *et al.*, 2014; Han *et al.*, 2014; Agharbaoui *et al.*, 2015; Song *et al.*, 2017). The availability of curated databases like miRBase (Kozomara *et al.*, 2018) has further facilitated systematic studies of miRNA-mediated

gene regulation, enabling comprehensive comparative analyses across plant species.

In contrast, **siRNAs** remain comparatively underexplored, largely due to the lack of widely adopted computational tools and dedicated resources similar to those for miRNAs (Lukasik *et al.*, 2016). While challenges persist, several studies have sought to elucidate the genomic origins and regulatory roles of siRNAs in plants. In *Arabidopsis*, tasiRNAs (Chen *et al.*, 2007; Howell *et al.*, 2007) and natsiRNAs (Zhang *et al.*, 2012b) have been shown to mediate transcriptional and post-transcriptional silencing through highly coordinated regulatory mechanisms. The discovery of 24-nucleotide (nt) phased siRNAs in maize anthers (Zhai *et al.*, 2015) revealed tissue-specific PHAS-24 siRNA networks that play essential roles in reproductive development. Interestingly, this strategy is similar to piRNAs in animal reproductive cells, where they associate with Argonaute/Piwi proteins to mediate mRNA cleavage (Farazi *et al.*, 2008).

Although several specialized siRNA databases exist, no centralized repository provides systematic and continuously updated siRNA annotations akin to miRBase for miRNAs (Kozomara *et al.*, 2018). The MPSS database archives plant PHAS loci from *Arabidopsis*, maize, and rice (Meyers Lab, 2019; Kakrana *et al.*, 2017), while Penn State University's plant small RNA genes repository (PSU Catalogue) compiles siRNA predictions across 47 plant species (Lunardon *et al.*, 2020; Axtell, 2013b). However, these collections are fragmented, and comprehensive *in silico* predictions for many plant genomes, including wheat, remain largely incomplete.

The absence of a standardized siRNA reference hinders large-scale comparative analyses and complicates the development and validation of computational models for siRNA classification. As a result, small RNA classes are often analyzed in isolation rather than as interconnected components of a broader regulatory network.

5.2.4 Wheat as a Model for iRNA Evolution and Stress Responses

As miRNAs and siRNAs play key roles in plant stress responses (Khraiwesh *et al.*, 2012), their thorough identification and target analysis are essential for uncovering the intricate regulatory networks that govern physiological processes. Wheat (*Triticum aestivum* L.), a major global staple crop, is an ideal model for studying iRNA-mediated stress responses due to its polyploid genome and the variation in freezing tolerance among cultivars. As a hexaploid species (2n=6x=42, AABBDD) resulting from the divergence and hybridization of three progenitor genomes, wheat offers a unique system for exploring small RNA evolution. Its complex genome facilitates comparative analyses of regulatory conservation, divergence, and functional innovation

across homologous gene networks (Marcussen *et al.*, 2014).

Although siRNAs have been less studied, recent research highlights their regulatory roles in wheat stress responses and genome stability. For instance, Yao *et al.* (2010) demonstrated that non-coding RNAs, excluding miRNAs, are upregulated under abiotic stress, guiding mRNA degradation to modulate stress-responsive pathways. Tang *et al.* (2012) extended this finding by identifying a complex network of siRNAs involved in wheat's response to cold stress, refining the differential expression of genes essential for stress tolerance. Sun *et al.* (2013) reported abundant TE-associated siRNAs in developing wheat grain, suggesting their involvement in gene regulation and genome stability maintenance. Kenan-Eichler *et al.* (2011) found that hybridization and polyploidization events lead to siRNA deregulation, influencing gene expression patterns that affect wheat's environmental adaptability.

Moreover, the availability of cold-acclimating and non-acclimating cultivars enables the dissection of small RNA-mediated regulation in stress adaptation, further establishing wheat as an exceptional model for investigating the role of small RNAs in both adaptive responses and evolutionary dynamics.

5.2.5 Computational Challenges in Small RNA Prediction in Wheat

Despite advancements in small RNA research, predicting miRNAs and siRNAs in wheat remains difficult due to the genome's large size (13 GB), high redundancy, and polyploid complexity (Appels *et al.*, 2018). Effective small RNA classification requires predictive models that integrate knowledge of their biogenesis, functional roles, and regulatory mechanisms and are computationally scalable for large genomic and transcriptomic datasets.

While miRNA annotation benefits from well-defined structural criteria (Axtell et Meyers, 2018), siRNA prediction remains constrained by the structural diversity of siRNA precursors and the lack of dedicated tools. To address both miRNA and siRNA prediction constraints, we integrate phase scoring, which detects siRNAs generated through consecutive Dicer digestion cycles (Howell *et al.*, 2007), and an established miRNA predictor (Wu *et al.*, 2019), into a scalable computational framework (Zaharia *et al.*, 2010) to ensure efficient and accurate small RNA identification.

Capturing the full complexity of small RNA dynamics in wheat requires a biologically meaningful dataset that accounts for genetic variation, temporal changes, and real-world environmental conditions. We lever-

age a large sRNA-seq dataset from wheat cultivars with varying freezing tolerance, sampled across warm and freezing months. By analyzing field-grown plants, we ensure our findings reflect miRNA and siRNA regulation under natural stress challenges.

5.2.6 Major Contributions

In this study, we present an integrated bioinformatics and statistical workflow for predicting miRNAs and siRNAs, analyzing their differential expression, and characterizing their functional roles in wheat cold acclimation. This workflow addresses key limitations in small RNA analysis by incorporating a scalable miRNA prediction tool for wheat's large genome, integrating a well-accepted scoring method for siRNA precursor identification, and unifying the analysis of miRNAs and siRNAs, which are often studied separately regardless of their shared biogenesis and functional interplay. The workflow was deployed on a large sRNA-seq dataset of wheat grown under natural field conditions. We performed a statistical functional analysis from late summer to the freezing season on winter and spring cultivars that differ in their *VRN-A1* gene expression levels and their ability to develop freezing tolerance, using the same wheat samples for which we already understand their cold response at the mRNA level (Li *et al.*, 2018). The primary contributions of this study include:

1. **Identification of Dual-Identity iRNAs** – Our dual-predictor approach revealed 90 iRNAs with both miRNA and siRNA characteristics, suggesting an evolutionary and functional link between the two RNA classes.

2. **Discovery of Novel iRNAs** – We identified 96 previously unreported miRNAs and siRNAs, expanding the known small RNA repertoire in wheat. Some novel iRNAs regulate cold responses, underscoring their potential role in stress adaptation.

3. **Genomic Origins of iRNAs** – We traced the genomic loci that generated miRNAs and siRNAs into distinct categories, including ancestral genomes, coding regions, TAS homologs, and transposable elements. Their distribution patterns provide insights into their evolutionary trajectories and regulatory roles.

4. **Identification of cold-response iRNAs** – We identified 66 cold-response iRNAs with distinct expression patterns across wheat cultivars, revealing their involvement in stress adaptation. These iRNAs

regulate key genes associated with cold tolerance, including kinases and transcription factors, and exhibit cultivar-specific regulatory roles.

5. **A Unified Small RNA Analysis Framework** – We developed a comprehensive bioinformatics workflow that integrates miRNA and siRNA predictions, differential expression analysis, and functional annotation.

This work serves as both a case study and a methodological framework for future small RNA research in complex plant genomes. By demonstrating the feasibility and necessity of integrating miRNA and siRNA analysis, we highlight the evolutionary significance and functional interplay of these small RNAs, thereby providing a more systematic methodology for investigating small RNA-mediated stress responses in plants.

5.3 Materials and methods

5.3.1 Small RNA preparation from wheat field samples

Four wheat cultivars that differ in their freezing tolerance and vernalization responses – winter-habit 'Norstar' (vrn-A1/vrn-A1), spring-habit 'Manitou' (Vrn-A1/Vrn-A1) and their reciprocal near isogenic lines (NIL), spring-habit Norstar 'Iso8S' (Vrn-A1/Vrn-A1) and winter-habit Manitou 'Iso11W' (vrn-A1/vrn-A1) – were grown under field conditions in two different years and were sampled from autumn to just before the freezing season. All four wheat cultivars were planted in late summer. Winter-habit wheat would undergo vernalization, a cold-promoted flowering process, and spring-habit wheat would eventually die from cold over winter Li *et al*. (2018). Air and soil temperature data are presented in the supplemental data section (See S1). Crown tissues were sampled and at least five crowns were used for each RNA extraction. Samples were ground to a powder in dry ice and total RNA was extracted using the mirVana™ miRNA Isolation Kit (Life Technologies). RNA quantity and quality were analyzed using NanoDrop spectrophotometer and Agilent 2100 bioanalyzer system, respectively. Small RNA libraries were constructed using the TruSeq Small RNA Sample Preparation Kit (Illumina) and the sequencing was performed on Illumina HiSeq 2000 platform. Each cultivar had 20 libraries, and a total of 80 sequencing libraries were constructed, representing sRNAs of two biological replicates at five time points between the beginnings of September and November from each of the four cultivars grown in two years, 2010 and 2013 in Saskatoon (see Li *et al*. (2018) for details of cultivar genotypes, field conditions and sample collection).

5.3.2 Bioinformatics approach for the identification of miRNAs and siRNAs in wheat cold response

We investigated wheat cold acclimation in the field by integrating several bioinformatic and statistical methods in the analysis workflow (Figure 5.1). We will describe the workflow briefly here as well as in detail for each incorporated method in the next sections. Sequencing data were generated by our collaborators. We performed quality control, prediction of interfering miRNAs and siRNAs from expressed sequences, differential expression, differential regulation and functional analysis among time series and cultivars. We incorporated the methods that we developed, including miRNA and siRNA predictions, removing biases, identification of cold response iRNAs, and we annotated the genes expressed in the same libraries for target gene analysis. Using two de novo prediction pipelines, 215 miRNAs and 127 siRNAs from the wheat field libraries were reported in this study. We used a large dataset consisting of 80 libraries which enabled us to predict iRNA candidates and eliminate noise and batch effects. To predict miRNA, we used miRLibSpark that embedded the gold-standard requirement for plant miRNA prediction. To predict siRNAs, we incorporated the H-score (Howell *et al.*, 2007) to search sustained phasing potentials in the genomic landscape. To understand the functions of the predicted miRNA and siRNA from wheat field samples, we studied their expressions. We devised an encoding scheme to translate differential expression profiles of spring and winter cultivars into cold response regulation. Differential regulation analysis among time series and cultivars showed 66 cold response iRNAs.

5.3.3 Identification of miRNA and siRNA candidates from sequenced small RNAs

These analyses were performed on the 80 sequencing libraries. Each library was checked for quality using FastQC (Andrews *et al.*, 2010). The adapters were removed using cutadapt 0.9 (Martin, 2011). Afterwards, putative iRNAs were predicted using miRNA prediction and siRNA prediction procedures illustrated in Figure 5.1 and described in the following paragraphs.

5.3.3.1 MiRNA prediction

MiRNA precursors are single-stranded RNA molecules expressed from the genome and exhibiting several structural features. They are cleaved by Dicer on the double-stranded hairpin into miRNA/miRNA* duplexes. The consensus determining features of a valid miRNA are specified recently (Meyers *et al.*, 2008; Axtell et Meyers, 2018). However, limited choices of the existing miRNA predictors (Axtell, 2013b; Lei et Sun, 2014) are capable of predicting novel and conserved miRNAs derived from organisms of big genomes using reasonable

time and computing resources, such as from wheat having 13 Giga base pairs of the genome. We designed and implemented mirLibSpark (Wu *et al.*, 2019), a scalable NGS plant miRNA prediction pipeline. Its miRNA prediction module compiles with the gold-standard requirement on plant miRNA predictions (Meyers *et al.*, 2008; Axtell et Meyers, 2018). The computational complexity has been optimized in mirLibSpark so that it takes only 2 to 3 hours to predict miRNAs from an average wheat sRNA library. The operations in mirLibSpark for this study are as follows. Low complexity sRNAs were removed using dustmasker (Morgulis *et al.*, 2006). Reads with size between 21 and 24 nts (Axtell et Meyers, 2018) were retained and mapped to the reference genome using Bowtie allowing zero mismatch and up to 15 mapped genomic loci. A set of known non-miRNA sequences (tRNA, rRNA, snoRNA) according to the annotation of the genome were removed. Small RNAs with at least 100 reads were chosen for folding predictions. For each sRNA, we considered two sequences that could include a potential pre-miRNA candidate as follows: 200 nts before the start and 500 nts after the end of the mapping; 500 nts before the start and 200 nts after the end of the mapping on the genome. The secondary structures of the extracted sequences (pri-miRNA, about 700 nts) were folded with RNAfold from the ViennaRNA v2.4.11 package (Lorenz *et al.*, 2011) to identify those having a hairpin-like shape, one of the fundamental characteristics of pre-miRNAs, using miRCheck (Jones-Rhoades et Bartel, 2004) and modified parameters complying with Meyers *et al.* (2008). Then, predicted pre-miRNA sequences were excised from the pri-miRNAs and resubmitted to RNAfold with an optimized hairpin and repredicted again using miRCheck. Precursors longer than 300 nts or containing a second loop that was larger than 5 nts were discarded (Axtell et Meyers, 2018). Finally, a dominantly expressed potential miRNA within the pre-miRNA candidate sequence was considered a true miRNA (Axtell et Meyers, 2018). This setting showed an improved accuracy for miRBase Arabidopsis annotations (87%) (Kozomara *et al.*, 2018; Wu *et al.*, 2019) and could identify the majority of the IWGSC-annotated miRNAs (52.5%) (Appels *et al.*, 2018; Rahmdani *et al.*, 2020). IWGSC stands for the International Wheat Genome Sequencing Consortium.

5.3.3.2 SiRNA prediction

Endogenous siRNAs are the major component of sRNA populations in plants (Liu *et al.*, 2014). SiRNAs are generated from Dicer's cutting of dsRNA precursors and the dsRNA region can come from either a single RNA forming hairpin, also known as a miRNA precursor or a dsRNA duplex, also known as a siRNA precursor (Howell *et al.*, 2007; Chakravarthy *et al.*, 2010; Kakrana *et al.*, 2018). Since Dicer dices along the dsRNA substrate one cycle after another until riding off the substrate, one to several cycles of phased siRNAs can be generated and a cluster of siRNAs is thus observed as regularly-spaced 21-nt or 24-nt duplex populations

aligned on the precursor genome (Howell *et al.*, 2007; Chen *et al.*, 2007). The following characteristics (Axtell, 2013b) of a valid siRNA precursor have been included in our analyses: expression of the precursor transcript; specific sizes of siRNAs originating from the precursor; strength of the phasing score from the precursor. We adopted a generalized phase score Howell *et al.* (2007) to calculate the potential of a selected-size sRNA (21 nts or 24 nts) being a siRNA at every sRNA-aligned position, considering both sRNA abundance and number of phased positions on the precursor. The phase score H-score Howell *et al.* (2007) to quantify the phasing potential of a given genomic position based on sRNA expressions was defined as follows. Given l, $\mathbf{G}$ and $\mathbf{L}$. l the PHAS size, $l \in \{18, 19, ..., 30\} \in \mathbb{N}^+$. $\mathbf{G}$ the genome, $\mathbf{G} = \{\mathsf{A}, \mathsf{C}, \mathsf{T}, \mathsf{G}\}^r$, and $r \in \{10^6, \ldots, 10^{10}\} \in \mathbb{N}^+$. $\mathbf{L}$ the ensemble of S the sRNA sequences in a library, $\mathbf{L} = \{S_1, S_2, \ldots, S_n\}$, where $S = (\alpha, q)$, and $\alpha = \{\mathsf{A}, \mathsf{C}, \mathsf{T}, \mathsf{G}\}^r$, $r < 50 \in \mathbb{N}^+$, $q \in \mathbb{N}^+$. Phase Score P_j for locus j on $\mathbf{G}$:

$$P_j = ln\left(1 + \sum_{i=1}^{8} K_{j+(i-1)\times l}\right)^{n-2}, P \geq 0$$

where $K_a \in \{\mathbb{N}^+, 0\}$ is the sum of the frequency q of any $S \in \mathbf{L}$ that is aligned to locus a on positive strand or locus $(a-2)$ on negative strand, and $S.r = l$; n the number of $K > 0$, $n \leq 8 \in \mathbb{N}^+$. Because canonical miRNA precursors, by definition, host only one pair of miRNA duplex, we set the parameters in H-score to predict siRNA precursors that at least 3 cycles of Dicer cuts are observed within 8 cycles. We thoroughly surveyed the architecture of the genome by developing a chromosome-walk algorithm. The operations of siRNA prediction procedure for this study are as follows. Low complexity sRNAs were removed using dustmasker (Morgulis *et al.*, 2006). Reads were retained if they mapped to the reference genome using Bowtie (Langmead *et al.*, 2009) allowing zero mismatch and up to 15 mapped genomic loci. After alignment, each sRNA was assigned to a CONTIG. The genomic CONTIGs were indexed from the whole genome, each spanning 5250 nts with 250 nts overlapping the CONTIGs before and after itself. For example, 6D__285740000_285745250 (or 6D__285740000 for short) is a CONTIG on chromosome 6D between positions 285740000 and 285745250. A CONTIG was retained if at least 3 sRNA species and at least 100 of total reads were aligned on the CONTIG to verify if the precursor was expressed, and at least one position had a phasing potential H-score greater than zero for the selected size of siRNAs. A putative precursor sequence was excised from the positive strand of the double-stranded CONTIG between the first and last Dicer cuts with additional flanking nucleotides. Random degradation may appear and such precursor candidates and the siRNAs derived from them were detected and excluded manually. Finally, all the sRNAs that had a phasing potential on the predicted precursors were considered functional siRNAs.

5.3.4 Sequence analyses for precursor and mature iRNAs

Each precursor is part of a CONTIG in our genomic indexing system based on genomic coordinates. The mature iRNAs were clustered together if they came from the same precursor CONTIG. Nucleotide sequence similarity analysis for the precursors was performed using bowtie (Langmead *et al.*, 2009) against the mRNA transcripts collected from the same field samples for sRNA analysis, designated XLOC database (Li *et al.*, 2018), as well as using BLASTN against the NCBI Nucleotide database with default parameters (Boratyn *et al.*, 2013). The annotations of the hits are the potential functions of precursors.

MiRBase family analysis for the mature sequences was performed using the miRFinder tool on Wheat MicroRNA Portal (with parameters as follows: Min sequence length 18, Min identity percentage 80, Min query coverage 0.9, Min subject coverage 0.9, Max of gaps 0) (Remita *et al.*, 2016) against miRBase v22 database (Kozomara *et al.*, 2018). Maximum of 2 mismatches were allowed. Hits were retained if they matched with the miRNAs of the 6 monocot species wheat *Triticum aestivum* or plants *Oryza sativa*, *Hordeum vulgare*, *Arabidopsis thaliana*, *Brachypodium distachyon* and *Zea mays*. A local version of microRNA Finder (Remita *et al.*, 2016) was used to search against additional wheat databases that were not available on the portal, such as literature collections (Agharbaoui *et al.*, 2015; Rahmdani *et al.*, 2020), IWGSC-annotated miRNAs (Appels *et al.*, 2018), collections based on ShortStack (PSU catalogue (Lunardon *et al.*, 2020)) and the predictions using ShortStack (Axtell, 2013b) on our 80 wheat libraries. Wheat is a hexaploid having 3 sets of genomes, denoted as A, B and D genomes. An iRNA was defined as an ancestral iRNA if it appeared in hexaploid (AABBDD) or in tetraploid (AABB) genomes because it suggests a more ancient origin (Marcussen *et al.*, 2014). Sequences originating from wheat coding regions were defined as mapping to the positive or negative strand of the expressed XLOC sequences. Sequences originating from wheat transposable elements and long terminal repeats loci were defined as mapping to the positive or negative strand of IWGSC-annotated coordinates, respectively. Considering both strands was to harmonize the siRNAs originating from both strands.

5.3.5 Expression statistical analyses of iRNAs

To perform differential expression analysis on the expressed iRNAs in 80 wheat libraries, we used DESeq2 (Love *et al.*, 2014; Hara *et al.*, 2017) because it provides a quantitative analysis for studies based on small replicate numbers and large dynamic ranges. It incorporates Bayes-based shrinkage estimation for dispersions and fold changes to improve the stability and interpretability of estimates. Read frequencies of the

expressed iRNAs prior to normalization and their metadata (cultivar, year, time point, replicate and soil temperature) were used as input for DESeq2. It normalized read frequencies based on the input read frequency population. We considered time point 1 of each experiment (for each year and each cultivar) as the baseline, and compared time points 2, 3, 4 and 5 with time point 1 of the same experiment. The significant difference in the normalized read frequencies between time point 1 and the four remaining time points in each cultivar was computed. An adjustment for multiple comparisons based on the false discovery rate (FDR) (Benjamini et Hochberg, 1995) was performed and iRNAs with FDR $\leq$ 5% were kept. To identify read frequencies that were inconsistent between the two biological replicates (r1 and r2), we used the gauge formula: [if r_1 and r_2 do not fulfill $abs(\frac{r1-r2}{2}) \leq (\frac{r1+r2}{2}) \times 20\%$, they are inconsistent], where abs is the absolute value. Most inconsistent entries were related to libraries Iso11W-2010-t3 and Iso11W-2013-t2. Finally, iRNAs with fold change (FC) lower than 0.5 or higher than 2.0 (ie. abs(log2FC $\geq$ 1)) based on consistent replicate entries were considered differentially expressed.

5.3.6 Assigning regulated iRNAs in genetic backgrounds

The differentially expressed iRNAs were grouped into cultivars Iso11w, Norstar, Iso8s and Manitou. Each group contains 20 libraries of the same cultivar collected at five time points in 2010 and 2013, respectively. If the expression of a RNA species was differential in only a cultivar, it was regulated specific to that cultivar. If a RNA species was differentially expressed in both spring habit cultivars (Manitou and Iso8S) but not in either winter habit cultivars (Norstar or Iso11W), it was spring-habit specific; and the other way around was winter-habit specific (Pan *et al.*, 2022). Similarly, we defined a RNA species as Manitou-background specific if it was commonly differentially expressed in both Manitou-background cultivars (Manitou and Iso11W) but not in either Norstar-background cultivars (Norstar or Iso8S); and the other way around was Norstar-background specific.

5.3.7 Differential regulation analysis based on iRNA expression patterns during field cold acclimation

We classified each iRNA by assessing its expression profile with an encoded vector clustering approach built on the time course expression in paired winter and spring cultivars. To do so, we first compared the expression of an iRNA in a cultivar at time points 2, 3, 4 and 5 with time point 1 of each year, and noted up- (u), down- (d) and not-regulated (0) according to fold change $\geq$ 2, $\leq$ 0.5 and others, respectively. For example, the pattern for an RNA X in a winter cultivar in a year through time points 2 to 5 was noted as [00uu]. The pattern for X in the counterpart spring cultivar in the same year was noted as [uuuu]. We then paired

the winter and spring cultivars in the same year to create an encoded expression profile vector, and thus X had an expression profile of [0u,0u,uu,uu]. This analysis allowed us to identify the differentially expressed RNAs as cold-responsive if they expressed at least partially coordinately between winter and spring cultivars (eg. up profile [0u,0u,uu,uu] or down profile [00,0d,dd,d0]); as unresponsive or constitutively expressed if they were not regulated at the same time points between the paired cultivars (eg. profiles [00,0u,0u,u0], [00,00,00,00]). If winter and spring cultivars have the coordinated trend for an iRNA, it shows that despite the difference in their supposed phenotypes, the iRNA behaves similarly in winter and spring cultivars. It must be responding directly to the common environmental stimuli, cold, rather than the consequential regulation of growth and development. If not coordinated, even if it is differentially expressed, it is not responding to cold, and therefore constitutive. We concluded the regulation of each RNA based on its profile and quoted up-, down-regulated and unresponsive in cold response as CRup, CRdw and Cstv, respectively. See Table S8 for profile and regulation conversions. Using DEGreport (Pantano, 2022) to show the coordinated expression changes of cold response iRNAs, samples were grouped by cultivars for calculating the mean sequence abundance as the representation at each time point. It then performs pairwise correlations between iRNAs to hierarchically cluster them.

5.3.8 Functional annotations of XLOC database

The wheat mRNA transcripts – 151690 XLOCs – collected from the same field samples for sRNA analysis (Li *et al.*, 2018) were used to annotate precursors and interfering targets in this study. Functional annotations of the XLOCs were based on Uniprot sequence similarity (Joshi et Xu, 2007). The Uniprot database (Consortium, 2019, 2021) – providing high-quality sets of protein identities, cross-reference IDs and functional annotations – was accessed in October 2020. The best 20 Uniprot hits for each XLOC sequence were identified using the local blastx program (Boratyn *et al.*, 2013; Conesa et Götz, 2008) against Uniprot's Ensembl Plants collection, which also hosts the RefSeq v1.0 assembly from the IWGSC (Yates *et al.*, 2020; Appels *et al.*, 2018). Peptide hits with more than 50% identity or e-value smaller than 0.0001 were retained (Bergman, 2007; Conesa et Götz, 2008). To benefit from annotations in other species while maintaining relevance to wheat, hits were restricted to monocots. The performance of annotation for chosen filtering parameters was assessed in Figures SF6 and SF7. Protein IDs, GO terms and KEGG pathways (Consortium *et al.*, 2004; Kanehisa *et al.*, 2017) were retrieved from the best Uniprot hits. To address the gaps in annotations resulting from the Uniprot update, we also incorporated the XLOC annotations provided by Li *et al.* (2018).

5.3.9 Functional enrichments of iRNAs through target genes

MiRNA and siRNA target genes were identified using the 2017 schema of psRNATarget, with a stringent Expectation threshold of 2.0 (Dai *et al.*, 2018), against the XLOC database. psRNATarget also reports cleavage sites. We annotated the XLOC database with KEGG pathways and GO terms including GOBP, GOMF and GOCC. The overall enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the hypergeometric distribution test (similar to (Vlachos *et al.*, 2015)). The wheat genome GO term (or KEGG pathway) background was constructed by taking into account all the terms covered by the expressed protein repertoire. The terms covered by the target genes of differentially regulated iRNAs during the time course were compared against the background. Terms supported by at least two iRNAs and p-value ≤ 0.01 were considered enriched. The same procedure was also applied for the targets of specific types of small RNAs, such as miRNAs and siRNAs, involved in cold response (CR, CRup, CRdw) and originating from ancestral genomes.

5.4 Results

5.4.1 Integrated miRNA-siRNA Prediction

RNA interference pathways are crucial for gene regulation, but genome-wide studies have mainly focused on miRNAs, neglecting siRNAs. To bridge this gap, we developed a computational framework that predicts both miRNAs and siRNAs from wheat small RNA sequencing data, and accounts for differential expression, regulation, and functional analyses (Figure 5.1). Predicting miRNA and siRNA subtypes simultaneously from the same datasets allows direct comparisons of their genomic origins, expression patterns, and regulatory roles. We incorporated a high accuracy, scalable miRNA predictor adhering to gold-standard plant miRNA annotation criteria (Wu *et al.*, 2019), to address computational challenges in large-genome species like wheat. We also created a scalable, Python-based in-house classifier that uses phasing scores and precursor structure constraints to detect phased siRNA loci (Howell *et al.*, 2007). By consolidating miRNA and siRNA predictions into a single analytical pipeline, we expanded the classification scope and supported the assumption that they share common biogenesis pathways and regulatory functions. Analyzing multiple libraries allowed for the removal of false iRNA families by retaining only those with recurrent detection across samples and sufficient read support, leading to the identification of 339 iRNAs, including 215 miRNAs and 127 siRNAs (Figure 5.2, Table S19). Three iRNAs were predicted by both the miRNA and siRNA classifiers and are therefore counted in both categories.

miRNAs and siRNAs are usually categorized based on their genomic origin, alongside the structural and functional characteristics of their precursors (Vaucheret et Voinnet, 2024), which reflect their biogenesis pathways and evolutionary trajectories. Through comparative analysis, miRNAs and siRNAs can also be classified as broadly conserved or lineage-specific, showing their conservation, diversification, and regulatory roles. We categorized the 339 iRNAs into three groups based on the genomic context of their precursor loci. An iRNA was classified as ancestral if it originated from the A and B progenitor genomes (Marcussen *et al.*, 2014). Accordingly, **105 ancestral iRNAs** were identified. An iRNA was categorized as originating from coding regions if its genomic locus overlapped with the mRNAs previously detected in the same wheat datasets (Li *et al.*, 2018). A total of **190 iRNAs** originated from precursors overlapping with expressed **coding transcripts** on either chromosome strand. An iRNA was TE-derived if it originated from transposable elements or long terminal repeat (LTR) regions of the wheat genome, as annotated by IWGSC (Appels *et al.*, 2018) (see M&M section 5.3.4 for details). A total of **45 iRNAs** were classified as **TE-derived**.

We further validated our predictions by multiple strategies and classified them into conserved and novel iRNAs. We also noted discrepancies in annotation, with some predicted miRNAs identified in databases as siRNAs and vice versa. **First**, to identify miRNAs and siRNAs conserved in wheat and other plant species, we compared the mature sequences of our predictions with miRBase (Kozomara *et al.*, 2018), allowing two mismatches using miRFinder (Remita *et al.*, 2016). Of the 339 iRNAs, 95 matched entries in miRBase (46 from wheat *T. aestivum*, 42 from other monocots, and 7 from additional plants). By further incorporating matches with conserved wheat miRNAs from literature, including IWGSC-annotated miRNAs (Appels *et al.*, 2018) and our previous studies (Agharbaoui *et al.*, 2015; Rahmdani *et al.*, 2020), we identified 144 iRNAs as conserved wheat miRNAs. However, some conserved wheat miRNAs were classified as siRNAs in our analysis. **Second**, we cross-referenced our predictions with other phase-score-based annotations, including the PSU plant small RNA catalog (a collection based on ShortStack) (Lunardon *et al.*, 2020), and analyzed our 80 wheat libraries using ShortStack (Axtell, 2013b), allowing two mismatches. Among the 339 sequences, 203 iRNAs were detected, but the classifications of some miRNAs and siRNAs were swapped. **Third**, by comparing precursor sequences with the NCBI Nucleotide database (Boratyn *et al.*, 2013), we found that 58 iRNAs originated from known miRNA precursors, while **33 originated from recognized TAS** precursors. Finally, there were **96 novel iRNAs** that had not been reported elsewhere, comprising 42 miRNAs and 55 siRNAs. Among the 339 iRNAs, **90 dual-identity iRNAs** were identified, with each mature sequence classified as both a miRNA and a siRNA according to at least one combination of the referenced databases. See Figure 5.2 for the summary.

5.4.2 Discovery of Dual-Identity Small RNAs

A key discovery of our dual-predictor approach was the identification of **90 dual-identity iRNAs**, which exhibited characteristics of both miRNAs and siRNAs. They were predicted or annotated inconsistently across databases, likely due to their dual characteristics as both miRNAs and siRNAs. These molecules may undergo flexible processing pathways, suggesting a potential evolutionary link between miRNA and siRNA biogenesis. Among them, three iRNAs were independently predicted by both our miRNA and siRNA classifiers. Cross-referencing with external databases (Figure 5.4) revealed that 80 predicted miRNAs had been previously classified as siRNAs, while seven siRNAs in our dataset were previously annotated as miRNAs. Transposable elements are known for their potential to generate small RNAs with both miRNA and siRNA characteristics through dual-processing mechanisms in plants (Piriyapongsa et Jordan, 2008). However, in our genomic context analysis, dual-identity iRNAs were **not** predominantly associated with transposable element loci (15.6% $\approx$ 13.3%, Table 5.1), suggesting that additional regulatory mechanisms influence their biogenesis. Instead, these dual-identity iRNAs were overrepresented in ancestral loci, indicating their evolutionary conservation and possible functional importance (40.0% $>$ 31.0%, Table 5.1).

5.4.3 Novel iRNA Discovery

In this study, we identified **96 novel iRNAs** (Figure 5.2, Table S19). These iRNAs have no prior annotations in public databases or predictive models. To better understand these novel iRNAs, we examined the enrichment of their genomic origins and their potential functional roles (Table 5.1).

Ancestral iRNAs, which make up 31.0% of all iRNAs, were enriched among conserved miRNAs (36.2%) but were underrepresented in novel siRNAs (17.7%). This suggests that miRNAs are more closely tied to ancient regulatory functions, while novel siRNAs may be more dynamically influenced by recent genomic changes and species-specific adaptations. Coding regions, which account for 56.1% of all iRNAs, contributed to 60.9% of known iRNAs but were less frequent in novel iRNAs (43.8%). This suggests that novel iRNAs originated from a broader range of genomic regions. Notably, 37 novel iRNAs were identified as cultivar-specific (Table S16), suggesting that certain newly emerged novel iRNAs may contribute to genotype-dependent regulatory networks.

Among the novel iRNAs, 16 were identified as cold-response (Table S19), with **14 classified as high-confidence cold-responsive iRNAs** due to their targeting of known acclimation-regulated mRNAs previously detected

in the same wheat datasets (Li *et al.*, 2018). These high-confidence iRNAs are more likely to have functional relevance in cold adaptation, as their predicted targets are directly involved in stress-responsive pathways (Table 5.3). These newly identified iRNAs may mediate the cold response by regulating mRNAs involved in signal transduction, structural integrity, metabolism, and gene expression, thereby influencing kinase signaling, cell wall remodeling, osmoprotection, protein stabilization, and transcriptional regulation. Several iRNAs, such as S-029 and S-348, targeted receptor-like kinases that mediate stress perception and activate downstream cold-response pathways. Others, including M-025 and M-313, regulated proteins involved in cell wall stabilization, enhancing freezing tolerance by reinforcing structural integrity. Metabolic adjustments could be facilitated by iRNAs like M-169, which regulated fructan metabolism for osmoprotection, and M-266, which targeted heat shock proteins to prevent cold-induced protein misfolding. Transcriptional control may be influenced by M-239 and M-320, which regulated stress-responsive transcription factors and serine/threonine kinases involved in gene activation under cold conditions.

The identification and confirmation of 96 new iRNAs represent a significant expansion in the catalog of small RNAs in wheat. By distinguishing between ancestral and newly emerged iRNAs, the study revealed that novel iRNAs might arise from a wider array of genomic regions. Novel iRNAs could emerge in a lineage-specific manner, potentially influencing phenotypic variation among wheat cultivars. Moreover, the identification of high-confidence cold-responsive iRNAs underscores their functional relevance in stress adaptation. These findings highlight the dynamic evolution and potential regulatory roles of novel iRNAs, uncovering a range of previously overlooked candidates for probing stress adaptation mechanisms and serving as molecular markers in the development of cold-resistant wheat cultivars.

5.4.4 Identification of Wheat siRNA-Producing Genes

We explored precursor identity as an alternative validation approach for our predicted siRNAs. Given the short length of siRNAs, their precursors provide a more stable reference for homology-based identification. Using BLASTN searches against the Nucleotide database (Boratyn *et al.*, 2013), we identified **eight** precursors with significant sequence similarity to known **TAS genes** (Table 5.6). Among them, six were homologous to wheat *TAS3b* genes, while the other two resembled *TAS3a* genes from different plant species, suggesting that they represent **novel TAS genes in wheat**. From these precursors, we determined that **33** predicted siRNAs were **tasiRNAs**. By homology searches, a straightforward and effective method, we validated siRNA precursors and expanded the known tasiRNA repertoire in wheat. However, similarity searches did not

detect other potential siRNA-producing genes, such as the PHAS and hc-siRNA genes. This absence may result from insufficient siRNA annotations for these subtypes in database entries or the genetic specificity of certain tissues or species.

Difficulties in Detecting TAS Using Phase Score Method. The phase score evaluates siRNA arrangement along a precursor by integrating contributions from the primary effector siRNA and the sibling siRNAs generated through multiple Dicer digestion cycles (Howell *et al.*, 2007). To assess whether the eight TAS genes exhibited phasing characteristics, we compared their phasing scores with the actual sRNA alignment patterns along their genomic contigs (Figure S17). Our analysis showed that the 21-nt sRNA species maintained a consistent 21-nt interval, with most TAS precursors undergoing at least seven phasing cycles, which indicates strong phasing signals.

High levels of effector siRNA expression indicate strong phasing activity; however, weak or absent sibling siRNAs may underestimate phasing signals and potentially exclude true siRNA loci, even if they share homology with known TAS genes. For instance, precursor 2A__598975000 exhibited seven phasing cycles, but its siRNA frequency peaked at merely 140 reads, with some sibling phasing positions supported by as few as five reads. This limitation highlights the inherent challenge of detecting siRNA loci using the phase score method. Enhancing computational strategies and increasing sequencing depth may mitigate these challenges and improve detection accuracy.

miRNA Trigger for siRNA Biogenesis? miRNAs, especially miR390, have been hypothesized to be essential triggers for siRNA biogenesis through targeted precursor cleavage (Axtell *et al.*, 2006). However, it is still uncertain whether miRNA cleavage in wheat is an absolute requirement. To explore this, we employed psRNATarget (Dai et Zhao, 2011) to predict cleavage sites on the validated TAS precursors. Our analysis identified two miR390 cleavage sites aligning with phasing borders in six of the eight wheat *TAS3a* and *TAS3b* precursors. Despite an exhaustive search, two TAS precursors contained only a single miR390 cleavage site, implying that tasiRNA production may not always depend on this miRNA trigger.

Several factors could explain the absence of detectable cleavage sites. Genetic variation between the Canadian wheat cultivars analyzed in this study and the *Chinese Spring* reference genome (Appels *et al.*, 2018) may have altered miRNA recognition motifs, obscuring expected cleavage patterns. Additionally, some tasiRNA precursors may initiate biogenesis through alternative pathways, relying on a single cleavage or

miRNAs other than miR390 (de Felippes *et al.*, 2017; Fei *et al.*, 2013; Guo *et al.*, 2018). These observations suggest that, while miRNA cleavage is a well-established trigger for tasiRNA production, it may not be the only mechanism, allowing for alternative pathways contributing to tasiRNA biogenesis.

5.4.5 Evolutionary Landscape of Wheat iRNAs

The evolutionary dynamics of small RNAs are shaped by their genomic origins and selective pressures. By analyzing chromosomal distribution, ancestral inheritance, and genomic context, we uncovered patterns that revealed how miRNAs, siRNAs, and dual-identity iRNAs emerged from distinct genomic landscapes under different constraints and followed unique evolutionary trajectories in wheat.

The chromosomal distribution of iRNA precursors in the wheat genome highlights key differences between miRNAs and siRNAs. While miRNAs were broadly distributed across all seven chromosomes and conserved across the three progenitor genomes (A, B, and D), siRNAs tended to accumulate in specific genomic regions and were notably absent in others (Figure 5.3). This disparity suggests that miRNAs undergo strong evolutionary conservation, whereas siRNAs emerge from specific genomic hotspots. The spatial organization of siRNAs might reflect functional compartmentalization, where siRNAs were preferentially associated with specific genomic regions. The absence of siRNAs in some regions suggests that these areas lack the necessary triggers for production.

The genomic origins of iRNAs further illuminate their evolutionary divergence. Among the 339 identified iRNAs, **105** originated from ancestral genomes, **190** from coding regions, and **45** from transposable elements (Figure 5.2; Table S16, sheet 1 *all*, see columns H: *isAncestral*, AL: *XLOCprecursor*, BI: *isTEorLTR*). These distinct origins suggest multiple evolutionary pathways, where some iRNAs are conserved while others evolve novel functions. To investigate these evolutionary trajectories, we conducted a statistical enrichment analysis to examine the distribution of iRNA categories across ancestral genomes, coding regions, and transposable elements (Table 5.1). Our findings, summarized below, highlight distinct mechanisms shaping the biogenesis of miRNAs, siRNAs, and dual-identity iRNAs in wheat.

- **Ancestral iRNAs** (background = 31.0%) were significantly enriched in miRNAs (37.2%), known iRNAs (36.2%), and dual-identity iRNAs (40.0%). In contrast, ancestral iRNAs were underrepresented in siRNAs (20.5%) and novel iRNAs (17.7%). This distribution suggests that miRNAs and dual-identity

iRNAs are more evolutionarily conserved for their regulatory functions, whereas siRNAs and novel iRNAs may have emerged more recently and are subjected to stronger selective pressures and rapid turnover.

- **Coding region-derived iRNAs** (background = 56.1%) were depleted by miRNAs (50.2%), while significantly enriched by known iRNAs (60.9%), siRNAs (65.4%), and tasiRNAs (78.8%). This observation suggests that siRNAs (and their subtypes) are more likely to originate from protein-coding genes and are directly associated with gene expression, whereas miRNAs emerging from intergenic regions exhibit greater stability.
- **TE (transposable element)-derived iRNAs** (background = 13.3%) were significantly overrepresented in miRNAs (18.6%) and depleted in siRNAs (3.9%). The low abundance of siRNAs originating from transposable elements implies that these loci have shorter lifespans, likely due to rapid purging or their transient role in gene regulation. Over time, some TE-derived sequences may undergo structural stabilization and acquire functional relevance, leading to their retention as miRNAs with more conserved and stable biogenesis pathways.

5.4.6 Identification of cold response iRNAs

With 339 iRNAs predicted from wheat cultivars varying in freezing tolerance, grown from late summer to freezing, the dataset provided a unique opportunity to investigate small RNA contributions to cold adaptation. By analyzing temporal expression dynamics, we identified miRNAs and siRNAs involved in the cold response through a series of analytical procedures. Temporal expression profiles were assessed across five time points (T1–T5), with pairwise comparisons between T1 and subsequent time points. We examined genetic and phenotypic differences between winter and spring cultivars using *encoded vector clustering* (see M&M section 5.3.7 for details). This method aimed to identify iRNAs regulated **coordinately** up- or down-regulated between spring and winter cultivars, isolate cold-response iRNAs from confounding environmental factors, and distinguish direct cold responses from *Vrn-A1*-dependent growth effects. In contrast, **non-coordinated** iRNAs, even if differentially expressed, were classified as constitutive iRNAs (Ctsv) and not considered cold-response. We identified **66 cold-response (CR) iRNAs** from 32 encoded profile clusters among 204 differentially expressed (DE) iRNAs (Figure 5.2, Table S8 sheet 4, M&M section 5.3.7). The remaining 138 DE iRNAs and 135 non-DE iRNAs were classified as constitutive iRNAs, indicating they do not respond to cold stress. CR iRNAs were annotated as up-regulated (CRup), down-regulated (CRdw), or mixed (CRupCRdw) based on winter-spring coordination. Many iRNAs showed significant differential expression

over time (Figure 5.5), including several cold-responsive iRNAs from established families, along with others that denote new regulatory elements crucial for cold adaptation.

This set of 66 cold-response iRNAs includes **58 CR miRNAs** and **9 CR siRNAs**, with one exhibiting dual identity. Notably, 25 of these iRNAs belong to 19 conserved miRBase families (Table 5.2), while 11 are conserved within wheat but are not cataloged in miRBase. Additionally, 13 were identified by ShortStack, 1 was classified as a tasiRNA, and 16 were determined to be novel (Table S19). Notably, 14 of the 16 novel cold-response iRNAs regulated validated acclimation-related mRNA targets (Table 5.3). We found that cold-response iRNAs were not significantly enriched in ancestral genomes (33.3% $\approx$ 31.0%, Table 5.1), suggesting that recent selective pressures have shaped their role in cold adaptation. This implies that wheat's small RNA-mediated cold response is primarily driven by newly evolved genomic elements rather than ancient regulatory pathways.

5.4.7 Functions of Cold-Response iRNAs in Acclimation

Many conserved cold-response miRNA families have been well-documented for their roles in helping plants adapt to cold stress (Table 5.2). To further expand this understanding and uncover additional iRNAs involved in cold acclimation, we integrated previously characterized mRNA datasets from the same wheat samples (Li *et al.*, 2018). The functional roles of the 66 cold-response miRNAs and siRNAs in cellular components, molecular functions, and biological pathways can be inferred through comprehensive functional analysis (Consortium *et al.*, 2004). This analysis involved predicting target genes, annotating them with Gene Ontology (GO) terms and KEGG pathways using UniProt IDs, and performing statistical tests (see M&M for details). This approach allowed us to classify several novel cold-responsive iRNAs as high-confidence candidates, including three siRNAs (S-029, S-348, and S-454) and several miRNAs not previously linked to cold response (Table 5.3). These siRNAs specifically targeted acclimation-related mRNAs identified by Li *et al.* (2018), providing further support for their regulatory roles in cold stress adaptation.

Building on these findings, we examined the broader functional contributions of cold-response iRNAs to wheat acclimation. These regulatory molecules participated in multiple biological processes essential for freezing tolerance, including kinase signaling, transcriptional control, stress adaptation, metabolic homeostasis, and hormone signaling. The following functional categories are summarized based on Figure 5.7, Table 5.2, and Table 5.3.

Kinase Signaling and Stress Perception. Kinase-mediated signaling acts as molecular sensors that trigger downstream defense responses upon detecting environmental stress (Chen *et al.*, 2021b). We found that cold-response iRNAs collectively targeted a substantial number of kinases (63 unique UniProt IDs) and transcription factors (90 unique UniProt IDs) (Table S19, UniProt IDs in Table S16).While some kinase targets were putative, others had well-documented roles in plant temperature responses. Notable examples included Casein kinase 1 (Praat *et al.*, 2021), Wall-associated receptor kinase (Wang *et al.*, 2019b), Coatomer (Zhou *et al.*, 2019), NB-ARC domain kinase (Zhu *et al.*, 2010), Leucine-rich repeat receptor-like protein kinase (LRR-RLK) (Yang *et al.*, 2014), and GsSRK kinase (Sawicki *et al.*, 2019). To explore the regulatory interactions between CR iRNAs and their kinase targets, we analyzed their expression patterns, identifying pairs with strong and statistically significant correlations (Table 5.4). Selected iRNA-kinase expression patterns across cultivars and time points were illustrated in Figure 5.8. These correlations showed that certain iRNAs inhibited kinase expression, whereas others might engage in feedback loops that increase their target's expression. iRNAs regulate these kinases, showing a precise and dynamic response to cold stress that helped plants effectively activate protective mechanisms.

Transcriptional Regulation. Transcription factors (TFs) and DNA-binding proteins coordinate the changes in gene expression necessary for cold acclimation by regulating stress-responsive pathways and developmental transitions. Our analysis showed that many cold-response iRNAs specifically targeted TF families that have been well characterized in cold adaptation, including SQUAMOSA promoter-binding proteins (SBP) (Zeng *et al.*, 2018), Calmodulin-binding transcription activator (CAMTA) (Vo *et al.*, 2015), NAC domain proteins (Lv *et al.*, 2016), GATA-type transcription factors (Zhang *et al.*, 2021), and MBD domain-containing proteins (Hu *et al.*, 2011). These TFs contributed to diverse regulatory networks, modulating plant development, stress-responsive gene expression, and hormonal signaling. Interestingly, transcription factors were preferentially regulated by miRNAs rather than siRNAs, suggesting that miRNAs play a specialized role in tuning transcriptional responses to environmental cues. Cold-response iRNAs also influenced post-transcriptional regulatory mechanisms. Some iRNAs target genes involved in mRNA splicing, RNA modification, and DNA replication. These functions help maintain efficient gene expression and genome stability, which are critical for sustaining cellular functions under stress. To further investigate these regulatory interactions, we identified miRNA and its target that show strong and significant expression correlations (Table 5.5). The temporal expression patterns revealed how cold-responsive iRNAs dynamically adjusted their gene expression programs in response to low temperatures across cold-acclimation time points (Figure 5.9).

Homeostasis and Stress Adaptation. Maintaining cellular homeostasis is essential for cold acclimation. Our analysis showed that cold-response iRNAs regulated genes involved in heat shock proteins, fatty acid β-oxidation, fructan metabolism, and ubiquitin-mediated protein degradation. These pathways helped stabilize cellular energy levels and ensured the efficient removal of damaged or misfolded proteins. In addition, hormonal signaling modulated stress responses and growth under unfavorable conditions. Cold-response iRNAs influenced hormonal signaling, including abscisic acid-mediated signaling, auxin-dependent growth regulation, and cytokinin-controlled cell division. By regulating homeostasis and hormonal signaling, iRNAs enabled wheat to maintain cellular integrity while balancing survival strategies with continued development in response to cold stress.

5.4.7.1 Expression Clustering of cold-response miRNAs and siRNAs

Through functional enrichment analysis (Figure 5.7), we found that these 66 CR iRNAs exhibit distinct functional associations. *CRup* iRNAs were primarily enriched in mRNA splicing, plant organ development, RNA modification, and response to salt stress, while *CRdw* iRNAs were linked to DNA replication, cytokinin-activated signaling, fatty acid beta-oxidation, and the negative regulation of hypoxia-induced intrinsic apoptotic signaling pathways.

To investigate the dynamic interaction between miRNAs and siRNAs, we applied hierarchical clustering on paired winter-spring expression profiles. This process organized CR iRNAs into eight temporal clusters (C1–C8) according to their similar expression patterns, using unsupervised learning in DEGreport (Pantano, 2022) (Figure 5.6). Each cluster displayed a unique temporal pattern of regulation, suggesting a chronological activation of stress-response genes, outlined as follows.

- C1 (11 miRNAs, 4 siRNAs): Fatty acid beta-oxidation, DNA replication.
- C2 (10 miRNAs): Ribosomal small subunit assembly, cell wall biogenesis, response to abscisic acid.
- C3 (8 miRNAs, 2 siRNAs): Photoreactive repair, UV protection, syncytium formation, glucose import.
- C4 (25 miRNAs, 2 siRNAs): Response to auxin, long-chain fatty-acyl-CoA metabolic process, proton export across the plasma membrane.
- C5–C8 (4 miRNAs, 1 siRNAs): Transmembrane transport, RNA modification, auxin-activated signaling pathways, and protein ubiquitination.

Notably, miRNAs and siRNAs were frequently co-expressed within the same clusters (Table S16, sheet *Cold response*, see column *Cluster*), suggesting that cold-response siRNAs functioned in tandem with miRNAs in modulating key stress regulatory pathways.

5.4.8 iRNAs in *Vrn-A1*-Mediated Vernalization

The *Vrn-A1* gene regulates flowering time in wheat by determining whether a cultivar requires prolonged cold exposure (vernalization) to transition from vegetative to reproductive growth. Norstar (*vrn-A1*) is a winter wheat cultivar with exceptional cold tolerance, while Manitou (*Vrn-A1*) is a spring wheat suited to shorter growing seasons. To dissect the genetic control of vernalization and wheat adaptation, we analyzed these cultivars along with their isogenic derivatives, Iso8S (spring-habit Norstar) and Iso11W (winter-habit Manitou), which differ only in their *Vrn-A1* genotype. Iso8S (*Vrn-A1*) converts Norstar into spring wheat, whereas Iso11W (*vrn-A1*) makes Manitou behave as a winter wheat. These isogenic lines provide an ideal system for investigating how iRNAs regulate vernalization, cold response, and flowering in wheat.

Among the 204 differentially expressed (DE) iRNAs, 121 exhibited cultivar-specific differential expression. In particular, we identified 11 Iso11W-specific, 23 Iso8S-specific, 27 Manitou-specific, 18 Norstar-specific, 15 spring-habit specific (Manitou and Iso8S), 5 winter-habit specific (Norstar and Iso11W), 6 Manitou-background specific (Manitou and Iso11W), and 16 Norstar-background specific iRNAs (Norstar and Iso8S) (Table S16, several sheets).

5.4.8.1 Differential iRNA Networks Under *Vrn-A1* for Growth Strategies

Winter-habit cultivars (Norstar and Iso11W) exhibited distinct iRNA expression profiles associated with cold survival, dormancy, and vernalization-controlled flowering. We found that cold-response miR9674 and miR5062 members were specifically differentially regulated in these cultivars (Table 7). Functional enrichment analysis (Figure SF3)showed that these iRNAs modulated meiosis, histone post-translational modifications (methylation, phosphorylation, acetylation), flowering processes (microgametogenesis, megagametogenesis), negative regulation of embryonic development, gene silencing, and vacuole-related transport.

Conversely, **spring-habit cultivars** (Manitou and Iso8S) expressed different sets of iRNAs, including miR166, miR167, miR319, miR393, miR396, and miR9778 (Table 5.7, functional enrichment in Figure SF3. Some miRNA families were known for their cold response (Table 5.2). These iRNAs were enriched in pathways related to

plant development, gibberellic acid homeostasis, chromatin assembly, DNA replication, and responses to oxidative stress. The involvement of these iRNAs in bilateral symmetry, meristem growth, and hormone metabolism suggests their role in rapid development and cold-independent flowering.

In conclusion, our findings suggest that *Vrn-A1*-dependent iRNA expression patterns contribute to the distinct developmental strategies of winter and spring habit wheat, emphasizing the importance of iRNA networks in shaping wheat adaptation to seasonal growth conditions.

5.4.8.2 iRNA Isoforms in Cultivar Adaptation

Cultivar specificity can sometimes arise from the differential expression of **isoforms** within a single miRNA or siRNA family. For example, among the 66 **cultivar-specific miRNAs** (including one dual-identity miRNA), the **miR159** family exhibited distinct isoform expression patterns that varied between cultivars. Similarly, within the 56 **cultivar-specific siRNAs**, members of the **TAS families** predominantly expressed different siRNA siblings derived from the same TAS precursors across cultivars (Table 5.8). These results showed that small RNA isoforms were not produced randomly but were tightly regulated by the cultivar's genetic background. This genetically programmed regulation probably affected small RNA-mediated pathways, including those governing vernalization (*Vrn-A1*) and cold adaptation, thereby shaping cultivar-specific responses and promoting adaptation.

5.5 Discussion

In this study, we used freezing-tolerant wheat cultivars as a case study to demonstrate the power of an integrated framework for small RNA research. By jointly analyzing different classes of small RNAs, we identified 339 iRNAs (215 miRNAs and 127 siRNAs), including 96 novel iRNAs, 33 tasiRNAs, 90 dual-identity iRNAs, and 66 cold-response iRNAs. These findings significantly expand the known repertoire of small RNAs in wheat.

We uncovered key regulatory elements in cold stress adaptation by integrating miRNA and siRNA prediction within a unified framework. The presence of 90 dual-identity iRNAs suggests an evolutionary and functional link between miRNAs and siRNAs, blurring the functional boundaries between these two classes. Identifying siRNA-generating genes in wheat further expands our understanding of the siRNA landscape. Notably, the discovery of TAS precursors lacking miR390 cleavage sites suggests that tasiRNA production may occur through non-canonical mechanisms, such as alternative miRNA triggers. Additionally, 66 cold-response

iRNAs were identified, many targeting key regulatory genes involved in cold adaptation.

To ensure the reliability of our predictions, we employed multiple validation strategies, including cross-referencing with established databases such as miRBase, ShortStack predictions, literature-reported wheat miRNAs, and phase-score-based annotations. Furthermore, their precursors were validated through BLASTN analysis, which indicated the presence of TAS homologs and further substantiated their classification as functional small RNAs. By integrating multiple validation strategies, we boosted confidence in small RNA identification and functional annotation, demonstrating the robustness of our bioinformatics approach.

The genomic mapping of iRNAs to ancestral loci, coding regions, and transposable elements indicates that small RNAs have evolved through various pathways, with selective pressures shaping their functional diversification. miRNAs and dual-identity iRNAs are more evolutionarily conserved, while siRNAs and novel iRNAs tend to emerge from recent genomic adaptations. The discovery of dual-identity iRNAs suggests an intermediate state in the evolutionary continuum between miRNAs and siRNAs. Cold-response iRNAs, in particular, seem to be a recent evolutionary adaptation rather than a product of conserved regulatory mechanisms. This finding further supports the hypothesis that small RNA-mediated regulation is an adaptive mechanism that enhances stress resilience in wheat.

Our analysis further revealed distinct roles of cold-response iRNAs in stress adaptation. These iRNAs target key regulatory genes, including kinases, transcription factors, and metabolic enzymes. The clustering of cold-response iRNAs (C1–C8) revealed distinct expression patterns, suggesting multiple regulatory layers in wheat's response to cold stress. Some clusters were enriched in auxin-related genes, while others were associated with metabolic adjustments and transcriptional regulation. The temporal co-expression of certain miRNAs and siRNAs within these clusters suggests a coordinated regulatory mechanism that orchestrates stress responses during the seasonal transition from warm to cold at different developmental and environmental stages. Moreover, cultivar-specific iRNAs highlight the genetic variability in wheat's response to environmental stimuli. Modulating cultivar-specific cold-response iRNAs may influence flowering time and stress responses, offering a potential avenue for targeted breeding and genetic modifications to develop wheat varieties optimized for diverse environmental conditions.

While this study identifies candidate biological functions of cold-responsive iRNAs primarily through target gene analysis, deeper biological interpretation would require explicit integration with domain-specific

plant stress literature and systematic cross-database analysis. Establishing direct links to molecular stress mechanisms, hormonal signaling pathways, and transcriptional regulatory networks would enable interpretation beyond observation-level patterns inferred from this dataset alone. Such synthesis, which relies on curated pathway knowledge and regulatory models, lies outside the methodological scope of this thesis and represents a natural direction for future interdisciplinary research.

Because this study relies on field-derived wheat samples, it is important to explicitly consider the implications of environmental variability for interpretation and reproducibility. Field conditions capture biologically realistic stress responses by integrating multiple regulatory signals simultaneously, thereby increasing the relevance of the identified iRNAs to agricultural contexts. However, this realism necessarily introduces constraints: environmental factors cannot be fully controlled, temperature dynamics may vary within and across days, average trends may not be strictly monotonic, and additional abiotic (e.g., water availability, UV exposure) or biotic stresses (e.g., insects, pathogens) may co-occur. Consequently, the observed iRNA responses reflect combined stress effects rather than isolated cold exposure. While this limits strict reproducibility under controlled conditions, it provides ecologically valid insights into stress-responsive regulation in wheat under natural environments.

Despite these advances, several challenges remain in identifying promising small RNA markers. The complexity and large size of the wheat genome pose challenges for small RNA prediction and classification. Computational predictors still struggle to accurately distinguish siRNA subtypes because of the absence of a dedicated reference database. Additionally, curation efforts and annotation criteria for siRNAs remain insufficient, leading to fragmented knowledge and limiting comparative analyses. Resource constraints further impact the development of bioinformatics tools for small RNA studies in wheat and other complex plant genomes.

Nevertheless, our study demonstrates the feasibility and importance of an integrated approach to small RNA analysis. Addressing current limitations by standardizing prediction criteria, curating small RNA classes and their precursors, and expanding the capacity of bioinformatics resources will enhance the reliability and applicability of future research. Ultimately, these efforts will support bioinformatics-driven breeding strategies, providing opportunities to enhance crop resilience in response to environmental stressors.

5.6 Figures and Tables

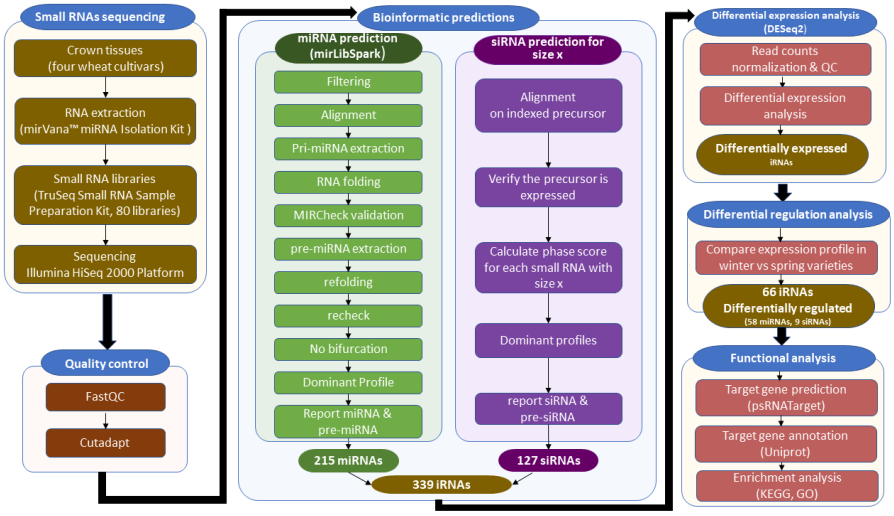


Figure 5.1: Bioinformatics analysis workflow of wheat miRNA and siRNA prediction and characterization. The workflow is divided into six steps: Small RNA sequencing, Quality control, Bioinformatics predictions, Differential expression analysis, Differential regulation analysis, and Functional analysis. The bioinformatics predictions contain miRNA predictions made by mirLibSpark and siRNA predictions made using in-house Python scripts. The numbers of predicted iRNAs were obtained from analyzing 80 wheat libraries.

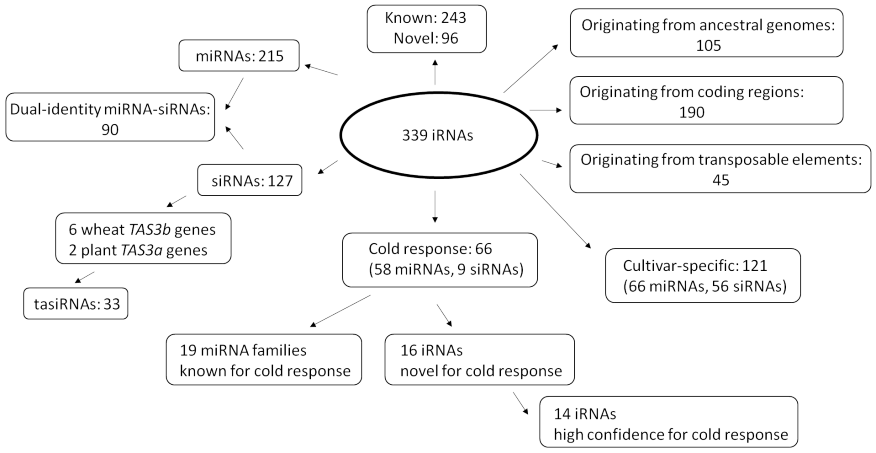


Figure 5.2: Classification of predicted iRNAs. A total of 339 iRNAs were identified from 80 wheat sRNA-seq libraries, including 96 novel siRNAs, 33 trans-acting siRNAs (tasiRNAs), and 90 dual-identity iRNAs. Expression profiling revealed 58 miRNAs and 9 siRNAs associated with cold responses; they belong to 19 miRNA families known to be involved in the cold response, and 14 iRNAs are novel but highly confident in cold response. We categorized the 339 iRNAs based on the genomic context of their precursor loci into three groups: originating from ancestral genomes (105), coding regions (190), and transposable elements (45). There are 121 cultivar-specific iRNAs. Note that some iRNAs are dual identities and could be counted twice in subcategories.

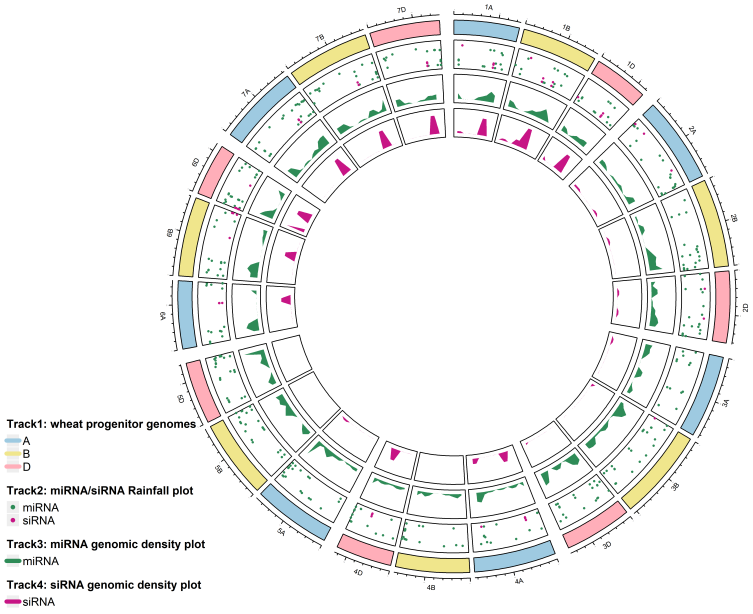


Figure 5.3: Chromosomal distribution of miRNAs and siRNAs in the wheat genome. This Circos plot displays genomic locations of miRNA and siRNA precursors and density across the three wheat progenitor genomes (A, B, and D). The outermost ring represents wheat chromosomes, with miRNA-originating loci mapped in the second ring and siRNA-originating loci in the third ring. The innermost heatmap tracks the genomic density of these small RNAs, revealing distinct clustering patterns. miRNAs are more evenly distributed, while siRNAs tend to form dense clusters, particularly in transposable element-rich regions. The plot highlights genomic hotspots with abundant phased siRNAs, suggesting potential regulatory roles. This figure provides an integrated view of small RNA chromosomal organization, helping to identify evolutionary trends and functional loci for further study.

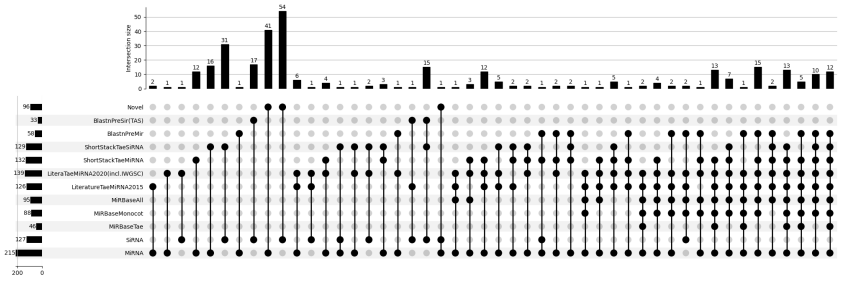


Figure 5.4: Cross-database comparison of predicted miRNAs and siRNAs. This UpSet plot illustrates the intersection of our predicted miRNAs and siRNAs with independent small RNA datasets and predictors from miRBase, literature sources, ShortStack predictions, and BLASTN-annotated precursors. (Left panel) The horizontal bar graph of total number of sRNA species in this study seen in different database - MiRBase-Tae/Monocot/All as the sRNA being annotated as wheat, monocot or plant miRNA in MiRBase v22.0 using miRFinder; LiteratureTaeMiRNA2015/2020 as the sRNA being annotated as wheat miRNA in literature (including IWGSC); BlastnPreMir/Sir as their precursors being annotated as miRNA or siRNA precursor in BLASTN similarity search (BLASTN only found similarity with TAS for siRNA precursors); and ShortStackMiRNA/SiRNA as the sRNA being predicted as miRNA or siRNA from our libraries using ShortStack, or annotated in PSU wheat sRNA database using miRFinder. Novel collects the sRNAs in this study not found in any databases above. (Right panel) The dots and lines on the bottom show which intersection is represented by the vertical bar graph above.

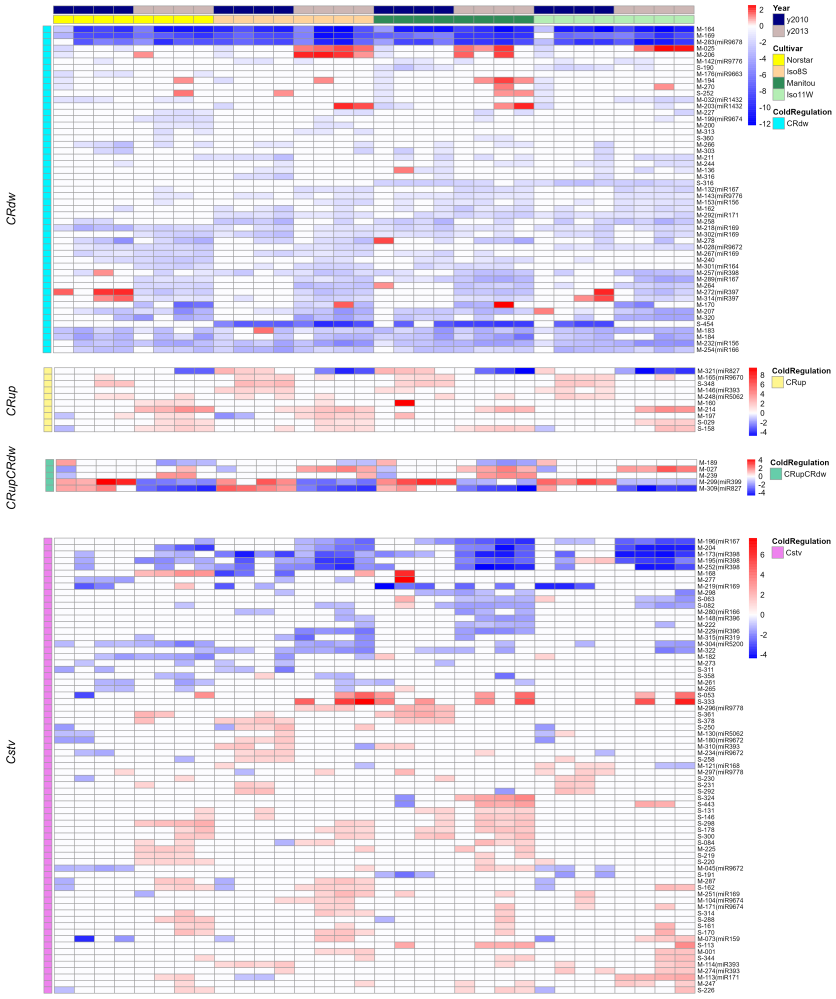


Figure 5.5: Temporal differential expression of miRNAs and siRNAs across four wheat cultivars in 2010 and 2013. iRNAs were classified into cold response categories (CRup, CRdw, CRupCRdw) using encoded vector clustering, which grouped them based on coordinated expression trends between winter and spring cultivars before plotting. Each panel separately displays these categories, revealing distinct regulatory responses and temporal dynamics. Each row represents the expression pattern of an iRNA across cultivars (Manitou, Iso11W, Norstar, and Iso8S) and time points in log2 fold change. Red, blue and white cells indicate expressions were up, down, or no change, respectively, as compared to their time point 1. All differentially expressed miRNAs and siRNAs are shown, except in the Cstv category that selected ones are shown by filtering changes in fewer than ($<$) 5 samples. MiR family of a miRNA, if available, is shown next to the ID.

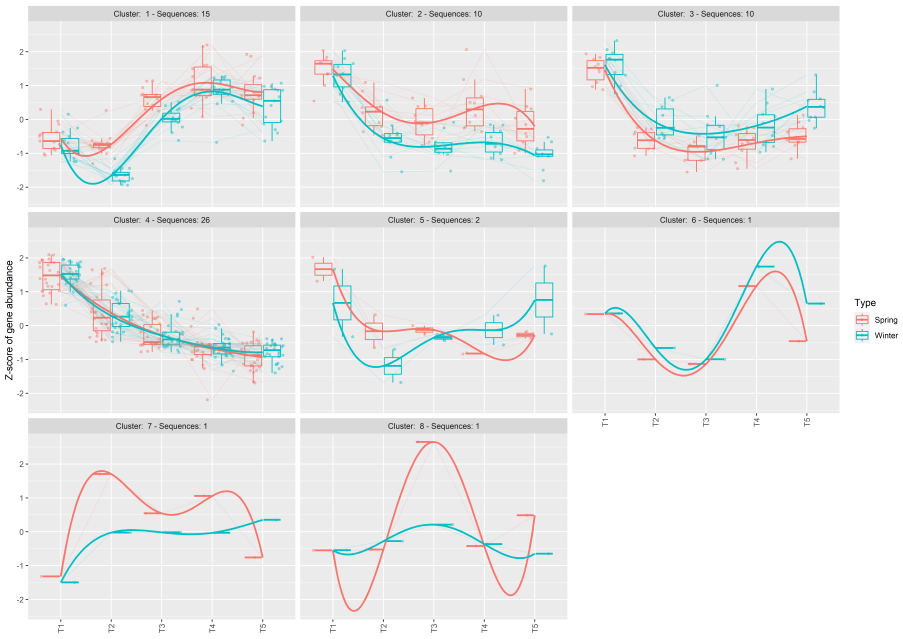


Figure 5.6: Winter-spring paired temporal expression profiles of cold response iRNAs. Identified 66 cold response iRNAs are clustered into 8 groups based on their differential expressions considering sampling time from warm to cold and cold hardiness between winter and spring cultivars.

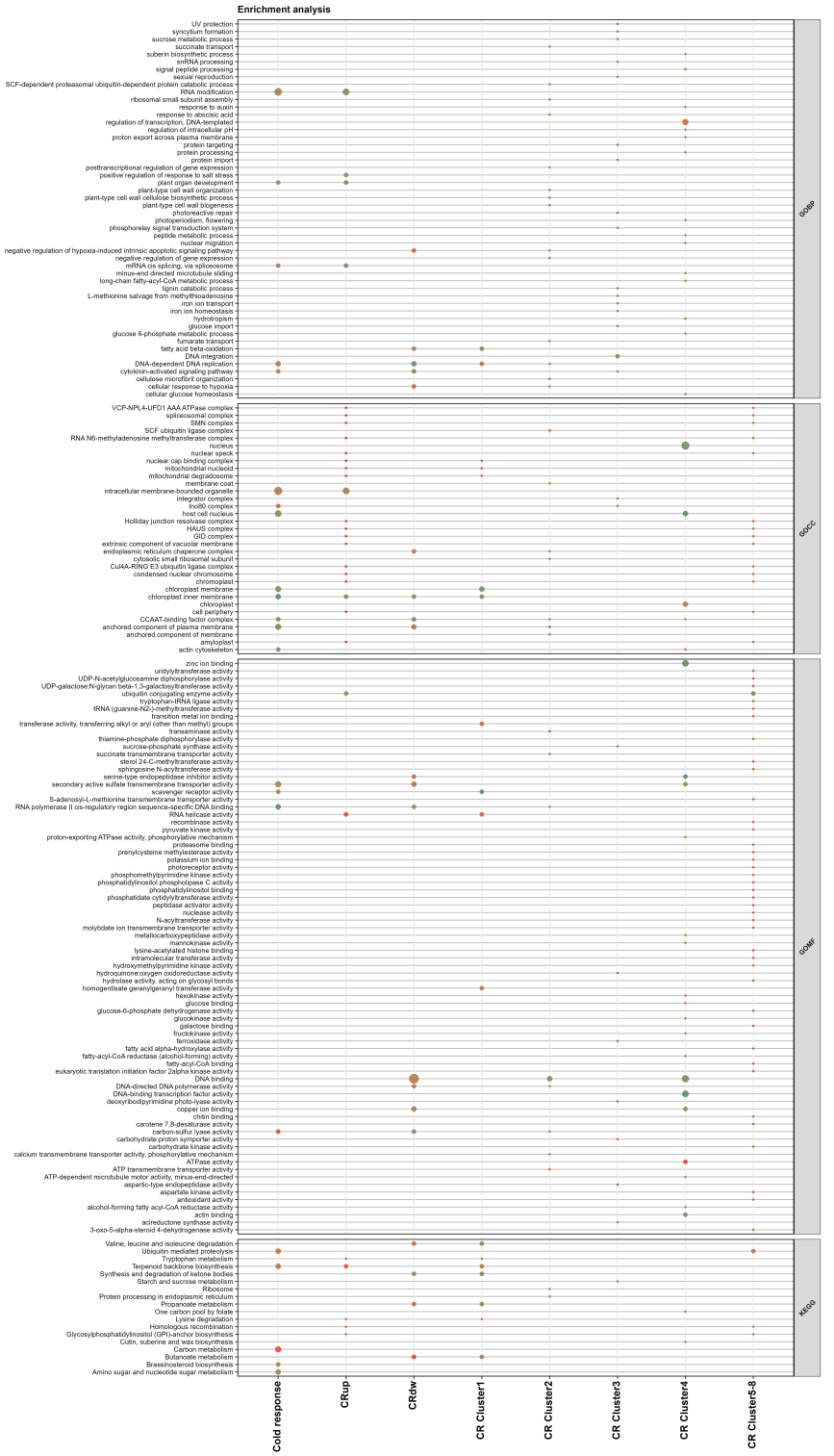

Enrichment analysis
GOBP
GOCC
GOMF
KEGG
Count
2
4
6
8
10
13
Pvalue
0.04
0.03
0.02
0.01
0.00
Cold response
CRup
CRdw
CR Cluster1
CR Cluster2
CR Cluster3
CR Cluster4
CR Cluster5-8

Figure 5.7: Functional enrichment of target genes regulated by cold-response iRNAs. Enrichment analysis was performed for different groups of cold-response iRNAs, including all cold-response iRNAs (**CR**), up-regulated (**CRup**), downregulated (**CRdw**), and eight expression clusters (**C1–C8**). **CR iRNAs** are enriched in kinase signaling, transcriptional regulation, and cold stress response. **CRup iRNAs** are associated with mRNA splicing, plant organ development, RNA modification, and response to salt stress. **CRdw iRNAs** are linked to DNA replication, cytokinin-activated signaling, fatty acid beta-oxidation, and the negative regulation of hypoxia-induced intrinsic apoptotic signaling. Among the expression clusters, **C1** is enriched in fatty acid metabolism and DNA replication; **C2** in ribosomal assembly, cell wall biogenesis, and abscisic acid response; **C3** in photoreactive repair, UV protection, and glucose import; **C4** in auxin response, fatty acid metabolism, and proton export; and **C5–C8** in auxin-activated signaling pathways, transmembrane transport, RNA modification, and protein ubiquitination. Enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the standard hypergeometric distribution test for the terms covered by the target genes of each RNA group. Terms with FDR ≤ 0.05 were considered significantly enriched.

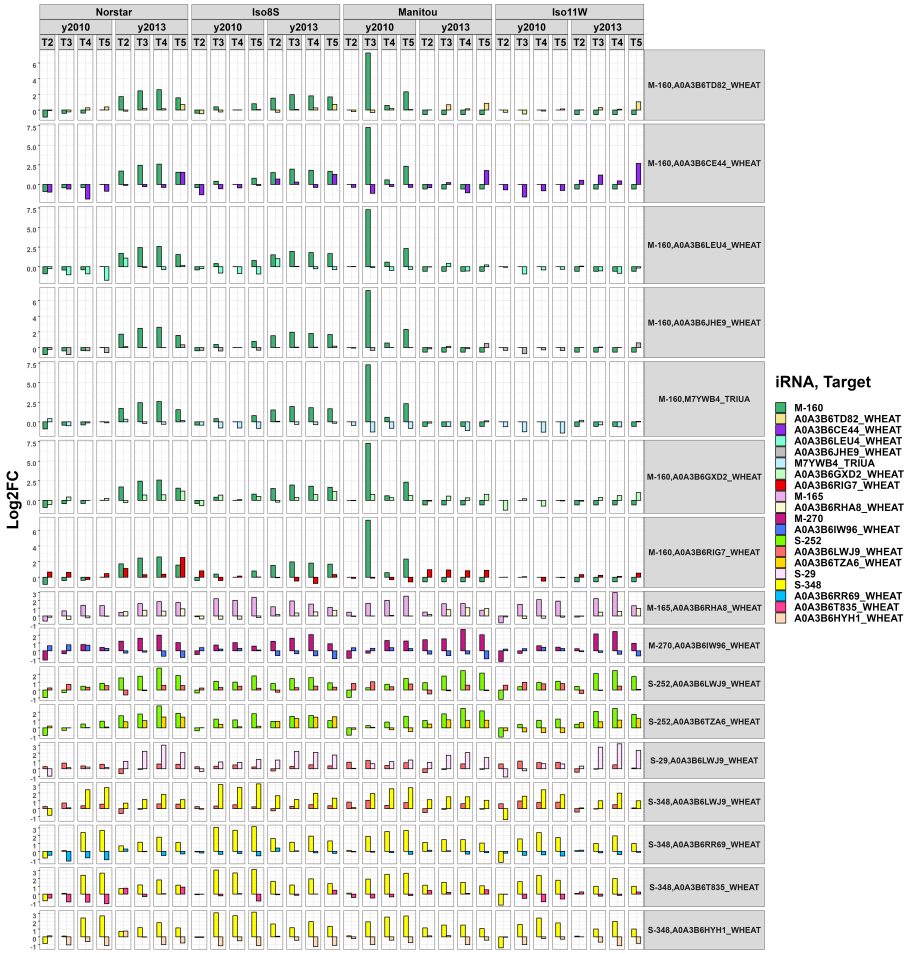


Figure 5.8: Gene expression patterns of cold response mature sequences and their kinase targets. Gene expression patterns of representative pairs of mature sequences, precursor, and downstream target kinase genes in 80 wheat libraries for selected cold response iRNAs. Log2 fold changes in the y-axis, and time point series in the x-axis.

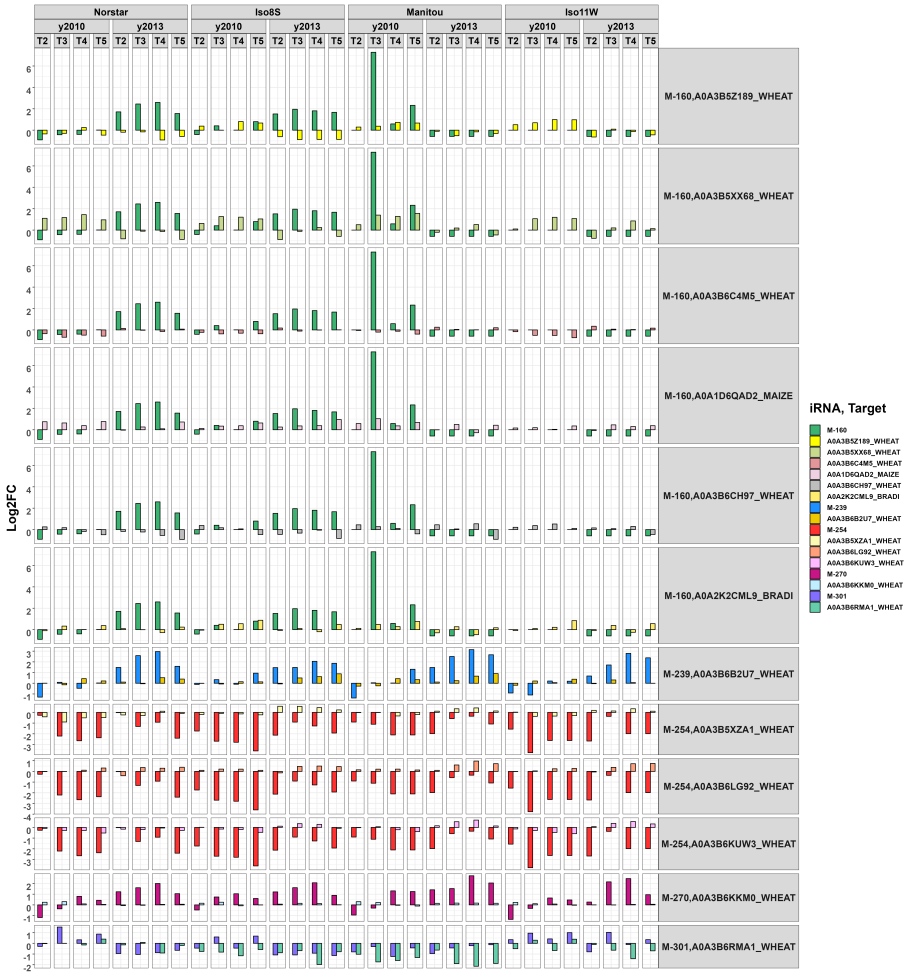


Figure 5.9: Gene expression patterns of cold response mature sequences and their transcription factor targets. Gene expression patterns of representative pairs of mature sequences, precursor, and downstream target transcription factor genes in 80 wheat libraries for selected cold response iRNAs. Log2 fold changes in the y-axis, and time point series in the x-axis.

Table 5.1: Overview of miRNA and siRNA prediction showing various categories of predicted iRNAs originating from (A) ancestral genomes, (B) coding regions, and (C) transposable elements. Over- and underrepresented: $p < 0.05$ by upper or lower side of the hypergeometric distribution test. [1] Includes transposable elements (TE) and long terminal repeats (LTR) loci.

(A)	Total number	Number of iRNAs originating from ancestral genomes	%	Comment
Novel iRNA	96	17	17.7	Underrepresented (p = 0.00051)
Known iRNA	243	88	36.2	Overrepresented (p = 0.00051)
miRNA	215	80	37.2	Overrepresented (p = 0.00068)
siRNA	127	26	20.5	Underrepresented (p = 0.00078)
Dual identity iRNA	90	36	40.0	Overrepresented (p = 0.0223)
iRNA originating from coding regions	190	81	42.6	Overrepresented (p < 0.00001)
Cold response iRNA	66	22	33.3	Similar to background (p = 0.37303)
All iRNA (background)	339	105	31.0	Background

(B)	Total number	Number of iRNAs originating from coding regions	%	Comment
Known iRNA	243	148	60.9	Overrepresented (p = 0.00307)
Novel iRNA	96	42	43.8	Underrepresented (p = 0.00307)
miRNA	215	108	50.2	Underrepresented (p = 0.00306)
siRNA	127	83	65.4	Overrepresented (p = 0.00508)
tasiRNA	33	26	78.8	Overrepresented (p = 0.00392)
All iRNA (background)	339	190	56.1	Background

(C)	Total number	Number of iRNAs originating from transposable elements[1]	%	Comment
miRNA	215	40	18.6	Overrepresented (p = 0.00005)
siRNA	127	5	3.9	Underrepresented (p = 0.00003)
Dual identity iRNA	90	14	15.6	Similar to background (p = 0.28205)
All iRNA (background)	339	45	13.3	Background

Table 5.2: Known cold-response miRNA families identified in this study. Among the 66 cold-response iRNAs identified, several belong to well-characterized miRBase families with established roles in plant cold acclimation. These miRNAs regulate key adaptive processes, including transcriptional control, signal transduction, stress response pathways, and structural stability. This table provides supporting literature for their involvement in cold response. CRdw: cold response down-regulated; CRup: cold response up-regulated; na: not available.

miR family	NewID	Cold response	Cold response in literature	Other abiotic stresses in literature	Target description
miR1432	M-032,M-203	CRdw	na	water (Giusti *et al.*, 2017)	DNA-dependent DNA replication, cell wall biogenesis
miR156	M-232,M-153	CRdw	(Tang *et al.*, 2012; Song *et al.*, 2017; Sunkar *et al.*, 2012; Xu *et al.*, 2016)	UVB (Guleria *et al.*, 2011)	Squamosa promoter-binding-like proteins
miR164	M-301	CRdw	(Lv *et al.*, 2010; Xu *et al.*, 2016; Tang *et al.*, 2012; Song *et al.*, 2017)	salt, mechanical (Guleria *et al.*, 2011)	PIF3, Circadian Clock, flower development (Li *et al.*, 2018)
miR166	M-254	CRdw	(Lv *et al.*, 2010; Tang *et al.*, 2012)	UVB (Guleria *et al.*, 2011)	Homeobox-leucine zipper protein HOX33
miR167	M-289,M-132	CRdw	(Xu *et al.*, 2016; Lv *et al.*, 2010; Tang *et al.*, 2012; Song *et al.*, 2017)	UVB (Guleria *et al.*, 2011)	Cytochrome P450
miR169	M-218,M-267,M-302	CRdw	(Xu *et al.*, 2016; Tang *et al.*, 2012; Song *et al.*, 2017)	UVB, salt (Guleria *et al.*, 2011)	Nuclear transcription factor Y subunit A-5
miR171	M-292	CRdw	(Song *et al.*, 2017; Sunkar *et al.*, 2012; Xu *et al.*, 2016)	UVB (Guleria *et al.*, 2011)	GRAS transcription factor
miR393	M-146	CRup	(Xu *et al.*, 2016; Tang *et al.*, 2012; Song *et al.*, 2017; Sunkar *et al.*, 2012)	UVB, abscisic acid, hypoxia (Guleria *et al.*, 2011)	Transport inhibitor response 1-like protein Os05g0150500
miR397	M-272,M-314	CRdw	(Xu *et al.*, 2016; Song *et al.*, 2017; Zhang *et al.*, 2009; Sunkar *et al.*, 2012; Wang *et al.*, 2016)	water, salt (Guleria *et al.*, 2011)	Laccase-12, VRN-A1 allelic effect (both MA and NO backgrounds)
miR398	M-257	CRdw	(Song *et al.*, 2017; Lv *et al.*, 2010; Wang *et al.*, 2014)	UVB, copper (Guleria *et al.*, 2011)	E3 ubiquitin-protein ligase listerin, snRNA processing
miR399	M-299	CRupCRdw	(Peng *et al.*, 2021)	salt (Guleria *et al.*, 2011)	ubiquitin-conjugating enzyme
miR5062	M-248	CRup	(Song *et al.*, 2017)	na	Protein argonaute MEL1, Potassium transporter, Reverse transcriptase, Carboxypeptidase
miR827	M-321,M-309	CRup	(Song *et al.*, 2017)	na	SPX domain-containing membrane protein Os02g45520, disease resistance protein RGA3
miR9663	M-176	CRdw	(Song *et al.*, 2017)	na	Polyglutamine-binding protein 1
miR9670	M-165	CRup	(Song *et al.*, 2017)	na	G-type lectin S-receptor-like serine/threonine-protein kinase SD2-5
miR9672	M-028	CRdw	(Song *et al.*, 2017)	na	F-box only protein 7
miR9674	M-199	CRdw	(Song *et al.*, 2017)	AI (Agharbaoui *et al.*, 2015)	phosphatase, NB-ARC, PPR_long, ANK_REP_REGION, and F-box domains-containing protein, mitochondrial protein Rf1
miR9678	M-283	CRdw	na	na	Serine protease SPPA
miR9776	M-142,M-143	CRdw	na	drought (Rahim *et al.*, 2022)	na

Table 5.3: High-confidence novel cold-response iRNAs with known functional targets identified as differentially expressed genes in Li *et al.* (2018). These novel cold-response iRNAs regulate stress perception, gene silencing, structural integrity, metabolic adaptation, protein stability, and transcriptional control.
Note[1]: Uniprot ID of the target with the highest psRNAtarget score.

NewID	Class	miR family	Cold response	Target Uniprot ID[1]	mRNA type (Li *et al.*, 2018)	Description
S-029	siRNA	noHit	CRup	A0A3B6BX82_WHEAT	NO backgrounds	Cysteine-rich receptor-like protein kinase 26
S-348	siRNA	noHit	CRup	A0A3B5Z507_WHEAT	NO backgrounds	G-type lectin S-receptor-like serine/threonine-protein kinase SRK
S-454	siRNA	noHit	CRdw	A0A3B6TTW8_WHEAT	NO backgrounds	Argonaute 1D
M-025	miRNA	noHit	CRdw	A0A080YUK3_WHEAT	MA backgrounds	Wall-associated receptor kinase 5, VRN-A1 allelic effect (both MA and NO backgrounds (Li *et al.*, 2018))
M-136	miRNA	noHit	CRdw	A0A1D5UIN7_WHEAT	Up-regulated	Disease resistance RPP8-like protein 2
M-143	miRNA	noHit	CRdw	W5D700_WHEAT	Up-regulated	Subtilisin-chymotrypsin inhibitor-2A, VRN-A1 allelic effect (NO background)
M-160	miRNA	noHit	CRup	A0A3B6B273_WHEAT	Down-regulated	VRN-A1 allelic effect (NO background), ABA-dependent pathway (Li *et al.*, 2018)
M-169	miRNA	noHit	CRdw	A0A3B6NIM8_WHEAT	NO backgrounds	Fructan 1-exohydrolase w1
M-200	miRNA	noHit	CRdw	A0A3B6NIR3_WHEAT	NO backgrounds	Mitochondrial protein Rf1
M-239	miRNA	noHit	CRupCRdw	A0A3B6B2U7_WHEAT	Up-regulated	GATA transcription factor 5
M-266	miRNA	noHit	CRdw	A0A3B6NSR2_WHEAT	NO backgrounds	Heat shock 70 kDa protein 17
M-270	miRNA	noHit	CRdw	A0A3B6PNX4_WHEAT	MA backgrounds	Acetylornithine deacetylase
M-313	miRNA	noHit	CRdw	A0A3B6LTE8_WHEAT	Down-regulated	COBRA-like protein 3
M-320	miRNA	noHit	CRdw	A0A3B6NQ17_WHEAT	NO backgrounds	Serine/threonine-protein kinase DDB_G0276461

Table 5.4: The cold response mature sequences and their kinase targets. Cold-response miRNAs and siRNAs' expression-correlated targets that were annotated as kinase are listed. We set the correlation strength to four levels from high to low by their Spearman correlation coefficient's absolute values: strong, moderate, fair and weak (Akoglu, 2018). The *p* value of Spearman smaller than 0.05 was labeled with a star (*).

iRNA	Target Uniprot ID	Description	kinase known in literature for cold response	Spearman correlation coefficient	pvalue	Correlation strength
M-160	A0A3B6TD82_WHEAT	Wall-associated receptor kinase 3	(de Oliveira *et al.*, 2014; Kazemi-Shahandashti et Maali-Amiri, 2018; Wang *et al.*, 2019b)	0.25	0.02773	fair*
M-160	A0A3B6CE44_WHEAT	LRR receptor-like serine/threonine-protein kinase FLS2	(Solanke et Sharma, 2008; Zhu *et al.*, 2010; Yang *et al.*, 2014; Kazemi-Shahandashti et Maali-Amiri, 2018)	0.26	0.02087	fair*
M-160	A0A3B6LEU4_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	-0.52	0	moderate*
M-160	A0A3B6JHE9_WHEAT	Coatomer subunit beta'-3	(Zhou *et al.*, 2019)	-0.23	0.04097	fair*
M-160	M7YWB4_TRIUA	Casein kinase I isoform beta	(Dong et Pei, 2014; Praat *et al.*, 2021)	-0.42	0.0001	fair*
M-160	A0A3B6GXD2_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	0.3	0.00613	fair*
M-160	A0A3B6RIG7_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	-0.5	0	moderate*
M-165	A0A3B6RHA8_WHEAT	Receptor-like serine/threonine-protein kinase (EC 2.7.11.1)	(Chen *et al.*, 2021b)	-0.22	0.05296	fair
M-270	A0A3B6IW96_WHEAT	TPX2 domain-containing protein;Os03g0212600 protein (Targeting protein for Xklp2 containing protein, expressed)	na	0.35	0.00161	fair*
S-252	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	0.39	0.0004	fair*
S-252	A0A3B6TZA6_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	0.24	0.03276	fair*
S-29	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	-0.21	0.06668	fair
S-348	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	(Yadav, 2010)	0.26	0.02102	fair*
S-348	A0A3B6RR69_WHEAT	Protein kinase domain-containing protein;NB-ARC domain-containing protein	(Zhu *et al.*, 2010)	-0.29	0.00932	fair*
S-348	A0A3B6T835_WHEAT	G-type lectin S-receptor-like serine/threonine-protein kinase SRK	(Chen *et al.*, 2021b)	-0.31	0.00585	fair*
S-348	A0A3B6HYH1_WHEAT	Wall-associated receptor kinase 5	(de Oliveira *et al.*, 2014; Kazemi-Shahandashti et Maali-Amiri, 2018; Wang *et al.*, 2019b)	-0.21	0.06585	fair

Table 5.5: The cold response mature sequences and their DNA-binding protein targets. Cold-response miRNAs and siRNAs' expression-correlated targets that were annotated as a DNA-binding or transcription factor are listed. We set the correlation strength to four levels from high to low by their Spearman correlation coefficient's absolute values: strong, moderate, fair and weak (Akoglu, 2018). The p value of Spearman smaller than 0.05 was labeled with a star (*).

iRNA	Target Uniprot ID	Description	TF/DNA-binding known in literature for cold response	Spearman correlation coefficient	pvalue	Correlation strength
M-153	A0A3B6SFZ3_WHEAT	GW8-B1;SBP-type domain-containing protein;Squamosa promoter-binding-like protein 16	(Wang *et al.*, 2009; Zeng *et al.*, 2018)	-0.3	0.006	fair*
M-153	A0A3B6RIB5_WHEAT	SBP-type domain-containing protein;GW8-B1	na	0.3	0.007	fair*
M-160	A0A3B5Z189_WHEAT	Polyubiquitin	(Dong *et al.*, 2006)	-0.64	0.000	moderate*
M-160	A0A3B5XX68_WHEAT	Helicase ATP-binding domain-containing protein;DNA helicase (EC 3.6.4.12)	(Tuteja *et al.*, 2012)	0.6	0.000	moderate*
M-160	A0A3B6C4M5_WHEAT	ANK_REP_REGION domain-containing protein;Calmodulin-binding transcription activator 5	(Vo *et al.*, 2015)	-0.48	0.000	fair*
M-160	A0A1D6QAD2_MAIZE	HAT transposon superfamily protein;rRNA N-glycosidase (EC 3.2.2.22)	na	-0.46	0.000	fair*
M-160	A0A3B6CH97_WHEAT	ERCC4 domain-containing protein	na	0.46	0.000	fair*
M-160	A0A2K2CML9_BRADI	DNA repair protein RAD51 homolog	na	0.38	0.001	fair*
M-160	A0A3B6C8P2_WHEAT	NAC domain-containing protein 8	(Hu *et al.*, 2010; Diao *et al.*, 2020)	-0.24	0.034	fair*
M-218	T1MVL7_TRIUA	CCAAT-binding transcription factor B (CCAAT-binding transcription factor family protein)	na	-0.28	0.014	fair*
M-218	A0A3B6C2Y1_WHEAT	CCAAT-binding transcription factor B (CCAAT-binding transcription factor family protein)	na	-0.25	0.028	fair*
M-239	A0A3B6B2U7_WHEAT	GATA-type domain-containing protein;Os04g0539500 protein	(Zhang *et al.*, 2021)	0.25	0.025	fair*
M-254	A0A3B5XZA1_WHEAT	Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9)	na	-0.33	0.003	fair*
M-254	A0A3B6LG92_WHEAT	Homeobox-leucine zipper protein HOX33	na	0.29	0.008	fair*
M-254	A0A3B6KUW3_WHEAT	START domain-containing protein;Homeobox-leucine zipper protein HOX10	na	-0.29	0.008	fair*
M-270	A0A3B6KKM0_WHEAT	Methyl-CpG-binding domain-containing protein 2;CW-type domain-containing protein;MBD domain-containing protein	(Hu *et al.*, 2011)	0.29	0.009	fair*
M-301	A0A3B6RMA1_WHEAT	NAC domain-containing protein	(Lv *et al.*, 2016)	0.22	0.050	fair*

Table 5.6: Homology-Based Identification and Classification of TAS Precursors.

Classification (*BLASTN*)	**TAS precursor** (CONTIG)
TAS3b in wheat	2A__598975000, 2D__450395000, 6A__406360000, 6B__446715000, 6B__446720000, 6D__285740000
TAS3a in other plants	4A__158175000, 4D__316535000

Table 5.7: Comparison of *Vrn-A1*-Dependent iRNA Functions in Winter- and Spring-Habit Wheat.

Allele **Cultivar** **Phenotype**	**Recessive *vrn-A1*** Requires vernalization **Winter-Habit** (Norstar ∩ Iso11W) Cold-tolerant Delayed flowering	**Dominant *Vrn-A1*** No vernalization required **Spring-Habit** (Manitou ∩ Iso8S) Rapid development Cold-independent flowering
Cultivar-specific iRNA (CR in bold face)	***miR9674***, ***miR5062***	***miR166***, ***miR167***, ***miR319***, ***miR393***, *miR396*, *miR9778*
Target	MEL1 (Argonaute; regulates meiosis and gene silencing) Histone H3-K9 (chromatin modification)	GID1 (gibberellin receptor) GRF (growth-regulating factor)
Enriched Biological Processes	Meiosis Histone modification (methylation, phosphorylation, acetylation) Gene silencing Dormancy Flowering delay (Microgametogenesis, Megagametogenesis) Vacuole transport Negative regulation of embryonic development	Gibberellic acid homeostasis Chromatin assembly DNA replication Oxidative stress response Bilateral symmetry Meristem growth Hormone metabolism Fatty acid degradation

Table 5.8: Cultivar-specific iRNA families. miR159 is a miRBase-defined miRNA family. tasiRNA families are defined by their origin from the same genomic segment homologous to known TAS genes. "Chr 4A (158175000)" denotes a precursor tag for a genomic segment on chromosome 4A, starting at position 158,175,000 and spanning 5,250 nucleotides. Other tags follow the same convention.

miRNA and tasiRNA family	**Manitou**	**Iso11W**	**Norstar**	**Iso8S**
miR159	S-217	M-159	M-073	M-073, M-238
Chr 4A (158175000)	S-378, S-361	–	S-274	S-221, S-361
Chr 6A (406360000)	–	–	S-265	S-447
Chr 6B (446715000)	S-129	S-129	S-358	S-358

5.7 Implications

Understanding the role of small RNAs in crop stress adaptation has critical implications for agricultural resilience. This chapter explores the coordinated prediction and analysis of miRNAs and siRNAs during cold acclimation in wheat, aiming to reveal regulatory iRNAs and their functional significance under realistic field conditions. By developing a pipeline to jointly analyze both small RNA types, this work bridges an often-overlooked gap in small RNA research and expands the toolkit for studying plant adaptation.

This methodology fulfills the second specific objective of this thesis by establishing a unified computational framework for the joint classification and analysis of miRNAs and siRNAs in wheat. It further contributes to the thesis by demonstrating that miRNA and siRNA co-regulation is not only feasible to study but essential for understanding the full regulatory network. It highlights how functional small RNA biomarkers can be evolutionarily conserved, stress-responsive, and closely linked to adaptation phenotypes. The outcomes of this work include:

- A unified computational framework for analyzing cold-responsive miRNAs and siRNAs in wheat.
- New insights into their regulatory interplay and evolutionary conservation during stress adaptation.
- Candidate iRNAs with potential utility in functional studies and breeding programs.

This analysis focuses on cold acclimation in wheat and is based on a large, well-designed field experiment. While the findings are robust within this context, broader applicability across species or stress types would require further validation. To address this, we compared our results extensively with existing literature to

highlight both conserved and novel regulatory elements.

In the wheat study, prediction validity and biological relevance were addressed through a layered strategy combining established annotation criteria, statistical modeling, and functional support rather than relying on single-step inference. miRNA candidates were identified using mirLibSpark under gold-standard plant miRNA criteria, while siRNAs were detected using the most widely adopted phase-scoring approach. Reproducibility was enforced through cross-library consensus across field-derived datasets, abundance filtering, and differential expression analysis using DESeq2 with false discovery rate control. Beyond differential expression, a customized differential regulation analysis was introduced to capture structured temporal patterns across cultivars and developmental stages. Functional relevance was further supported by target prediction restricted to sample-matched mRNA transcriptomes, updated UniProt-based annotation, pathway enrichment, and literature evidence for conserved and cold-responsive iRNAs. Together, these steps reduce false positives while explicitly defining the analytical assumptions under which biological interpretation is supported, and they illustrate how statistical and functional validation are integrated to compensate for environmental variability inherent to field-based data.

The co-regulation patterns identified between miRNAs and siRNAs are based on correlative data, such as expression profiles and shared target predictions. While suggestive of functional interaction, experimental validation remains necessary. One potential strategy involves transient expression assays in wheat protoplasts, introducing synthetic precursors or overexpression constructs of the candidate miRNA and siRNA individually and in combination. Changes in target mRNA levels could then be quantified using qRT-PCR to assess cooperative or antagonistic regulatory effects.

The computational framework used in this chapter relies on traditional scoring methods to quantify phasing signals for siRNA detection. While effective for identifying well-phased loci, these methods have not been systematically evaluated for their ability to capture diverse precursor properties associated with siRNA biogenesis. In particular, they often overlook structural, sequence, and contextual features that may distinguish functional phased loci. This limitation motivates the development of a more flexible and biologically informed approach, presented in the next chapter as `siWalk`, a biologically informed machine learning model designed to improve siRNA classification by integrating multiple biogenesis-related features.

CHAPTER 6

LEVERAGING MACHINE LEARNING FOR IMPROVED SIRNA PREDICTION IN ARABIDOPSIS

siRNAs have strong potential as biomarkers for gene regulation, silencing, and environmental responses, but computational limitations hinder their identification. Unlike miRNAs, which benefit from defined biogenesis pathways and curated datasets, siRNAs lack standardized resources, making their identification and interpretation challenging. Their complex precursors, which originate from diverse genomic regions and show condition-specific expression, often overlap with miRNAs. Current tools rely on phase-scoring and expression periodicity, but these heuristics lack robustness across species and conditions, limiting downstream reliability.

To fulfill the third specific objective—developing a robust, biologically informed framework for siRNA classification—this chapter presents `siWalk`, a machine learning approach guided by siRNA biogenesis features. Integrating sequence composition, precursor structure, phasing scores, and expression patterns, `siWalk` enables generalizable and interpretable prediction of phased siRNA loci. A key innovation is its evaluation of structural constraints by combining nucleotide context, secondary structure stability, and phasing, enhancing both accuracy and biological relevance. The model was trained on a curated *Arabidopsis* dataset spanning diverse library types and conditions to capture broad biogenesis signatures. The methodology consists of:

- Develop a machine learning–based classifier (`siWalk`) that leverages biogenesis-informed features to enable accurate prediction and classification of siRNAs.
- Identify key sequence and structural features that distinguish siRNAs from other small RNAs and optimize feature selection for improved classification.
- Validate `siWalk`'s predictions using *Arabidopsis* sRNA-seq datasets and benchmark against existing siRNA prediction tools.

The manuscript is currently under internal review and is being prepared for submission to a peer-reviewed journal. Chao-Jung Wu conducted the algorithm implementation, experimental design, and manuscript writing under the supervision of Professor Abdoulaye Baniré Diallo. The manuscript and `siWalk` manual are included in Appendix C.

6.1 Abstract

Plant small interfering RNAs (siRNAs) regulate gene expression but are difficult to identify due to confounding factors with other RNA molecules, limited resources, the absence of prediction criteria, and shortcomings in current prediction methods. Existing methods rely on phase scoring, which measures alignment periodicity, but this approach often fails to distinguish siRNAs from other small RNAs and leads to high false-positive rates. We propose to improve *Arabidopsis* siRNA prediction by incorporating more features derived from phased siRNA precursors beyond periodicity into a machine learning classifier. Leveraging a Random Forest classifier, siWalk integrates biogenesis knowledge with mathematical modeling of siRNA precursors. Its algorithms detect phased positions, define and tag phasing boundaries, compute key attributes, unify related segments, and rank siRNAs by calculating weights for their effector properties. SiWalk predicts siRNA precursors from sRNA-seq alignments with greater precision and accuracy than existing phase scoring methods. It validates highly conserved siRNA loci by highlighting the influential expressional and structural features for siRNA prediction. By focusing on structure-only features, siWalk expands its capacity to identify localizations of effector siRNAs in a sequence without requiring sRNA-seq data. SiWalk, with its robust learning and sorting kernel, enriches the toolkit for genome-wide annotation of phased siRNA precursors in diverse tissues and treatments and for selecting functional siRNAs in *Arabidopsis*.

Keywords: machine learning in siRNA prediction, siRNA localization, siRNA features, functional genomics, agritech

6.2 Introduction

Plant small interfering RNAs (siRNAs) play key roles in post-transcriptional gene silencing (Agrawal *et al.*, 2003; Li *et al.*, 2017), regulating critical processes such as development (Zhang *et al.*, 2019; Liu *et al.*, 2020), stress responses (Khraiwesh *et al.*, 2012) and pathogen defense (Llave, 2010; Zhang et Zhu, 2011). The generation of siRNAs begins with the formation of double-stranded RNA (dsRNA) from sources such as viral infections, transposable elements, RNA-dependent RNA polymerase (RDR) activity on single-stranded RNA (ssRNA) transcripts (Baulcombe, 2004; Vaucheret et Voinnet, 2024; Axtell, 2013a) or the self-folding of ssRNA (Kakrana *et al.*, 2018; Piriyapongsa et Jordan, 2008). Dicer-like enzymes cleave double-stranded RNA (dsRNA) into 21- to 24-nucleotide siRNAs (Kim *et al.*, 2005; Fukudome et Fukuhara, 2017) through sequential cycles. This process produces siRNA **phasing** clusters from dsRNA substrates, which are regularly spaced and periodically aligned on the precursor genome (Howell *et al.*, 2007; Chen *et al.*, 2007) (illustrated in Fig-

ure 6.1). In some pathways, microRNAs (miRNAs) trigger siRNA production by cleaving ssRNAs, which are then converted into dsRNA by RDR (Allen *et al.*, 2005; Axtell *et al.*, 2006; Chen *et al.*, 2010). Finally, selected siRNAs are loaded into the RNA-induced silencing complex (RISC), where they function as active **effectors** and guide the complex to silence target genes (John *et al.*, 2004; Dai *et al.*, 2018).

SiRNAs are promising biomarkers for monitoring plant health and improving agricultural outcomes. Many existing siRNA prediction methods essentially predict siRNA precursors from small RNA sequencing (sRNA-seq) data, and assign the most abundant siRNA species as the effector if its position corresponds to Dicer processing cycles. They rely on one of three main phase scoring schemes to evaluate siRNA production from expressed genomic regions: Howell (Howell *et al.*, 2007), Chen (Chen *et al.*, 2007) and Guo (Guo *et al.*, 2014). These scoring approaches represent the earliest efforts to disentangle confounding factors associated with other RNA molecules and identify phasing patterns among small RNAs aligned to a precursor. Howell *et al.* (2007) introduced the phase transformation score, a metric that quantifies phasing strength by linking the periodic alignment of RNAs to their abundance. Chen *et al.* (2007) achieved siRNA detection by using the hypergeometric distribution to evaluate the occupancy of in-phase positions by expressed siRNAs among all in-phase positions in a search window. This approach ignores sequencing read accumulation and instead emphasizes rigorous statistical assessment of phasing significance. Guo *et al.* (2014) advanced siRNA prediction with the software PhaseTank, which connects miRNA triggers to siRNA loci and their downstream targets—an area beyond the scope of this study. Guo's phase scoring approach integrates siRNA count, abundance and phased ratio to quantify phasing strength and quality.

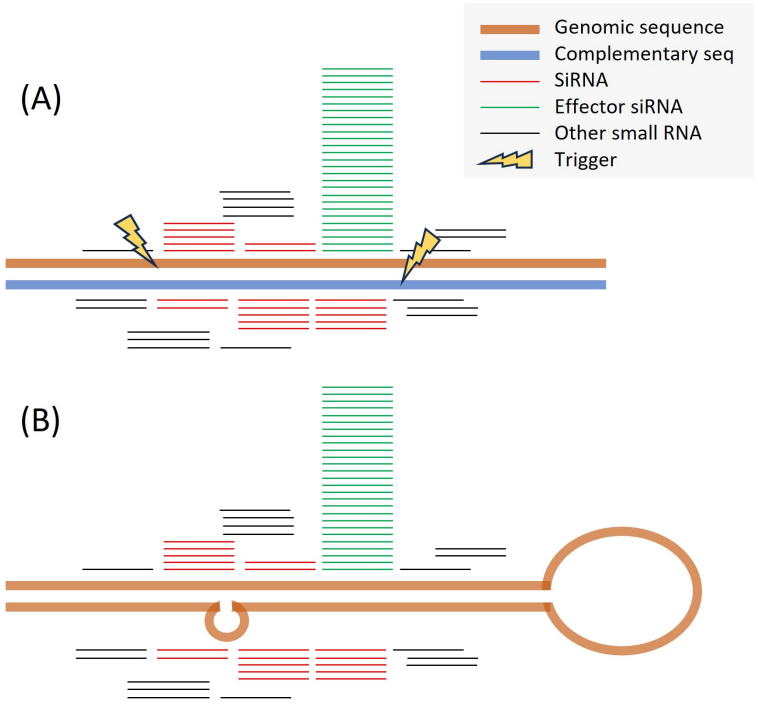


Figure 6.1: Phasing patterns of small RNAs aligned to genomic sequences that produce dsRNA precursors. Duplexes align periodically at regular intervals and arise from two contexts: (A) complementary genomic strands, and (B) a folded single-stranded RNA derived from one genomic strand. Other RNAs may align randomly. Duplex production may be triggered by signals such as miRNA-mediated cleavage. Phase scores quantify phasing patterns by measuring the periodic alignment of small RNAs.

While phase scoring methods have laid the foundation for subsequent advancements in siRNA prediction, they are not without notable limitations. First, their ability to differentiate phased RNAs from others has not been evaluated. Although originally developed in *Arabidopsis*, these scoring schemes have been widely adopted for plant siRNA discovery in various crops (Zhang *et al.*, 2012a; De Paoli *et al.*, 2009; Zheng *et al.*, 2014; Dotto *et al.*, 2014; Kakrana *et al.*, 2018). They typically rely on fixed thresholds without quantitative justification. Evaluating phase scores across varying thresholds would improve understanding of their optimal use for siRNA precursor prediction.

Second, interpreting phase scores remains challenging because they alone do not fully distinguish siRNAs from other RNAs that align to adjacent genomic regions. Since scoring schemes focus on alignment patterns, they often overlook the structural and biochemical properties that are crucial for siRNA biogenesis.

Members of the siRNA superfamily are traditionally classified based on various characteristics into distinct types, such as miRNAs, piRNAs (piwi-interacting), ta-siRNAs (trans-acting), nat-siRNAs (natural antisense), hc-siRNAs (heterochromatic) and mirtrons (intronic), among others (Piriyapongsa et Jordan, 2008; Farazi *et al.*, 2008; Rorbach *et al.*, 2018; Axtell et Meyers, 2018; Kakrana *et al.*, 2018; Agrawal *et al.*, 2003). The characteristics for classification include the mechanism of dsRNA formation, their genetic origin (e.g., coding, heterochromatic, terminal-repeat, intronic, viral), the post-transcriptional modifications they regulate (e.g., cleavage, methylation), precursor structure (e.g., hairpin, length) and tissue specificity (e.g., reproductive tissues). Incorporating these characteristics into a classifier would improve siRNA prediction beyond relying only on alignment. Building on the study by Leclercq *et al.* (2013), which improved miRNA prediction by training a Random Forest classifier on 100 diverse features, siRNA prediction can similarly benefit from leveraging machine learning's capacity to learn a wide range of features and to distinguish between classes.

Third, although singly phased miRNAs have well-defined prediction criteria (Meyers *et al.*, 2008; Axtell et Meyers, 2018; Wu *et al.*, 2019), the criteria for multiply-phased siRNAs remain poorly established. This gap contributes to higher false-positive rates and less reliable siRNA detection. Establishing such criteria falls within the domain of biology; however, identifying key features important for classification may support this effort.

Fourth, siRNA research is constrained by the availability of specialized software and databases. Unlike miRNA, which has over 160 analysis tools (Lukasik *et al.*, 2016), only a few are available for siRNA: Notable software includes ShortStack (Johnson *et al.*, 2016) and PHASIS (Kakrana *et al.*, 2017), and databases such as sRNAanno (Chen *et al.*, 2021a) and MPSS (Nakano *et al.*, 2020). A new prediction tool with enhanced performance would enrich the bioinformatics toolkit and advance siRNA research.

Fifth, existing resources assign different identifiers and coordinates to the same phasing events (Table 6.1), creating barriers to data standardization and comparative analyses. Assigning each locus a standardized genomic tag would resolve discrepancies in siRNA precursor identification and facilitate better integration across studies.

Sixth, for a given precursor, one might be curious about the potential siRNA production sites. However, sRNA-seq data, typically used by existing methods to predict precursors and assign effectors based on expression rather than directly predicting them, are not available in this scenario. Thus, localizing effector

sites directly from the precursor sequence requires a new classifier independent of sRNA-seq.

In this study, we present siWalk, a novel bioinformatics method that leverages machine learning to predict siRNAs with improved precision, accuracy, and biological insight. SiWalk addresses limitations of conventional methods by integrating comprehensive knowledge of siRNA biogenesis, including structural and biochemical properties that influence interactions and stability, and the periodicity and abundance of sequence alignments. These elements are incorporated into a hybrid machine learning and mathematical scoring method for precise precursor identification and effector localization. The primary contributions of this study include:

- SiWalk applies supervised machine learning to siRNA prediction by integrating current biogenesis knowledge with mathematical modeling. It predicts siRNA precursors from sRNA-seq data and improves precision and accuracy compared to existing phase-score–based methods. Furthermore, it predicts the most likely localizations of effector siRNAs within a sequence without requiring sRNA-seq data.
- Algorithmic procedures identify phased positions, define locus boundaries, compute key attributes, and merge related phasing segments.
- Feature analysis reveals key structural and nucleotide motifs potentially crucial for distinguishing true siRNAs, providing insights into siRNA biogenesis and biological relevance.
- A standardized genomic tag is proposed for reducing discrepancies in siRNA-generating loci identification across datasets.

By improving precursor prediction, advancing effector siRNA localization, refining phase boundary and tag, and demonstrating model generalization, siWalk enriches the toolkit for annotating genome-wide phased siRNA precursors in diverse tissues and treatments and reduces the barriers associated with selecting functional siRNAs in *Arabidopsis*.

6.3 Materials and methods

6.3.1 Datasets

6.3.1.1 Reference contigs of Arabidopsis siRNA-generating loci

Since no central repository for plant endogenous siRNA precursors was available, we obtained the coordinates of *Arabidopsis* phased siRNA precursors from four studies: 21 precursors from Chen *et al*. (2007), 18 from Howell *et al*. (2007), 37 from Nakano *et al*. (2020), and 35 from Chen *et al*. (2021a). All identified loci corresponded to 21-nt (nucleotide) siRNAs. Many loci from these sources overlapped but did not align exactly, likely due to variations in expression levels between libraries, precursor boundary definitions and annotation approaches.

To streamline the analysis, we pre-indexed the entire genome into **contigs**, each spanning 5250 nucleotides with a 250-nt shared region between consecutive contigs, ensuring that events at the junctions were captured. We unified overlapping precursors using the standardized contig tags to reduce ambiguity (Table 6.1). This process yielded 51 reference contigs of siRNA-generating loci, each 5250-nt long (Table 6.2).

Table 6.1: Example of a reference contig. A standardized genomic contig represents this Chromosome (Chr) 1 phasing locus, which has been assigned varying IDs and genomic coordinates in four different sources.

ID	Chr	Start	End	Length	Contig	Source
At1g62910	1	23299057	23300958	1901	1-23295000-23300250	(Chen *et al*., 2007; Howell *et al*., 2007)
PHAS21-7	1	23299479	23300850	1371	1-23295000-23300250	(Chen *et al*., 2021a)
AT_TAIR10.1.23299482.23300923	1	23299482	23300923	1441	1-23295000-23300250	(Nakano *et al*., 2020)

Table 6.2: Literature-supported *Arabidopsis* siRNA-generating contigs (Chen *et al.*, 2007; Howell *et al.*, 2007; Nakano *et al.*, 2020; Chen *et al.*, 2021a). (A) Complete list of reference contigs and (B) contigs of TAS (trans-acting siRNA) genes. *Notes*. All identified loci correspond to 21-nt Dicer call. No phasing locus was found in mitochondrial or chloroplast genomes. Refer to file **ath_ref_contigs.xlsx** for data source mapping.

(A) Reference contig			
1–4180000-4185250	1–23410000-23415250	3–1955000-1960250	5–5460000-5465250
1–4295000-4300250	1–23415000-23420250	3–1965000-1970250	5–9785000-9790250
1–4350000-4355250	1–23450000-23455250	3–1970000-1975250	5–15555000-15560250
1–4365000-4370250	1–23485000-23490250	3–5860000-5865250	5–15550000-15555250
1–5295000-5300250	1–23490000-23495250	3–6390000-6395250	5–15755000-15760250
1–17885000-17890250	1–23505000-23510250	3–6915000-6920250	5–16635000-16640250
1–17890000-17895250	1–23585000-23590250	3–8525000-8530250	5–17560000-17565250
1–18545000-18550250	1–25275000-25280250	3–8530000-8535250	5–17565000-17570250
1–23175000-23180250		3–9415000-9420250	5–20130000-20135250
1–23205000-23210250		3–9870000-9875250	5–23390000-23395250
1–23200000-23205250		3–9865000-9870250	5–24305000-24310250
1–23295000-23300250			
1–23300000-23305250	2–7190000-7195250	4–5760000-5765250	
1–23305000-23310250	2–11720000-11725250	4–8145000-8150250	
1–23385000-23390250	2–16535000-16540250	4–8380000-8385250	
(B) Contig of TAS			
1–18545000-18550250	2–11720000-11725250	3–5860000-5865250	5–20130000-20135250
	2–16535000-16540250	3–9415000-9420250	5–23390000-23395250

6.3.1.2 Arabidopsis sRNA-seq libraries

Arabidopsis sRNA-seq libraries were obtained from NCBI Gene Expression Omnibus repository: two from GSM518432 and GSM738727 (Lei et Sun, 2014), and 27 from GSE44622 (Jeong *et al.*, 2013). Using the least stringent criterion (at least three nearby phase positions occupied by 21-nt RNAs), 50 of 51 reference contigs were expressed in these 29 libraries.

To improve robustness and reduce study-specific annotation bias, these libraries were drawn from multiple

independent studies and encompass diverse experimental conditions and treatments. This heterogeneous collection was intentionally selected to ensure broad coverage of siRNA precursor expression scenarios, rather than condition-specific patterns. By aggregating libraries across studies, siWalk is trained on a wide range of expression contexts, which supports generalization despite inherent class imbalance in siRNA prediction tasks.

6.3.1.3 Curation of Arabidopsis positive and negative sets

As shown in Table 6.1, phasing events are dynamic with variable start and end positions at the same locus. To capture the variability across libraries and experiments and to provide examples for classifier training, we curated phasing segments with precise boundaries from sRNA-seq libraries. This approach ensured the retrieval of relevant precursor sequences and corresponding effector siRNAs for feature generation. sRNA-seq libraries were aligned to genome using Bowtie (Langmead *et al.*, 2009) allowing no mismatches, and partitioned into contigs using SAMtools (Danecek *et al.*, 2021). We computed six phase scores (Supplementary Section "Phase Scores") for each search window by analyzing the RNA alignments within that region. Within a contig, nearby in-phase segments and those sharing the same putative effector siRNA sequences were merged into a single segment (see Section 6.3.2). A general filter (requiring the segment to be expressed, phase positions to be occupied, and sufficient effector reads to be present) was applied to reduce instances for training efficiency. Significant phasing events were identified by modeling phase scores with gamma distributions and applying a local false discovery rate (FDR) threshold. A phasing segment was assigned to the positive set of siRNA-generating precursors:

- if its contig was among the reference contigs, and
- if at least two of its six phase scores were significant (FDR < 0.05), and
- if its **effseq** was present in at least two libraries, where effseq is the most abundant 21-nt siRNAs considering phase drift by 1 or 2 positions;
- otherwise, it was placed in the negative set.

The final training dataset, compiled from 29 *Arabidopsis* sRNA-seq libraries, includes 10683 positive and 174856 negative segment examples (Table 6.3).

Table 6.3: Training dataset.

Class	Segment count	%Total	Ratio	Corresponding to contig count
Positive	10683	5.76	1	49
Negative	174856	94.24	16	6034
Total	185539	100		By genome size ∼27000

6.3.2 Algorithms for determining phasing boundaries

Figure 6.2 illustrates how phasing boundaries are determined and the core steps of the implementation. An in-phase position is at a multiple of the $DicerCall$ interval from a starting position pos. A phased position is an in-phase position occupied by RNAs of $DicerCall$ length. The search window for a phasing segment spans from pos to $pos + DicerCall \times (cycle - 1)$. The booleans $pos.watson$ and $(pos + 2).crick$ denote the RNA duplex counterparts of $DicerCall$ length expressed on the Watson or Crick strand. List functions Min, Max and Unique return the minimum, maximum and unique values in a list. Length returns the length of the list. The default search window spans 169 nucleotides, corresponding to 9 Dicer cleavage $cycles$, each spacing 21-nt known as $DicerCall = \{18, 19, \ldots, 31\}$.

After identifying phased positions (Algorithm 1) and establishing locus boundaries (Algorithm 2), a function computes attributes (e.g., nucleotide frequency, folding mode), which are stored in the segment object. Segments with shared effector siRNAs are aggregated (Algorithm 3).

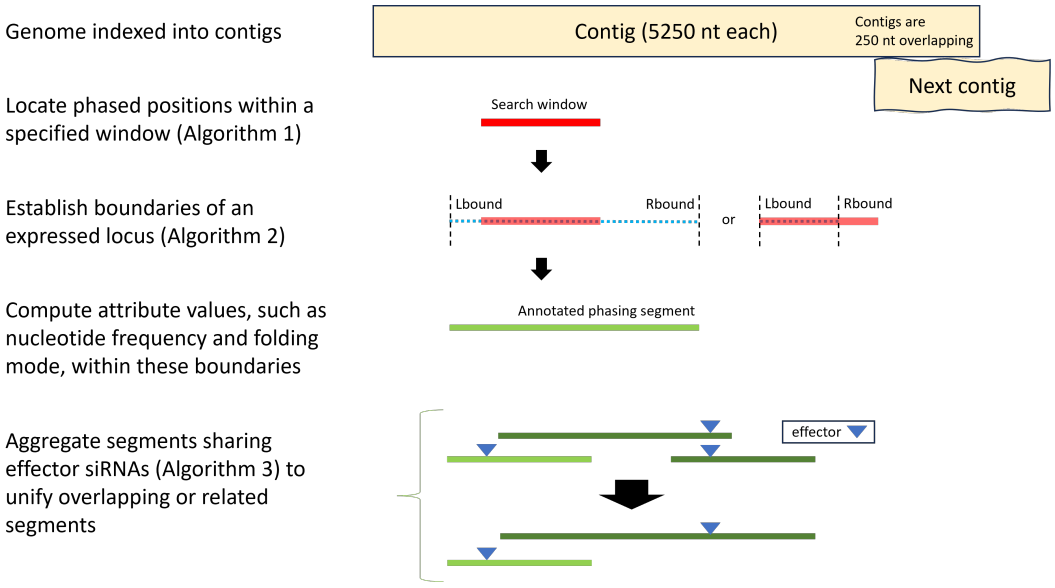


Figure 6.2: Illustration of algorithms for identifying phasing event boundaries. A phasing segment is identified through: Labeling the source contig, Identifying expressed phasing positions within an initial search window, Refining boundaries by searching nearby positions, Computing features to annotate phasing segments, and Aggregating segments that share effector siRNAs.

Algorithm 1 Get Phased Positions within a Window

function GetPhasedPositions(pos, $DicerCall$, $cycle$): list of integers
    Initialize an empty list *L* and a positive integer *posNext*
    for i in {0, 1, 2, .., $cycle - 1$} **do**
        posNext $\leftarrow pos + DicerCall \times i$
        if *posNext.watson* or *(posNext + 2).crick* **then**
            Append *posNext* to *L*
        end if
    end for
    return *L*
end function

Algorithm 2 Get boundaries of an expressed locus

function GetBoundaries(pos, $DicerCall$, $cycle$): 2 integers
    Initialize *Lbound* = 0, *Rbound* = 0 and *posTmp* = 0
    Initialize empty lists *L* and L_0
    ▷ Step 1: Collect expressed phase positions from *pos*
    L ← GetPhasedPositions(pos, $DicerCall$, $cycle$)
    ▷ Step 2: Scan nearby in-phase positions
    if Length(*L*) $>$ 1 **then**
        for *direction* in {1, -1} **do** ▷ 1 for downstream, -1 for upstream
            posTmp ← $pos + DicerCall \times cycle \times$ *direction*
            L_0 ← GetPhasedPositions($posTmp$, $DicerCall$, $cycle$)
            while Length(L_0) $>$ 1 **do**
                Append L_0 to *L*
                posTmp ← $posTmp + DicerCall \times cycle \times$ *direction*
                L_0 ← GetPhasedPositions($posTmp$, $DicerCall$, $cycle$)
            end while
        end for
        ▷ Step 3: Determine boundaries
        Lbound ← Min(*L*) $-DicerCall$
        Rbound ← Max(*L*) $+DicerCall$
    end if
    return Max(0, *Lbound*), *Rbound* ▷ Return adjusted Lbound and Rbound
end function

Algorithm 3 Aggregate segments

```
function SOFT_AGGREGATE(G: list of segment objects): list of segment objects
    Initialize empty lists L, T and M
    L ← Unique(segment.effpos in G)
    for effpos in L do
        Create a new segment object S with S.effpos ← effpos
        T ← segments in G where segment.effpos = effpos
        S.Lbound ← Min(Lbound of segments in T)
        S.Rbound ← Max(Rbound of segments in T)
                                          ▷ Aggregate remaining attributes with specific rules
        for all remaining attributes in segment do
            S.attribute ← Min/Max(attribute values in T)
        end for
        Append S to M
    end for
    return M
end function
```

6.3.3 Feature vectors and training

Each training example was represented by **245 raw features** (summarized in Table 6.5, see Supplementary Section "Feature definitions and complete list" for details).

These features, categorized into **expressional** and **structural**, were derived from the genomic alignment of small RNAs, encompassing primary sequence, abundance, alignment patterns, precursor structural layout, folding motifs, and molecular interactions. Features of interest included expression statistics from sRNA-seq alignments that capture phasing, such as phase scores, the count of phase positions expressed by 21-nt RNAs, their accumulation at phase positions, and strand-specific expression proportions from the precursor. Precursor length, G/C content, nucleotide composition, and mono-, di-, and tri-nucleotide frequencies within the primary effector siRNA sequence, as well as nucleotide types at key effector positions, were also considered (Lück *et al.*, 2019). These sequence-based features can influence the strand selection and RISC loading process, ultimately affecting siRNA functionality (Takeda *et al.*, 2008). Structural motifs (loops,

bulges, hairpins, triplet folding modes) (Leclercq *et al.*, 2013) and miRNA targeting (Axtell *et al.*, 2006) were included to capture the molecular interaction behavior of siRNA precursors, which may impact their stability and processing by Dicer-like enzymes.

Features were obtained using computational tools, string pattern searches and statistics. Phasing potential embedded in alignment patterns was evaluated by phase scores (Supplementary Section "Phase Scores"). RNAfold was used to determine the minimum free energy (MFE) and structural layout of candidate precursors. The precursor hairpin structure was analyzed by MIRcheck (Jones-Rhoades et Bartel, 2004). To identify miRNAs that trigger siRNA production, *Arabidopsis* miRNAs were retrieved from miRBase v22.1 (Kozomara *et al.*, 2019), and cleavage on candidate precursors was predicted using miRanda (John *et al.*, 2004).

We used scikit-learn (Pedregosa *et al.*, 2011) for data preprocessing, feature selection, classifier training and evaluation. Features were encoded into **1502 attributes** (Supplementary Section "Feature to attribute"). The rank scores, representing the importance of individual features for precursor classification, were provided by the chi-squared test (McHugh, 2013), and the most relevant attributes were selected by SelectKBest.

SiWalk utilizes a **Random Forest** classifier, which aggregates multiple decision trees on subsamples to improve overall accuracy (Breiman, 2001). The rationale for selecting it over alternative classifiers such as **Gradient Boosting** (Friedman, 2001), **Logistic Regression** (Yu *et al.*, 2011), and **Support Vector Machine** (Chang et Lin, 2011) is detailed in the Results section. Pairing with **AdaBoost**, which focuses on misclassifications and combines the weighted votes of weak classifiers (Freund et Schapire, 1997), can further enhance accuracy. The optimal parameters maximizing precision are listed in Table 6.4.

Table 6.4: Hyper-parameters of classifiers optimized for precision using Grid Search.

Classifier	Parameters
Gradient Boosting	learning_rate = 0.2, max_features = log2, min_samples_leaf = 4, and n_estimators = 200
Random Forest	max_depth = 30, max_features = log2, min_samples_leaf = 6, and n_estimators = 200
Logistic Regression	C = 0.1, max_iter = 10000, and solver = sag
Support Vector Machine	C = 0.01, degree = 4, gamma = 0.1, and kernel = poly

The performance of each classifier, trained on all or the subset of best attributes, was evaluated using a 5-fold cross-validation strategy. In standard split, data were split into training and testing sets based on class

labels to ensure proportional representation in each fold. In stratified split, data were divided to preserve the distribution of both class and library labels, ensuring testing on unseen libraries and reducing the risk of overfitting to specific library characteristics. Metrics assessed included precision, sensitivity, specificity, accuracy, Matthews correlation coefficient (MCC, addressing class imbalance) and F1 score, as described previously (Leclercq *et al.*, 2013).

Table 6.5: SiWalk feature summary. The 245 raw features are categorized by their expression-dependent and structural properties. The numbers of features considered for precursor prediction and effector localization are specified.

Feature property	Precursor prediction	Effector localization	Description
Expression:			
Phase score	14	0	Howell, Chen, and Guo scores, and phasing-related expression statistics
Ratio	1	0	Proportion of forward reads relative to the total reads
Nucleotide frequency	65	0	Expression-weighted nucleotide total frequency of the precursor or positional frequency on the effector
Abundance	2	0	Coordinate of the effector, the most abundant siRNA allowing for drift
Structural:			
Length	3	2	Precursor length and relative position of the effector siRNA on the precursor
Nucleotide type	46	46	Nucleotide composition at specific positions on the effector (as mono-, di-, or tri-nucleotides)
Base pairing	5	5	Percentage of bases forming base pairs around the effector (-4 to +4 from start and end of the effector on the precursor)
MFE	5	1	MFE of precursors at varying lengths
Triplet	44	44	Frequency of effector triplet folding mode
Hairpin	9	3	Hairpin-forming potential of the precursor
Bulge	25	25	Bulge length around the effector
Loop	23	23	Loop length around the effector
MiRNA targeting	3	3	Number and score of miRNA targeting the precursor

6.3.4 siRNA indication score

Drawing from Leclercq *et al.* (2013)'s approach to miRNA localization, which showed that scoring improved localization, we found that combining machine learning with scoring also enhances siRNA localization. Take

a 300-nt precursor, each nucleotide is a potential starting position for the effector siRNA. Considering siRNAs of 19 to 23 nt provides 1500 possible position-length effector candidates. To identify the best candidates, we adopted and refined their approach as follows.

The process first sorts candidates using machine learning classifications, followed by a secondary sorting based on indication scores. Presenting a short list (size = t) of top candidates increases the likelihood of including the true effector.

The classification for each candidate of length l at position p on the precursor (length = $n \gg t$) is determined by a machine learning model M, with the siRNA length *l* set to a range of $DicerCall \pm 2$, where $DicerCall = \{18, 19, .., 31\}$ represents the interval of Dicer cleavage. The range for l is $19 \leq l \leq 23$ if $DicerCall = 21$.

$$siRNA_{true} = \left\{ (p,\, l) \middle| M(p,l) = True,\ \ 19 \leq l \leq 23,\ 0 \leq p \leq n - l + 1 \right\} \tag{6.1}$$

The **indication score** is then used to sort $siRNA_{true}$ or to identify additional siRNAs until the list reaches size t. It is calculated in three steps (Leclercq *et al.*, 2013):

- The base score for each candidate is calculated as the sum of the products of importance, correlation, and value for each **structural** feature i associated with the siRNA (Chao *et al.*, 2022). Feature importance is determined by their ranking score from the chi-squared test. Correlation between a variable and a binary classification is assessed using the Point-Biserial Correlation (Kornbrot, 2014).

$$score(p,l) = \sum_i importance_i \times correlation_i \times value_i \tag{6.2}$$

- The consensus scores for p being the starting position $S(p)$ or the ending position $E(p)$ are defined as in (Leclercq *et al.*, 2013)

$$S(p) = \sum_{l=19}^{23} score(p,l) \tag{6.3}$$

$$E(p) = \sum_{l=19}^{23} score(p - l + 1, l) \tag{6.4}$$

- The most likely siRNA start and end positions are determined by the largest combinations of start and end position scores (Leclercq *et al.*, 2013). The $argsort$ function returns a sorted list of most likely siRNAs from a given precursor.

$$\{siRNA_1, siRNA_2, \dots\} = \left\{ (p,\, l) \middle|_{19 \le l \le 23,\ 0 \le p \le n-l+1} \underset{}{argsort} \; [S(p) + E(p + l - 1)] \right\} \tag{6.5}$$

6.4 Results

We developed siWalk, a hybrid classifier that leverages machine learning and integrates mathematical scoring to enhance siRNA prediction. SiWalk operates in two modes:

- **Prediction mode**, where it predicts genome-wide siRNA precursors from sRNA-seq data and assigns the effector siRNAs.
- **Localization mode**, where it evaluates a given long RNA sequence and predicts the most likely effector siRNA localizations.

We curated phased siRNA precursors originating from *Arabidopsis* reference contigs. A set of expressional and structural features were extracted from the examples and ranked by their importance for precursor classification. Using these features, machine learning models from various algorithms were trained and evaluated, and a final model was selected for its predictive performance. In localization mode, a sorting process was incorporated alongside machine learning to compensate for the absence of sRNA-seq data, as expression-derived phasing boundaries were unavailable for feature generation. Here, we presented the features for phased siRNA prediction, the training and evaluation of siRNA classifiers, and the delivery of siWalk.

6.4.1 Evaluation of individual predictive features

Feature construction was shaped by previous studies of miRNA duplex and hairpin characterization (Leclercq *et al.*, 2013), siRNA precursor prediction (Chen *et al.*, 2007; Howell *et al.*, 2007; Guo *et al.*, 2014) and properties relevant to siRNA production (Axtell *et al.*, 2006; Lück *et al.*, 2019; Takeda *et al.*, 2008). The predictive power of **1502 attributes**, generated through the encoding of **245 raw features**, was evaluated.

We observed that the attributes ranked as highly important by the chi-squared test generally showed strong positive or negative correlations with precursor classes (Figure 6.3). As shown in Table 6.6, important attributes for precursor classes included scoring (e.g., k, Guob, effector abundance, pvalb, Howell, Howellb), miRNA targeting, siRNA nucleotide compositions, precursor length and MFE. Key nucleotides included siRNA mono-, di-, and tri-nucleotides at the 5' end positions 1 and 2 (attribute names involving 5p1, 5p2), combinations of the first nucleotides at both ends (5p1, 3p3), and triplets at the center (md); They involved positions known to affect thermodynamics in determining precursor cleavage and strand selection (Takeda *et al.*, 2008). The nucleotide at position 3 of the 3' end (3p3) marked the cleavage site of the complementary strand. Triplet folding modes and nucleotide expression frequencies may also influence accurate precursor prediction. Since some attributes were structural and did not require expression data, using them was expected to help determine effector **localization** within the precursor. Among the key structural attributes (Table 6.6), miRNA targeting, precursor MFE, effector nucleotide composition, triplet folding mode, and the start position within the precursor could be significant factors driving effector siRNA production.

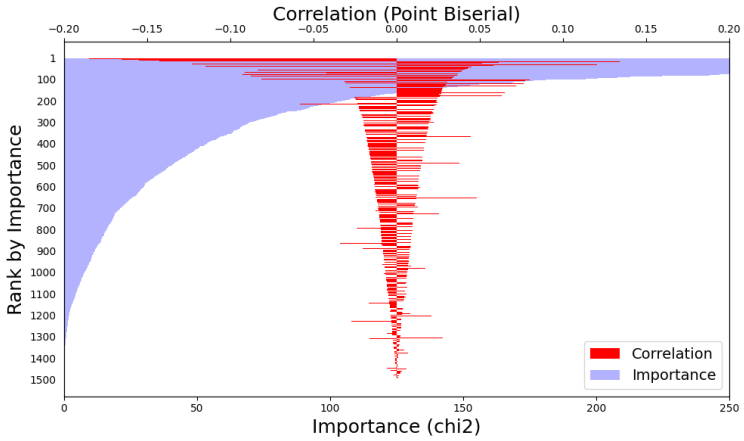


Figure 6.3: Attributes ranked by importance and their correlation with precursor prediction.

Table 6.6: Top predictive features for precursor **pred**iction and effector **loca**lization.

Attribute	Description	Property	**Usage**	Rank score
k	number of phase positions expressed by 21-nt RNAs within the 9-cycle window	phase score	pred	3822.9
ext_k	updated number of phasing positions after increasing the boundary beyond the 9-cycle window	phase score	pred	2447.4
n	number of all positions expressed by 21-nt RNAs within the 9-cycle window	phase score	pred	1143.5
Guo_b_ccdf	Guob, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	1045.7
twohitBool_True	miRNA targets the precursor twice or more	miRNA targeting	pred, loca	999.5
eff_frq	frequency of the most abundant effector siRNA allowing drift	phase score	pred	850.3
length	length of the precursor	length	pred	807.5
md3C	expression-weighted percentage of a nucleotide base type at **md3** on effector	nucleotide frequency	pred	803.0
eff5p1di_TA	di-nucleotide base type at **5p1** on effector	nucleotide type	pred, loca	783.1
BestScore	the highest targeting score among miRBase microRNAs	miRNA targeting	pred, loca	715.0
md2C	expression-weighted percentage of a nucleotide base type at **md2** on effector	nucleotide frequency	pred	705.0
3p3G	expression-weighted percentage of a nucleotide base type at **3p3** on effector	nucleotide frequency	pred	696.8
eff5p1tri_TAA	tri-nucleotide base type at **5p1** on effector	nucleotide type	pred, loca	669.7
pval_b	pvalb, the minimum score in a precursor's scanning windows	phase score	pred	639.8
max_mfe	the maximum MFE of folding sequences in a precursor's scanning windows	MFE	pred	616.4
premfe	MFE of the precursor in [Rbound, Lbound], calculated by RNAfold	MFE	pred, loca	614.6
NBtripletC	occurrences of triplet C around eff_seq	triplet	pred, loca	554.6
5p1T	expression-weighted percentage of a nucleotide base type at **5p1** on effector	nucleotide frequency	pred	531.9
5p1C	expression-weighted percentage of a nucleotide base type at **5p1** on effector	nucleotide frequency	pred	527.1
eff5p2tri_ATG	tri-nucleotide base type at **5p2** on effector	nucleotide type	pred, loca	525.3
eff5p1_T	mono-nucleotide base type at **5p1** on effector	nucleotide type	pred, loca	514.0
eff5p1_3p3_TG	the first and the last nucleotides (**5p1, 3p3**) on effector	nucleotide type	pred, loca	513.5
effmd4tri_AAG	tri-nucleotide base type at **md4** on effector	nucleotide type	pred, loca	499.7
GC%	expression-weighted percentage of G and C combined in a precursor	nucleotide frequency	pred	487.3
effmd1tri_ATG	tri-nucleotide base type at **md1** on effector	nucleotide type	pred, loca	485.6
Howell_ccdf	Howell, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	481.1
5p4C	expression-weighted percentage of a nucleotide base type at **5p4** on effector	nucleotide frequency	pred	479.1
Howellb_ccdf	Howellb, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	479.0
effmd2tri_GTA	tri-nucleotide base type at **md2** on effector	nucleotide type	pred, loca	470.9
3p4C	expression-weighted percentage of a nucleotide base type at **3p4** on effector	nucleotide frequency	pred	454.4
effmd2tri_AGA	tri-nucleotide base type at **md2** on effector	nucleotide type	pred, loca	450.2
eff_pos	coordinate of the most abundant effector siRNA allowing drift	abundance	pred	449.1
dist_5p	eff_seq relative position on precursor	length	pred, loca	447.9
5p5C	expression-weighted percentage of a nucleotide base type at **5p5** on effector	nucleotide frequency	pred	443.8
eff3p3_G	mono-nucleotide base type at **3p3** on effector	nucleotide type	pred, loca	427.4
dist_3p	eff_seq relative position on precursor	length	pred, loca	427.4
5p2T	expression-weighted percentage of a nucleotide base type at **5p2** on effector	nucleotide frequency	pred	421.3
anyhitBool_False	boolean, any miRNA targeting this precursor	miRNA targeting	pred, loca	409.6
5p1G	expression-weighted percentage of a nucleotide base type at **5p1** on effector	nucleotide frequency	pred	396.4
effmd5tri_TGG	tri-nucleotide base type at **md5** on effector	nucleotide type	pred, loca	387.9
5p2C	expression-weighted percentage of a nucleotide base type at **5p2** on effector	nucleotide frequency	pred	381.4
eff3p1tri_GGT	tri-nucleotide base type at **3p1** on effector	nucleotide type	pred, loca	372.3
effmd4tri_TTG	tri-nucleotide base type at **md4** on effector	nucleotide type	pred, loca	367.7
eff3p3tri_GAG	tri-nucleotide base type at **3p3** on effector	nucleotide type	pred, loca	361.9
eff3p3di_AG	di-nucleotide base type at **3p3** on effector	nucleotide type	pred, loca	353.8
eff5p4tri_GGA	tri-nucleotide base type at **5p4** on effector	nucleotide type	pred, loca	337.2
md4T	expression-weighted percentage of a nucleotide base type at **md4** on effector	nucleotide frequency	pred	333.8
eff5p1_3p3_GA	the first and the last nucleotides (**5p1, 3p3**) on effector	nucleotide type	pred, loca	332.9

To explore whether these important features can support siRNA precursor classification, we examined how classifier performance changes with increasing features. Selecting top-ranked features from a set of 245 showed that predictive performance plateaued between 45 and 115 features with reduced computation time (Figure 6.4), suggesting that this range captures the essential features for classification. Considering the performance stabilization within this range and the practical constraints of evaluating multiple subsets, we selected 100 attributes and extended the analysis to various classifiers. As shown in Table 6.7A, this subset often reduced computation time across classifiers compared to using the full set. Further evaluations are presented in the following sections.

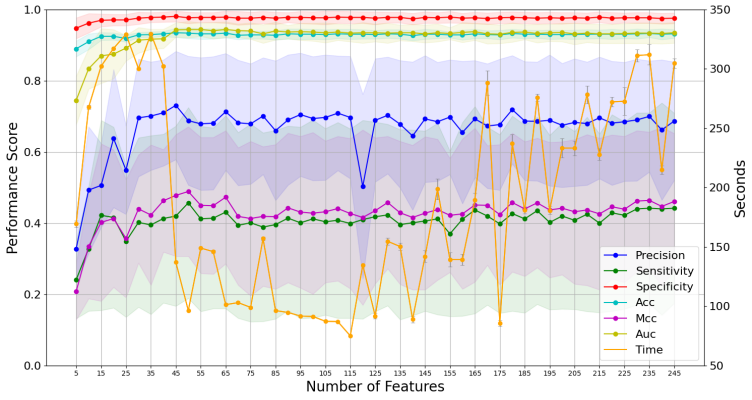


Figure 6.4: Relationship between the number of features and classifier performance. The Gradient Boosting classifier with AdaBoost was evaluated on a reduced training dataset (10093 examples) using 5-fold cross-validation to assess performance of leading features at five intervals of increasing feature counts.

Table 6.7: Time efficiency of various prediction methods. (A) Total time spent on Grid Search with 5-fold cross-validation for a dataset of 185539 examples. Using a subset of 100 attributes often reduces computation time compared to using the full set with the same classifier. (B) Average prediction time per library across 29 libraries. RFAs100, which uses 100 structural features, processes libraries more slowly than ShortStack and PHASIS due to its larger feature count and the additional time required for feature generation. Abbreviations *na*: not applicable (total time, run once); *nd*: not determined; Gradient Boosting with 100 features (GB100), all features (GBall) and AdaBoost (GBA); GBA with 100 structural features (GBAs100); Random Forest (RF); Logistic Regression (LR); Support Vector Machine (SVC).

(A) Grid Search	Total (minutes)	Std Dev
GB100	2.2	na
GBall	7.7	na
GBAs100	34.7	na
GBAall	62.2	na
RF100	7.4	na
RFall	6.4	na
RFAs100	165.7	na
RFAall	303.7	na
LR100	0.5	na
LRall	16.3	na
SVC100	265.7	na
SVCall	1900.7	na
(B) Per Lib	Average (minutes)	
ShortStack	12.5	6.9
PHASIS	0.4	nd
RFAs100	86.-	101.-

6.4.2 Training and evaluation of siRNA classifiers

We assessed the predictive performance of several machine learning models, including Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVC), each paired with complete or partial feature subsets, or structural-only attributes (Table 6.8). These models were compared to phase scoring methods with FDR and established siRNA prediction software, ShortStack v3.8 and PHASIS v3.3. Since segment-level information was not available for external software, we mapped their predictions onto contigs for comparing with the 50 expressed reference contigs. To maintain a 1:16 class ratio similar to our training dataset (Table 6.3), we arbitrarily assigned 818 negative contigs.

Although the actual class distribution in real-world scenarios is unknown, the 1:16 ratio reasonably approximates the assumption that most genomic segments express or regulate other RNA types, with only an unspecified fraction producing siRNAs. We believe that modeling siRNA production on an imbalanced dataset reflects the biological reality, and ensures better applicability of the classifier.

Table 6.8: Performance evaluation of various classifiers with imbalanced examples (5.76% positive, 94.24% negative) **(A)** Phase scoring with FDR < 0.05, predicting a segment dataset of 185539 examples. **(B)** Existing phase scoring software such as ShortStack (Guo $>= 30$) and PHASIS (pval $<= 0.005$), predicting 50 expressed reference contigs and 818 negatives contigs across 29 sequencing libraries. **(C)** Machine learning using either the top 100 attributes, the full feature set, or the top 100 structure-derived attributes, predicting the same segment dataset with 5-fold cross-validation. **(D)** Machine learning with representative models (GBA100 and RFAs100), training and predicting the library-stratified segment datasets with 5-fold cross-validation. Abbreviations: Gradient Boosting with 100 features (GB100), all features (GBall) and AdaBoost (GBA); GBA with 100 structural features (GBAs100); Random Forest (RF); Logistic Regression (LR); Support Vector Machine (SVC); GBA100 on library-stratified datasets (GBA100-GS); Machine learning (ML).

(A) Phase score	TP%	TN%	FP%	FN%	Precision	Sensitivity	Specificity	Accuracy	MCC	F1
Howell	1.022	90.841	3.401	4.736	0.231	0.177	0.964	0.919	0.160	0.201
Howellb	1.739	79.770	14.472	4.019	0.107	0.302	0.846	0.815	0.094	0.158
Guo	5.725	7.082	87.161	0.033	0.062	0.994	0.075	0.128	0.063	0.116
Guob	5.740	18.608	75.634	0.018	0.071	0.997	0.197	0.243	0.116	0.132
pval	0.579	89.895	4.347	5.179	0.118	0.101	0.954	0.905	0.059	0.108
pvalb	1.075	88.093	6.150	4.683	0.149	0.187	0.935	0.892	0.109	0.166
(B) Existing software										
ShortStack (Guo)	0.727	87.216	7.024	5.033	0.385	0.126	0.926	0.879	0.152	0.167
PHASIS (pval)	2.275	91.694	2.546	3.485	0.599	0.395	0.973	0.940	0.427	0.422
(C) ML model										
GB100	3.446	93.248	0.994	2.312	0.805	0.598	0.989	0.967	0.671	0.673
GBall	2.720	93.456	0.786	3.038	0.798	0.472	0.992	0.962	0.591	0.583
GBA100	4.460	93.225	1.017	1.298	0.857	0.775	0.989	0.977	0.794	0.795
GBAs100	3.270	93.401	0.841	2.488	0.802	0.568	0.991	0.967	0.657	0.662
GBAsall	4.755	93.427	0.815	1.003	0.873	0.826	0.991	0.982	0.836	0.840
RF100	2.673	93.835	0.407	3.085	0.894	0.464	0.996	0.965	0.622	0.597
RFall	1.444	94.140	0.102	4.313	0.945	0.251	0.999	0.956	0.469	0.390
RFA100	2.915	93.813	0.430	2.843	0.904	0.506	0.995	0.967	0.653	0.631
RFAs100	3.840	93.806	0.436	1.918	0.911	0.667	0.995	0.976	0.766	0.765

Table 6.8 continued from previous page

	TP%	TN%	FP%	FN%	Precision	Sensitivity	Specificity	Accuracy	MCC	F1
RFAall	3.473	93.949	0.293	2.285	0.939	0.603	0.997	0.974	0.735	0.721
LR100	1.970	92.919	1.323	3.787	0.680	0.342	0.986	0.949	0.445	0.431
LRall	2.959	92.297	1.945	2.798	0.643	0.514	0.979	0.953	0.543	0.555
SVC100	2.083	93.870	0.372	3.675	0.880	0.362	0.996	0.960	0.542	0.501
SVCall	5.051	92.362	1.880	0.707	0.774	0.877	0.980	0.974	0.806	0.813
(D) Stratified										
GBA100-GS	2.677	93.392	0.453	3.478	0.857	0.430	0.995	0.961	0.589	0.569
RFAs100-GS	3.958	93.495	0.349	2.198	0.923	0.641	0.996	0.974	0.757	0.756

6.4.2.1 Boosting predictive power

As shown in Table 6.8A-B, phase scoring methods provided low F1 and MCC, highlighting their limitations. Although Guo identified most positives (TP approaching 5.76%), it suffered from low specificity due to a bias towards predicting positives. The accuracies of Howell, Guo, pval, ShortStack, and PHASIS remain below 94.24%, making them worse than random guessing (Figure 6.5). In contrast, machine learning approaches delivered higher precision, sensitivity, and accuracy, with fewer false positives (Table 6.8C). Take GBA100, it achieved an *accuracy gain* of 0.0346 over random guessing (60% of maximum 0.0576); other models showed similar gains. Recognizing that preselected fixed thresholds may not be optimal, we assessed ROC curves of scoring methods across various score thresholds when predicting precursor coordinates from 29 sRNA-seq libraries (Figure 6.6). We compared these methods to GBA100 by varying probability thresholds. GBA100 achieved a higher area under the curve, consistently outperforming all phase scoring methods except PHASIS, highlighting the enhanced ability of machine learning to differentiate between positives and negatives.

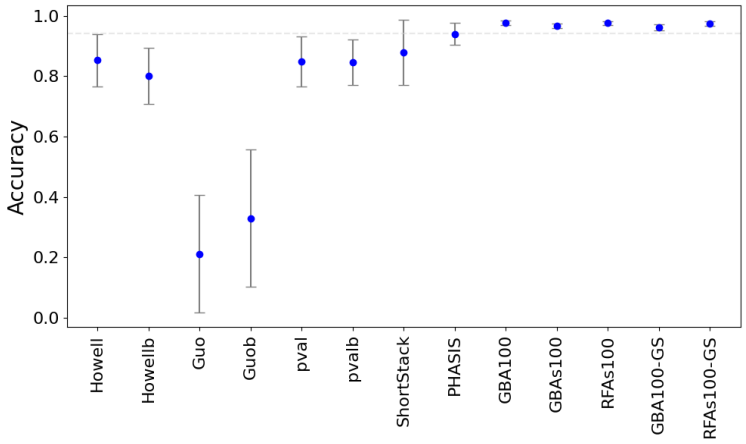


Figure 6.5: Accuracies of selected classifiers compared with random guessing (0.9424). Since negative examples make up 94.24% in the dataset, a naive classifier that always predicts negative would achieve an accuracy of 0.9424. In contrast, true random guessing under the observed class distribution (94.24% negative, 5.76% positive) yields an expected accuracy of approximately 0.8914, computed as $p_{\text{neg}}^2 + p_{\text{pos}}^2$.

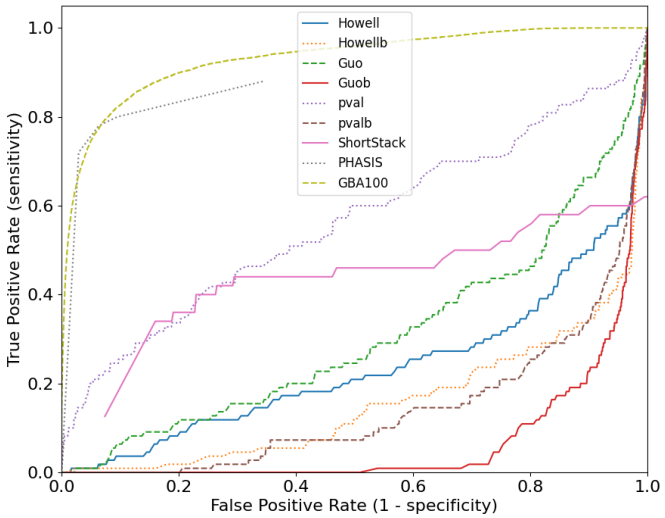


Figure 6.6: ROC curves (Receiver-Operating Characteristic) of selected classifiers predicting from sRNA-seq libraries, showing their ability to distinguish true positives from false positives.

6.4.2.2 Minimizing false positives

Given the predominance of the negative class in the training set, a naive classifier is inclined to predict examples as negatives over positives. Therefore, we prioritized **precision** to minimize false positives, at the same time considering **MCC** to address class imbalance and enhance overall performance, along with runtime efficiency when selecting the final classifier. The trade-off between precision and MCC is illustrated in Figure 6.7. Phase scoring methods clustered in the lower-left corner of the plot, showing poor precision; They detected true positives but struggle with false positives, likely due to reliance on limited expression features. ShortStack had an MCC of 0.152, indicating marginal improvement over random guessing (MCC = 0), and a precision of 0.385, indicating many incorrect positive predictions. PHASIS, located at the center of the plot, performed reasonably well with an MCC of 0.43 and a precision of 0.6 but may still be unsatisfactory. In contrast, all machine learning methods outperformed the scoring methods, with the GBA and RFA families leading in precision and MCC, as shown in the upper-right corner of the plot, suggesting superior predictive capabilities.

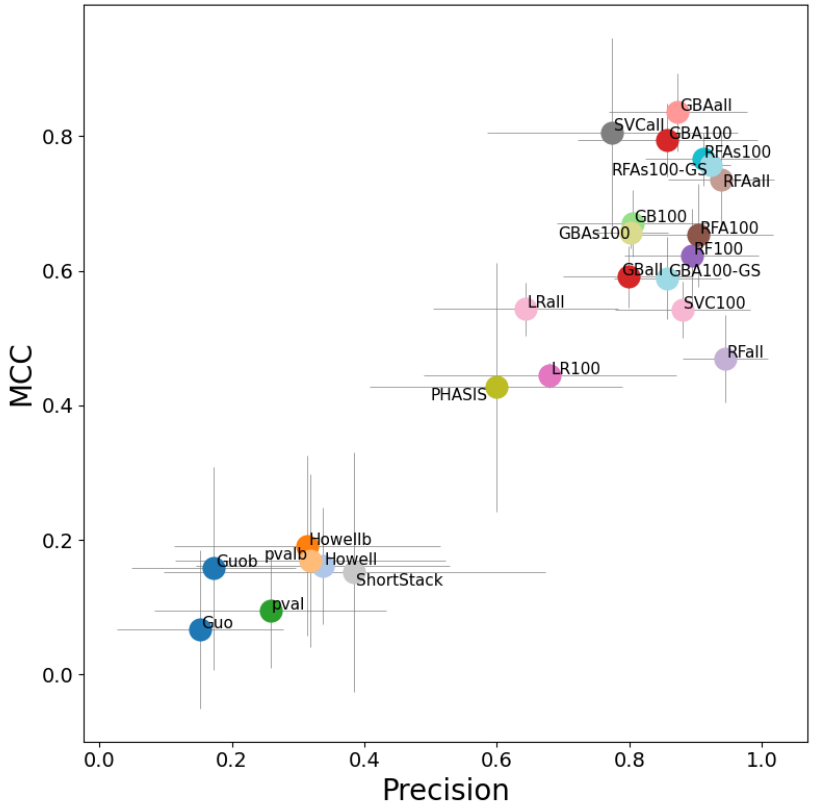


Figure 6.7: Comparison of Precision and MCC among classifiers. GB, LR, RF, and SVC refer to Gradient Boosting, Logistic Regression, Random Forest, and Support Vector Machine.

6.4.2.3 Versatile classifier for precursor and effector prediction

The **GBA100** model, using the 100 best overall attributes, achieved a precision of 0.857 and an MCC of 0.794. The **RFAs100** model, using 100 best structural attributes, offered a precision of 0.911 and an MCC of 0.766. These two models emerged from the GBA and RFA families as the leading candidates for final consideration. While both models performed strongly, RFAs100 excelled at detecting true positives, and GBA100 provided better overall performance. To facilitate the prediction of siRNA production sites for a given precursor, **RFAs100 is the optimal choice, offering a versatile model** that relies solely on static structural features without reliance on expression features.

6.4.2.4 Generalization across experiments

The *Arabidopsis* libraries spanning various tissues and treatments were sourced from distinct studies (Lei et Sun, 2014; Jeong *et al.*, 2013). To test model generalization across diverse experimental conditions, preparations, sequencing depths, and biological contexts, performance was evaluated on unseen libraries via stratified library sampling (Table 6.9, Figures 6.5 and 6.7). Compared to counterpart models trained on standard split datasets, GBA100-GS and RFAs100-GS models maintained strong precision and accuracy on library-stratified datasets, with a slight decrease in MCC. Similar performance between standard and stratified cross-validation highlights model robustness and the ability to generalize.

Table 6.9: Performance between standard and stratified split strategies. Extracted from Table 6.8. *GS*: training on library-stratified datasets.

	TP%	TN%	FP%	FN%	Precision	Sensitivity	Accuracy	MCC
GBA100	4.460	93.225	1.017	1.298	0.857	0.775	0.977	0.794
GBA100-GS	2.677	93.392	0.453	3.478	0.857	0.430	0.961	0.589
RFAs100	3.840	93.806	0.436	1.918	0.911	0.667	0.976	0.766
RFAs100-GS	3.958	93.495	0.349	2.198	0.923	0.641	0.974	0.757

6.4.2.5 Reliable prediction of conserved precursors

The **RFAs100** and GBA100 models showed strong reliability in identifying conserved precursors, as demonstrated by their substantial overlap with highly conserved reference contigs (Figure 6.8, left panel) and agreement with predictions from established methods (Figure 6.8, middle and right panels). This suggests

that both models are reliable tools for genomic precursor predictions, producing results consistent with established benchmarks

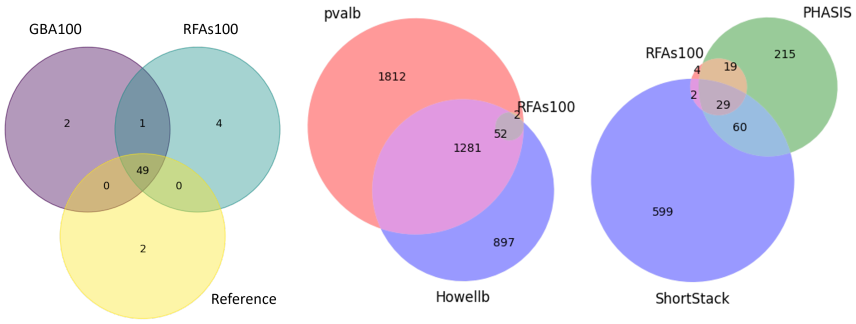


Figure 6.8: Comparison of contigs predicted by siWalk RFAs100 classifier with reference contigs, GBA100 classifier, scoring methods Howellb and pvalb, and existing tools ShortStack and PHASIS. Background consists of 6083 contigs based on the training dataset; or ∼27000 contigs based on genome size (Table 6.3).

6.4.3 Effector siRNA localization in precursor

We selected **RFAs100**, a machine learning model that can predict precursors without relying on expression-derived features. As most existing siRNA predictors rely on expression data and focus on precursors, we tested whether this model could locate effector production sites without such data. We developed a siRNA sorting process (see Section 6.3.4), powered by the **RFAs100** classifier and the indication score, to predict potential effectors directly from sequence information. We demonstrated it using the ta-siRNA sequence UUCUUGACCUUGUAAGACCCC (Chen *et al.*, 2021a; Fahlgren *et al.*, 2006). The genomic sequence surrounding the ta-siRNA was extracted, and siRNA candidates of various lengths at each position were evaluated using RFAs100, along with indication scores computed from $S(p)$ and $E(p)$ (Figure 6.9). Among the six identified effector candidates, three clustered near the known site, and two were positioned ∼20 nucleotides downstream. This process successfully approximates effector localization, showcasing its ability to predict biologically relevant siRNA localization.

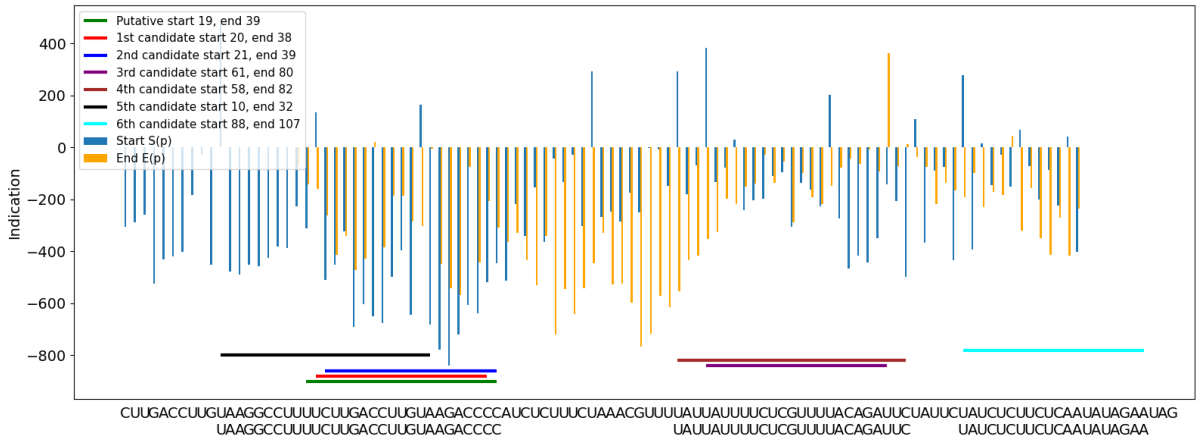


Figure 6.9: Effector siRNA localization in an example precursor. Six candidates are identified by siWalk as the most likely effectors for a precursor (Chr3, positions 5862187-5862334), extracted and trimmed around the known ta-siRNA sequence UUCUUGACCUUGUAAGACCCC (positions 19-39). This ta-siRNA is derived from the TAS3 precursor PHAS21-21 (Chen *et al.*, 2021a).

We assessed the offsets of the best (1st) candidates for each available positive precursor and an equal number of randomly selected negatives from the training dataset (Figure 6.10). To facilitate the analysis, each precursor was standardized to 121 nucleotides, with the start position of putative siRNAs centered at position 50 ± 5 for consistent comparison. SiRNAs located far from the center were excluded. This yielded 6386 positive examples from an initial 10683. The negative set exhibited a uniform distribution of offsets ranging from -50 to 50 nucleotides, and the positive set demonstrated a clear peak near zero between -10 to 10. The distributions between the two populations were significantly different, suggesting that the sorting process reliably identifies potential effector siRNAs.

6.4.4 The siWalk program

SiWalk is an open-source python program available on GitHub (https://github.com/bioinfoUQAM/siWalk). It offers modules for genome-wide precursor prediction, effector localization without expression data, and customized training using user-labeled data, as shown in Figure 6.11.

SiWalk can either use pre-trained models provided with the program or be trained on a user-labeled pre-

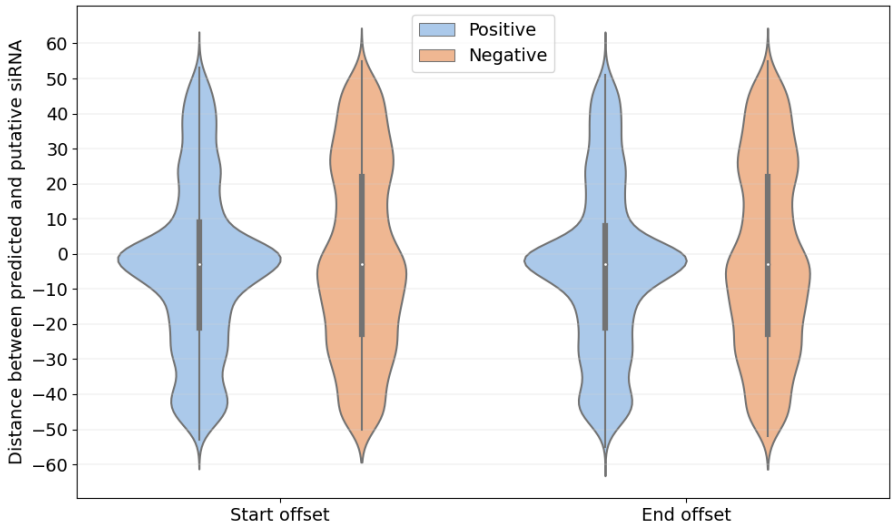


Figure 6.10: Offset distribution of the 1st effector candidates for 6386 positive and negative precursors, each standardized to 121 nucleotides with putative siRNAs centered. Significant differences in the distributions of absolute offsets between positive and negative populations were confirmed by the Kruskal-Wallis test (Kruskal et Wallis, 1952): Start offset $p = 2.97 \times 10^{-67}$; End offset $p = 5.25 \times 10^{-73}$.

cursor dataset. Due to its 3.6 GB size, RFAs100 is not distributed with siWalk; instead, the instructions for model generation are provided. The GBA100 and GBAs100 models are available for reference. Optionally, users can label segment classes to create a training dataset, and a script is provided to train a custom model. Training on a dataset derived from 29 *Arabidopsis* libraries takes approximately 30 minutes on a single 3.2 GHz CPU.

In **prediction** mode, siWalk accepts SAM files as input, which are generated by aligning sRNA-seq libraries to a reference genome. The feature annotation stage takes 86 $\pm$ 101 minutes per library using 64 CPUs at 2.40 GHz with Spark parallel processing (Zaharia *et al.*, 2016), based on processing 29 *Arabidopsis* SAM files. The precursor prediction stage takes around 1 minute per annotated *Arabidopsis* library. Although siWalk requires extra time for feature generation, it can be sped up by adding CPUs. SiWalk's robustness and generalization enable users to apply it to investigate genome-wide siRNA precursors expressed in their own sequencing datasets.

In **localization** mode, siWalk accepts a long nucleotide sequence as input and returns six siRNA candidates. Predicting siRNA from a 121-nt precursor takes approximately 8 minutes. Users can provide a precursor of

any length to predict the localization of effectors, but note that processing time increases with precursor length since the scanning is conducted at each nucleotide.

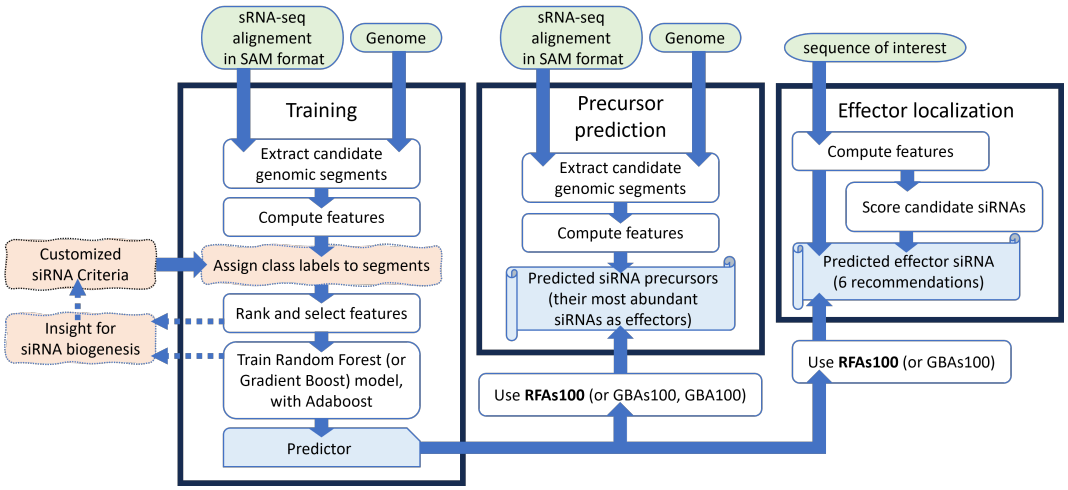


Figure 6.11: Workflow of the siWalk algorithm. Three modules are provided. **Precursor prediction**: genome-wide prediction of precursors and assigning effectors. **Effector localization**: predicting effector localization in a sequence. **Training**: training a custom model with user-labeled siRNA examples. The miRBase as an input for computing features is not shown. Wiggled boxes indicate operations during training that could be performed by the user.

6.5 Discussion

Predicting siRNAs from precursors is complex due to the difficulty of pinpointing precise Dicer cleavage sites and understanding how the active effector is selected by RISC. While the formation of double-stranded siRNA precursors is increasingly understood, the sequence and structural signals guiding these processes remain unclear. As portrayed in this study, siWalk is an accurate, reliable and generalized tool for the prediction of conserved siRNA precursors and for localizing effectors within a sequence. SiWalk identifies potential siRNA precursors across the entire genome, enabling researchers to investigate siRNA expression patterns and explore regulatory mechanisms influenced by RNA interference. This facilitates the study of siRNAs' involvement in stress responses and other regulatory processes. By revealing influential features, siWalk facilitates biologists to hypothesize the relationship between biogenesis and feature importance and to uncover their evolutionary origins. These insights could inform the development of refined criteria for siRNA prediction, rooted in the biological significance of key features. The siRNA production sites identified by siWalk could also guide RNA interference vector design. This could enhance the amplification of desired

siRNA, thereby reducing off-target toxicity. On the other hand, its framework is tailored and benchmarked for regularly phased siRNAs, restricting its capacity to predict irregularly phased siRNAs, which arise from variations in precursor origins. It is possible to apply siWalk in other plant species, but its scope is constrained by the absence of comprehensive ground truth datasets similar to miRBase for miRNAs. Therefore, the training in this study has been limited to *Arabidopsis*, leaving its applicability to other plant species to be determined. Despite some limitations, accurate siRNA precursor and localization prediction by siWalk opens opportunities to drive advancements in identifying biomarkers of plant health and agricultural innovations, such as enhancing crop resilience and productivity.

6.6 Implications

Accurate identification of functional siRNAs is vital for understanding RNA interference pathways and for leveraging siRNAs as biomarkers or therapeutics. This chapter introduced `siWalk`, a machine learning–based framework trained on biogenesis-related features including phasing, expression, and precursor structure. By integrating these signals, `siWalk` enables more robust, interpretable, and generalizable predictions than traditional rule-based systems. This methodology fulfills the third specific objective of the thesis by introducing a biologically informed machine learning model that improves the accuracy and robustness of siRNA classification. It further contributes to RNA biomarker prediction by combining biogenesis-aware features with computational rigor, enabling interpretable and generalizable predictions. The advances of this methodology include:

- A machine learning classifier that improves siRNA prediction accuracy.
- A curated training dataset of validated phased siRNA loci from diverse *Arabidopsis* libraries.
- An interpretable model that reveals biogenesis-relevant features and supports rational siRNA precursor design for therapeutic or experimental delivery vectors.

To mitigate overfitting, siWalk was designed with safeguards at the feature, data, and validation levels. Feature selection was guided by established siRNA biogenesis principles reported in the literature rather than data-driven feature discovery, ensuring biological interpretability and limiting model flexibility. Training data were derived from a large and heterogeneous collection of 29 Arabidopsis thaliana small RNA libraries spanning multiple studies and experimental conditions. Positive training examples were constructed using literature-supported phased contigs as references, while negative examples were generated through a

controlled computational procedure. Model performance was evaluated using both standard and stratified cross-validation schemes to assess generalization across data partitions.

Beyond classification performance, a central contribution of siWalk is its ability to rank biologically motivated features according to their predictive importance. This ranking provides quantitative insight into which biogenesis-related features contribute most strongly to siRNA production, offering biologists a data-driven perspective to refine or extend existing classification criteria. In this way, siWalk establishes a positive feedback loop between computational modeling and biological knowledge. In addition, siWalk uniquely enables prediction of siRNA localization within a given precursor, extending beyond binary locus classification to provide finer-grained, biologically actionable predictions.

The development and generalization of `siWalk` are constrained by the absence of a centralized, experimentally validated database for endogenous siRNAs—akin to miRBase for miRNAs. Without such a resource, it is challenging to define consistent ground truth across species or to benchmark performance beyond the Arabidopsis datasets used in this study. As a result, `siWalk` currently focuses on regularly phased siRNA loci in Arabidopsis, and its performance on irregular phasing patterns or in other plant systems remains untested. The model further depends on curated features rooted in current biogenesis knowledge, which may bias predictions toward well-characterized siRNA biotypes and overlook atypical or divergent forms. This reflects a broader challenge in the field, where limited experimental annotations hinder comprehensive model training. Nevertheless, a key strength of `siWalk` lies in its interpretability. By identifying and weighting biologically relevant features, the model can inform biologists about which attributes are most predictive of siRNA functionality. This creates a positive feedback loop: feature importance analysis may help refine siRNA classification criteria, guide future experimental designs, and support efforts toward standardizing siRNA annotation across species. The next chapter shifts from small RNA biomarker discovery to application, presenting a methodology for assessing genome integrity in RNA-based therapeutics delivered by recombinant adeno-associated viruses (rAAVs).

CHAPTER 7

ARCTICFOX: A QUANTITATIVE PROFILING FRAMEWORK FOR RAAV GENOME POPULATION

Small RNAs hold significant promise as therapeutic molecules. Viral delivery systems such as Adeno-Associated Virus (AAV) (Naso *et al.*, 2017; Wright, 2008) and *Potato virus X* (Zhao *et al.*, 2016) have been used for gene therapy and gene knockdown, enabling targeted delivery of small RNAs into both animal and plant cells. However, viral genome packaging is inherently error-prone, resulting in heterogeneous viral populations. In any given production batch, some viral particles correctly package the intended genome, while others accumulate mutations, deletions, or undesired insertions—leading to truncated, rearranged, or incomplete genomes. These defects complicate the functional validation of small RNAs and raise concerns about the reproducibility and safety of RNA-based interventions. Such inconsistencies introduce a critical need for systematic and quantitative quality control. Traditional short-read sequencing methods often fail to capture full-length AAV genomes, while long-read sequencing technologies, though more comprehensive, introduce their own biases that complicate downstream analysis (Tai *et al.*, 2018; Tran *et al.*, 2020; Wang *et al.*, 2019a).

To address the fourth specific objective of this thesis—establishing a computational framework to evaluate the integrity of RNA therapeutics delivered via viral vectors—this chapter introduces `ArcticFox`, a profiling system that leverages long-read sequencing (LRS) data to classify and quantify residual DNA species in recombinant AAV (rAAV) products. `ArcticFox` provides essential quality metrics for RNA cargo integrity and supports the development of safer and more consistent therapeutic manufacturing processes. To achieve this, the methodology includes:

- Develop a bioinformatics framework (`ArcticFox`) for assessing genome integrity in rAAV-based RNA delivery.
- Use long-read sequencing (PacBio, Oxford Nanopore) to detect truncated or rearranged viral genomes and correct for systematic sequencing biases.
- Establish standardized integrity metrics, including positivity rate, intact genome rate, and length distribution bias correction.

This work was conducted during an industrial research internship and led to two public conference contributions. The first was a poster presented in 2021 at the *Annual International Conference on Research in*

Computational Molecular Biology (*RECOMB*), titled *A Quantitative Profiling Framework for rAAV Genome Population*. The second, a more technical poster, was presented in 2023 at the *Intelligent Systems for Molecular Biology and the European Conference on Computational Biology* (*ISMB-ECCB*), titled *Assessing Functional Impurities in rAAV Production Platforms by Long-Read Sequencing*. Additional laboratory validations are currently underway. Once complete and cleared for release, these results will be incorporated into a comprehensive manuscript. Dr. Hui-Wen Liu identified the necessity of the methodology and, together with other scientists at the internship, provided guidance to Chao-Jung Wu (CJW) and produced standard samples used in the pipeline development. CJW implemented the analytical framework, conducted the experiments, and wrote the manuscript under the supervision of Professor Abdoulaye Baniré Diallo. Printed versions of the publicly available RECOMB abstract and ISMB poster are included in Appendix D.

Disclaimer: In accordance with the confidentiality agreement between UQAM and Biogen, only authorized content is included from the internship. Figures, tables, and their legends reflect previously published material. Additional contextual explanation is provided using publicly available information. The Introduction and Conclusion of this chapter are original, literature-based, and not governed by the confidentiality agreement.

7.1 Abstract

Recombinant adeno-associated virus (rAAV)-mediated gene therapy has been applied to treat human diseases. However, the rAAV capsids contain heterogeneous mixtures of full-length and truncated genomes and, depending on the manufacturing process, residual host cell and plasmid DNA. Therefore, methods are needed to characterize the encapsidated DNA of rAAV in order to support process development and batch release. The emerging long-read sequencing (LRS) has achieved AAV single-genome resolution. Here we propose a Python-based LRS profiling framework to classify and quantify residual DNA species in rAAV products. We designed a reference that contains universal genetic components that are commonly used in rAAV production, including AmpR, KanR, Rep and Cap genes along with HPV18, Ad5 and hg38 genomes. We assessed impurities of rAAV production from public and in-house LRS datasets. Analyzing the lambda fragments supplemented in these datasets showed that sequencing introduced size biases, which couldn't be fully corrected by regression but is improvable within library preparation. Functional potential of impurities were assessed through indicators derived from long-read alignments, which enabled us to quantitatively compare impurities between manufacturing batches. We demonstrated that LRS provides informative met-

rics for rAAV production and can facilitate process development to ensure therapeutic product safety and quality.

Keywords: rAAV genotyping, residual DNA, long-read sequencing, PacBio, Nanopore, gene therapy vector quality

7.2 Introduction

Adeno-associated virus (AAV), first discovered during adenovirus preparation in the 1960s, is a non-enveloped virus with a genome of single-stranded DNA (ssDNA) at 4.7 kb (Atchison *et al.*, 1965; Hoggan *et al.*, 1966). Since then, tremendous discovery has been made and now recombinant AAV (rAAV) has become one of the most promising tools for delivering therapeutic human genes, given its lack of pathogenicity, target specificity, and long-term efficacy in multiple clinical trials and products (Hermonat et Muzyczka, 1984; Naso *et al.*, 2017). The typical structure of a rAAV cargo vector contains the gene-of-interest (GOI) around 4 kb in between the 5'-ITR and the 3'-ITR (inverted terminal repeats). Current manufacturing platforms to generate rAAV for clinical use include transient expression in human HEK293 cell line with plasmids, one carrying cargo vector, one carrying the AAV Rep/Cap genes and one carrying helper genes from adeno or herpes viruses; adenovirus-infected producer HeLa cell line; and baculovirus-infected insect Sf9 producing system (Naso *et al.*, 2017). AAV accumulates in both media and cellular lysate from producer cells. An AAV production will contain not only cellular debris but also heterogeneous populations of rAAV particles that range from containing the intact cargo (full capsids) to those without the ITR-flanked transgene (empty capsids) and populations of various sizes of the genome in between. Full capsids can be efficiently separated from empty capsids using anion exchange chromatography after initial AAV purification by heparin or AAV-specific affinity columns during Good Manufacturing Practice (GMP)-compliant commercial manufacturing (Naso *et al.*, 2017; Potter *et al.*, 2014). However, one of the big challenges for rAAV manufacturing is the co-purification of residual DNA. AAV can package various amounts of host cell DNA, AAV genes, helper plasmid DNA and selection markers (e.g. KanR or AmpR) along with the cargo DNA. Previous studies have shown the presence of heterogeneous genomic populations composed of truncated transgenes and contaminating sequences (Chadeuf *et al.*, 2005; Tai *et al.*, 2018). These residual DNA are considered as impurities and more clinical experience is required to evaluate their adverse effects on product quality and safety (Wright, 2008). Studies have shown that by choosing and modifying manufacturing platforms, the impurity profiles change (Tran *et al.*, 2020, 2022). Therefore, methods to assess product purity and integrity are

essential for the optimization of manufacturing platforms. Current assessment methods for purity include electrophoresis such as alkaline agarose gel and capillary electrophoresis; genome identity can be verified by target-based approaches such as restriction fragment length polymorphism, quantitative polymerase chain reaction (qPCR), droplet digital PCR (ddPCR), and Sanger sequencing. These quality control methods (QC) are easy to implement and provide useful readouts at relatively low cost and short turn-around time. However, electrophoresis-based methods are limited by their low resolution. On the other hand, while target-based methods are sensitive, accurate, and can quantify vector genome copy, whether rearrangements or unknown contaminating sequences are present in an AAV production batch remains unclear.

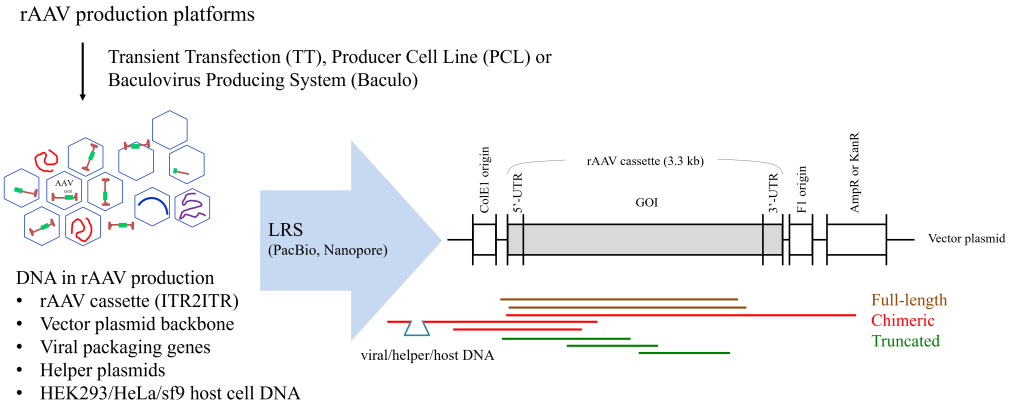


Figure 7.1: rAAV profiling by long-read sequencing. The rAAV capsids contain heterogeneous mixtures of full-length and truncated genomes and residual host cell and plasmid DNA, depending on the manufacturing process (Naso *et al.*, 2017). The emerging long-read sequencing (LRS) has AAV single-genome resolution (Tai *et al.*, 2018).

Next generation sequencing, including short-read sequencing (SRS) and the emerging long-read sequencing (LRS), is a powerful tool and could offer a solution (Oikonomopoulos *et al.*, 2020). While SRS gives valuable information regarding nucleotide variants and the extent of contaminating sequences, it poses limitations for assessment of full-length cargo because fragmentation of DNA input is required prior to sequencing library construction (Lecomte *et al.*, 2015). All next-generation sequencing platforms experience some degree of base-calling errors. However, advancements in library preparation methodologies have significantly

reduced the error rates associated with long-read sequencing over the years (Oikonomopoulos *et al.*, 2020). For genotyping purposes, error rates below 10% may be acceptable. In addition, GC bias of SRS results in low coverage of sequence alignment in GC-rich sequences and is known to aggravate genome assembly (Chen *et al.*, 2013). Sequencing through the ITR regions is also challenging due to its strong secondary structure (Tran *et al.*, 2020).

Recently, studies based on LRS revealed the presence of truncated genomes and contaminating sequences fused with transgenes in the defective cargo (Tai *et al.*, 2018). The source of residual DNA is from packaging components such as AmpR, KanR, Rep and Cap genes along with HPV18, Ad5 and human genomes (Lecomte *et al.*, 2015; Tai *et al.*, 2018; Tran *et al.*, 2020), depending on the manufacturing process (Dobrowsky *et al.*, 2021). These unintended sequences raise concerns regarding potential long-term expression in patient tissues, warranting close monitoring and control. Long-read sequencing has helped clarify the heterogeneity and complexity of residual DNA in rAAV genomes. To quantify this residual DNA, we developed a Python-based software for characterizing these sequences in purified rAAV products using a predefined reference residual DNA genome. ArcticFox accepts FASTQ data from long-read platforms, providing quantitative insights into impurities for cross-batch comparison using metrics derived from sequence alignments.

7.3 Materials and Methods

7.3.1 ArcticFox overview

ArcticFox includes modules specifically designed for the quantitative profiling of rAAV genome populations, as illustrated in Figure 7.2. It is compatible with sequencing data generated from major long-read platforms, such as PacBio and Oxford Nanopore. The LRS pipeline for assessing functional impurities in rAAV comprises three analytical modules integrated into the Python-based software ArcticFox: Dataset Validation, Genotyping and HG Analysis. ArcticFox is available as an open source project in GitHub at https://github.com/bioinfoUQAM/ArcticFox

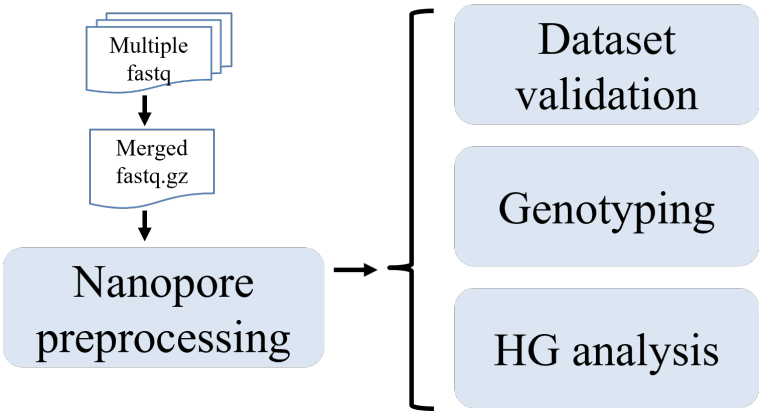


Figure 7.2: LRS pipeline for assessing functional impurities in rAAV, including three analytic modules embedded in the python-based software ArcticFox: Dataset validation to validate usable dataset, analyze lambda fragments and report picking up by template length and reading through by template length. Genotyping to analyze residual DNA based on several pre-defined chromosomes, reporting output metrics. HG analysis to analyze residual DNA coming from host genome hg38, reporting output metrics.

7.3.2 Datasets

To evaluate the consistency and utility of long-read sequencing for profiling rAAV genome populations, 18 LRS datasets from 4 rAAV production experiments were collected (Table 7.1). These datasets represent various rAAV production systems and sequencing data from PacBio and Nanopore platforms. In-house batches, Sample01 and Sample05, were prepared using different AAV production methods. Lambda fragments were added to the encapsidated AAV genomic DNA, followed by library preparation for Nanopore and PacBio sequencing. Selected public PacBio datasets were retrieved from the NCBI Sequence Read Archive. Lambda DNA readouts analyzed template length bias in sequencing procedures, and the experiments demonstrated the ability to compare genetic components using ArcticFox's default reference-genome settings.

7.3.3 Design of reference genomes

The reference genomes supporting the analytic modules are illustrated in Figure 7.3. The first genome is for lambda calibration, containing 14 BstEII-digest fragments derived from the lambda genome. The sec-

Table 7.1: LRS datasets of rAAV productions from public and in-house sources, including (a) in-house, (b) SRA PRJNA383145 (Tai *et al.*, 2018), (c) SRA PRJNA608034 (Tran *et al.*, 2020), and (d) SRA PRJNA794197 (Tran *et al.*, 2022), representing various rAAV production systems; and sequencing data obtained from PacBio and Nanopore long-read sequencers.

Sample name	Transgene	Sequencer	Dataset source
Nanopore.Sample01	BIIB-X	MinION Mk1B	a
Nanopore.Sample05	BIIB-X	MinION Mk1B	a
PacBio.Sample01	BIIB-X	PACBIO Sequel II	a
PacBio.Sample05	BIIB-X	PACBIO Sequel II	a
pt2018_78	EGFP	PacBio RS II	b
pt2018_79	EGFP	PacBio RS II	b
pt2018_80	EGFP	PacBio RS II	b
pt2018_81	EGFP	PacBio RS II	b
pt2020_10 (Lib1)	Cas9	PACBIO Sequel	c
pt2020_09 (Lib2)	Cas9	PACBIO Sequel	c
pt2020_08 (Lib3)	Cas9	PACBIO Sequel	c
pt2020_07 (Lib4)	Cas9	PACBIO Sequel	c
pt2020_06 (Lib5)	mCherry	PACBIO Sequel	c
pt2020_05 (Lib6)	mCherry	PACBIO Sequel	c
pt2020_04 (Lib7)	EGFP	PACBIO Sequel	c
pt2020_03 (Lib8)	EGFP	PACBIO Sequel	c
pt2022_03	EGFP	PACBIO Sequel II	d
pt2022_04	EGFP	PACBIO Sequel II	d

ond genome is for genotyping, containing 9 chromosomes: ITR2ITR, AmpR, KanR, Rep, Cap, HPV18, Ad5, lambda integral genome and backbone. The third genome is the human genome, containing 22 autosomal chromosomes, 2 sex chromosomes and 1 mitochondrial genome. The default transgene in ITR2ITR is Cas9 derived from pX602 (Addgene), or a user-specified sequence. The default backbone is pX602 with the transgene removed, or a user-specified backbone. Lambda DNA-BstEII digest (New England Biolabs) is often supplemented as a sequencing control during library preparation.

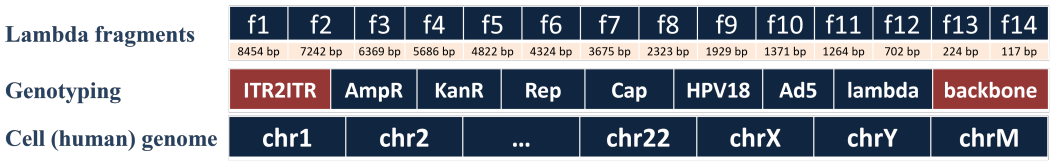


Figure 7.3: Design of reference genomes. The transgene in ITR2ITR is default Cas9 derived from pX602 (Addgene) or specific. The backbone is default pX602 removing transgene or specific. Lambda DNA-BstEII digest (New England Biolabs) is often supplemented as sequencing control during library preparation.

7.3.4 Correcting size biases of read abundance

We suggest that the size-biased read abundance could be modeled by regression as follows. Given A_o: Observed read abundance, L_t: Theoretical template length, L_o: Observed read length and R_o: Read abundance ratio (against baseline f6) obtained by analyzing lambda fragments, the corrected read abundance A_c in formula 3 can be obtained by determining formula 1 and 2, where formula 1 models the size bias regarding picking up template based on template length trained on lambda; and formula 2 models the size bias regarding complete reading through template based on template length trained on lambda. The model of formula 1 is a polynomial regression, and that of formula 2 is a linear regression. Formulas:

$$R_o = f(L_t) = \beta_0 + \beta_1 L_t + \beta_2 L_t^2 + ... \tag{7.1}$$

$$L_t = g(L_o) = \gamma_0 + \gamma_1 L_o \tag{7.2}$$

$$A_c = h(L_o) = A_o / R_o = A_o / f(L_t) = A_o / f(g(L_o)) \tag{7.3}$$

7.3.5 Functional quantification

Positivity rate. The positivity rate for each chromosome is defined as the percentage of reads in a batch aligned to that chromosome:

$$P_j = \frac{R_j}{R_{\text{total}}} \times 100\% \tag{7.4}$$

where P_j represents the positivity rate for chromosome j, R_j is the number of reads containing chromosome j, and R_{total} is the total number of reads in the library, excluding lambda reads, where j = {ITR2ITR, AmpR, KanR, Rep, Cap, HPV18, Ad5, backbone}. Take ITR2ITR as an example,
Positivity rate = reads containing ITR2ITR / total reads in library.

Intact rate. The intact rate for each chromosome is the percentage of reads that are quasi-full-length compared to all reads containing that chromosome:

$$I_j = \frac{Q_j}{R_j} \times 100\% \tag{7.5}$$

where I_j represents the intact rate for chromosome j, Q_j is the number of quasi-full-length reads of chromosome j, and R_j is the total number of reads containing chromosome j. Take ITR2ITR as an example,
Intact rate = reads are full length ITR2ITR / reads containing ITR2ITR; full-length threshold is 90%.

Mass percentage. Mass percentage for each chromosome is inferred by the proportion of nucleotides aligned to the chromosome relative to the total nucleotide count in the batch:

$$M_j = \frac{N_j}{N_{\text{total}}} \times 100\% \tag{7.6}$$

where M_j represents the mass percentage of chromosome j, N_j is the nucleotide count aligned to chromosome j, and N_{total} is the total nucleotide count in the library, excluding lambda nucleotides. Take ITR2ITR as an example, Mass percentage = nucleotides aligned to ITR2ITR / total nucleotides in library

7.4 Results

7.4.1 Size bias related to sequencing

Using known-length DNA fragments for spike-in analysis calibrates sequencer bias related to template length (Shirane *et al.*, 2013; Castro-Wallace *et al.*, 2017). According to the Nanopore manufacturer's guidelines, the spike-in procedure involves BstEII digestion of the lambda genome, producing 14 fragments (f1 to f14) of theoretically equal abundance. During sample preparation, lambda DNA fragments were added to each rAAV

sample prior to sequencing. After sequencing, read abundance for each fragment was measured, and ratios were calculated against the 4324 bp fragment (Figure 7.4A). If sequencing is unbiased for picking up templates based on template lengths, these ratios should be close to 1. The completeness of reading for each DNA template is assessed by the full-length rate as the reading through rate, which should be 100% if there is no length bias (Figure 7.4B).

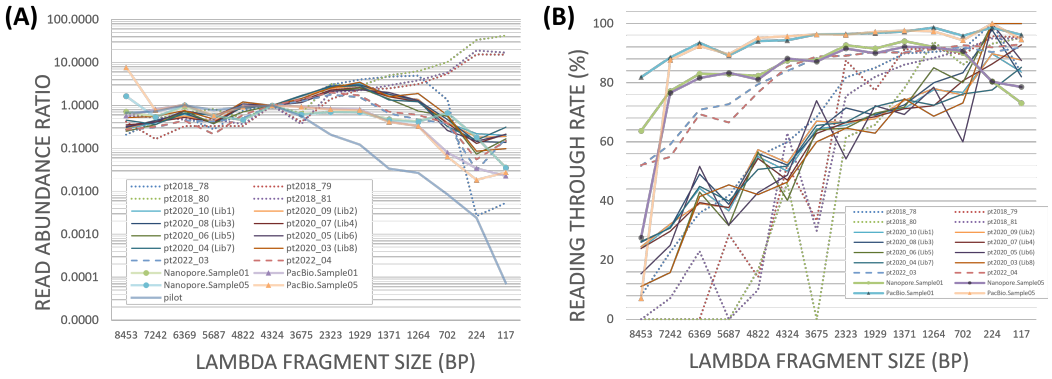


Figure 7.4: Size bias related to sequencing (A) regarding picking up template based on template length and (B) regarding complete reading through template based on template length. The BstEII-digestion of lambda genome gives rise to 14 fragments representing a range of template sizes that are theoretically of equal abundance ratio. After sequence identification, read abundance of each lambda fragment was measured and the abundance ratios were calculated against the fragment of 4324 bp. The theoretical reading through rate is 100%.

Picking up bias can be corrected if the ratios show a steady trend without fluctuations (Figure 7.5A2), otherwise negative ratio and overfitting occur in the polynomial regression model (Figure 7.5A1). For the regressions of reading through shown in Figure 7.5B, the values of the slope γ_1 and the R^2 are close to 1, which indicates that the major trend is L_t equals to $L_o + \gamma_0$. However, reading through is challenging, as the square root of the mean squared error can reach several hundred base pairs, particularly when the theoretical length of observed small reads spans a wide range. Therefore, regressions have limited power to mathematically rescue the size biases and are only applicable for rescuing a steady trend of picking up bias. In-house and public datasets show that size biases are similar within experiments and improve over time (Figure 7.4), suggesting that library preparation protocols may reduce these biases.

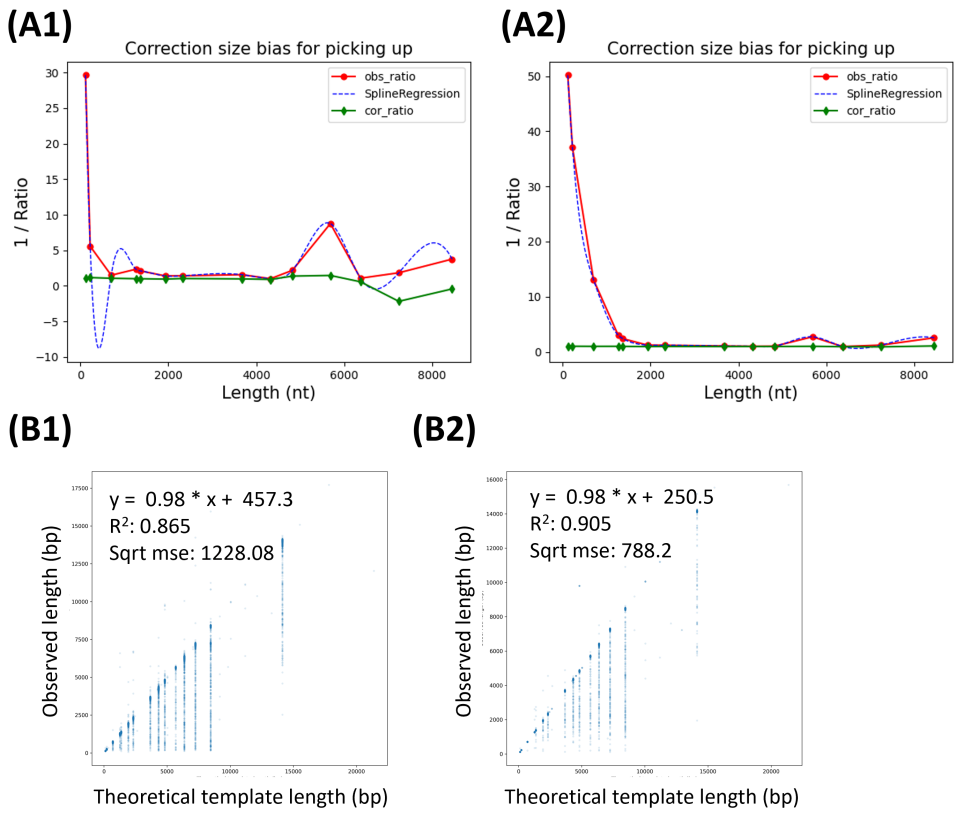


Figure 7.5: Correcting size biases of read abundance estimated from training on lambda in sample01 regarding (A) picking up and (B) reading through in (1) Nanopore and (2) PacBio. Picking up bias can be corrected if the ratios show a steady trend without fluctuations, otherwise negative ratio and overfitting occur in the polynomial regression model. Reading through is very difficult to correct, especially that the theoretical lengths of an observed small read have a wide range of possible values. Sqrt mse indicates on average how far (bp) the prediction of the linear regression model is from the theoretical length.

7.4.2 Multiple manufacturing batches

The changes of residual DNA profiles between batches and manufacturing processes can be captured by comparing the ArcticFox output metrics for the common genetic components (Figure 7.6). Users can then select the most effective manufacturing process based on this comparative analysis, ensuring optimal quality and consistency in their rAAV products.

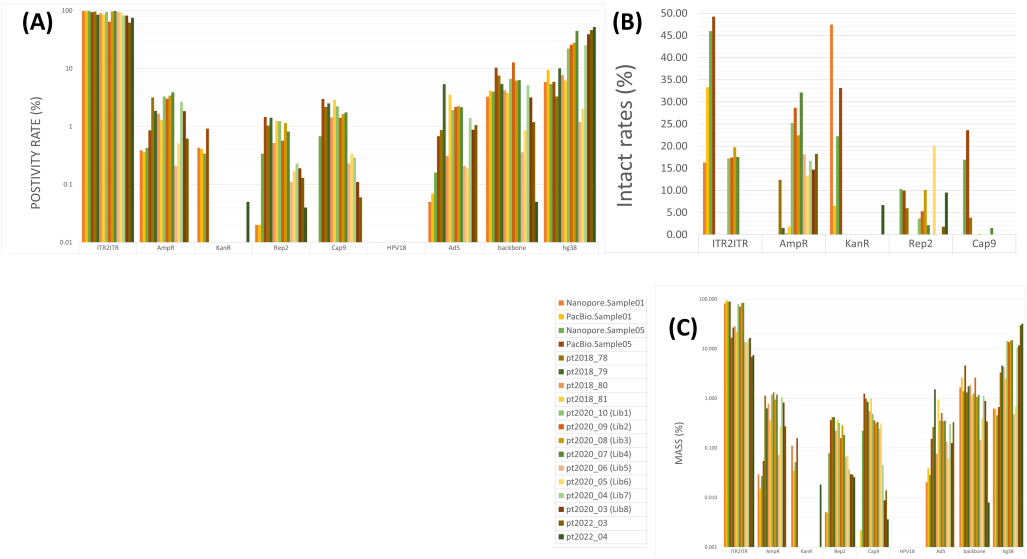


Figure 7.6: LRS sequencing provides several output metrics to evaluate functional potential of impurities, such as (A) Percentages of DNA populations based on reference genome positivity rates, (B) Intact rates of reference genomes, and (C) Mass percentages of reference genomes inferred from base count. These impurity metrics of the universal genetic components - AmpR, KanR, Rep and Cap genes and host genomes – allowed for quantitative comparison of impurities from different manufacturing batches; while the positivity rate can be compared if the ITR2ITR transgene between batches are the same.

7.5 Conclusion

LRS provides an opportunity to quick profile the drug substance before delivery or assess the outcome of a production procedure change or a vector design change. The universal reference genomes capture the genetic components that are commonly used in rAAV production. In addition to mass percentages of reference genomes, LRS provides the positivity rates of DNA populations and intact rates of reference genomes, which could not be obtained via short-read sequencing (SRS).

LRS is prone to size bias. Supplement of Lambda could be used to evaluate the quality of the library preparation and validate a usable dataset. The biases are similar within an experiment but different between experiments, indicating that the bias comes from library preparation rather than the sequencers. The size biases could be modeled by regression. However, overfitting may happen due to insufficient datapoints (e.g., 14 datapoints from standard fragments); and it is difficult to predict the theoretical lengths of an ob-

served small read for modeling reading through. The correction power of regression is therefore limited. Improving library preparation is more intuitive than correcting the biases mathematically.

7.6 Implications

Small RNAs hold great promise as biomarkers in both research and therapeutic contexts. Earlier chapters addressed their classification and characterization; this chapter extends that vision to quality control, where RNA delivery must be validated at nucleotide resolution. `ArcticFox` introduces a profiling methodology that quantifies genome integrity in rAAV-delivered RNA therapeutics using long-read sequencing. The advances of this work include:

- A bioinformatics pipeline for enhanced rAAV genome integrity analysis using long-read sequencing, reporting standardized integrity metrics to support industrial quality control.
- Detection and correction of systematic biases in viral vector sequencing workflows.
- Demonstration of the practical utility of long-read sequencing for genome-level quality control in therapeutic manufacturing.

This work fulfills the fourth specific objective by establishing a computational framework to assess the integrity of RNA therapeutics delivered via viral vectors. It addresses a key translational challenge—ensuring that RNA delivery vectors are free from artifacts—bridging computational analysis and experimental validation. This reinforces the thesis's central hypothesis: that small RNA biomarker discovery and application benefit from biologically informed, robust computational tools. While `ArcticFox` offers a powerful approach to profiling rAAV-based therapeutics with LRS, several considerations remain. The current implementation is optimized for rAAV, but its modular design allows adaptation to other viral vectors, including plant virus-based delivery systems. Its accuracy depends on input library quality, which can be affected by sample preparation biases. Although `ArcticFox` incorporates spike-in controls and correction models, some artifacts persist, potentially distorting genome representation.

Although prior studies (e.g., using PacBio) revealed heterogeneity in rAAV genomes, NGS-based genome-level quality control in therapeutic manufacturing remains underexplored. This project is among the first to apply LRS to industrial datasets and to introduce standardized integrity metrics tailored for rAAV evaluation. These contributions demonstrate the practical value of LRS and the need for dedicated bioinformatics solutions in quality assurance pipelines for RNA therapeutics.

CONCLUSION

This thesis set out to enhance the discovery and application of small RNA biomarkers by developing computational methodologies that are biologically informed, scalable, and robust. Across four interconnected methodologies, we addressed critical gaps in the classification, analysis, and delivery of small RNAs by integrating knowledge from plant biology, genomics, computer science, and machine learning. Our work exemplifies the importance of interdisciplinary approaches in bioinformatics, particularly in the context of high-throughput sequencing and functional RNA research.

The **first** main contribution of this thesis is the development of `mirLibSpark`, a distributed miRNA prediction pipeline introduced in Chapter 4. Built on Apache Spark, this methodology scales efficiently to multi-library datasets and large plant genomes. It addresses the bottleneck of miRNA analysis in data-intensive studies by integrating prediction, validation, and downstream annotation into an automated and reproducible workflow. The **second** main contribution, presented in Chapter 5, extends the analytical scope to study the joint behavior of miRNAs and siRNAs during cold acclimation in wheat, offering new insights into regulatory coordination and evolutionary conservation. This methodology not only demonstrates the biological value of unified *de novo* small RNA analysis but also identifies computational limitations in existing siRNA detection methods, motivating the development of more accurate and interpretable alternatives. The **third** main contribution, described in Chapter 6, is the introduction of `siWalk`, a machine learning model trained on features tied to siRNA biogenesis. By moving beyond heuristic phase scoring, `siWalk` improves classification accuracy and interpretability, laying the groundwork for more rigorous siRNA annotation practices. The **fourth** main contribution, detailed in Chapter 7, is the development of `ArcticFox`, a quality control framework for assessing the genome integrity of RNA therapeutics delivered via recombinant adeno-associated viruses. This work represents one of the first applications of long-read sequencing to profile viral therapeutic payloads at high resolution, and the first to establish quantitative metrics for evaluating residual genomes in industrial RNA therapeutic production, thereby linking computational analysis directly to quality control and deployment.

Together, these methodologies illustrate a cohesive strategy for advancing RNA biomarker research. They emphasize the importance of computational scalability, biological interpretability, and practical utility, offering new ways to explore and understand RNA-based regulation. The work flows from predictive modeling to real experimental datasets, demonstrating that scalable and biologically informed methods can support the

full biomarker lifecycle, from classification and validation to therapeutic development. By integrating mechanistic insight with robust computational design, this thesis highlights how interdisciplinary approaches can accelerate both the understanding and application of RNA-based technologies.

Despite the broad methodological coverage presented in this thesis, certain limitations and scope boundaries remain. While the proposed frameworks collectively span much of the RNA biomarker discovery process—from high-throughput prediction to therapeutic quality control—they do not encompass all aspects of the process. Although some datasets were generated in-house or in collaboration, this thesis does not examine how experimental design choices—such as tissue selection, time points, or sequencing depth—might influence the quality or interpretability of small RNA analysis. Experimental validation also remains a structural bottleneck: while predictions were supported by literature, conservation, or enrichment analyses, wet-lab confirmation of small RNA function or target interactions was beyond the scope of this work. In addition, functional annotation was computational and did not involve biochemical validation. Finally, while the thesis focuses deeply on miRNAs and siRNAs, other RNA classes such as tRNA fragments, and lncRNA-derived sRNAs remain unexplored. These boundaries reflect intentional scope decisions aimed at maximizing methodological depth and coherence, while identifying clear opportunities for future expansion.

In addition to these scope boundaries, this thesis operates within several structural and methodological constraints that reflect the current state of small RNA research. First, all computational models and pipelines rely on existing reference genomes, annotation databases, and community-defined criteria, which continue to evolve as sequencing technologies and biological standards advance. Second, the scarcity of experimentally validated siRNA ground truth necessitated the use of inferred labels and curated proxies, introducing an unavoidable dependency on available annotations. Third, scalability was achieved through abstraction and distributed computation. While this improves efficiency, it may require environment-specific tuning to ensure portability across infrastructures. Finally, the biological scope of the work is intentionally centered on plant systems, where miRNA and siRNA biogenesis follow mechanistic principles distinct from those in animals. These limitations reflect the current technical and biological boundaries of small RNA research.

Despite these constraints, the frameworks developed in this thesis establish a foundation for systematic extension. Within plant systems, the proposed computational architectures can be generalized to additional species, environmental conditions, and larger population-scale datasets. As experimentally validated reference databases expand, particularly for siRNAs, predictive models can be recalibrated to improve ro-

bustness and reduce annotation bias. The modular design of the pipelines further enables integration of additional interfering RNA classes and emerging molecular origins without restructuring the core architecture. Beyond discovery, the quantitative strategies introduced here support in silico evaluation and optimization of siRNA-producing constructs and viral delivery vectors, thereby strengthening the computational bridge between biomarker identification and therapeutic implementation.

Looking ahead, this work opens three promising directions for advancing small RNA research. First, the unified framework for miRNA and siRNA analysis can be extended to cover additional RNA subclasses. For relatively well-characterized types such as tRNA fragments and snoRNA-derived sRNAs, it may be more strategic to integrate outputs from established tools and curated databases rather than developing new models from scratch. In contrast, under-characterized siRNA subclasses, including nat-siRNAs and hc-siRNAs, warrant the development of new tools and a deeper analysis, as their annotation and biogenesis remain poorly understood. Second, to support generalization beyond *Arabidopsis*, high-confidence ground truth datasets should be curated by combining phasing score–based detection with false discovery rate filtering and database cross-referencing. These resources would facilitate model transfer to other plant species while preserving interpretability and reliability. Third, future prediction tools should aim to classify siRNA precursors based on their biogenesis pathways, distinguishing Type I, II, and III origins—respectively associated with RNA-dependent RNA polymerase activity, intramolecular foldbacks, and transcript hybridization. This precursor-based classification would enhance the annotation of diverse siRNA loci and clarify regulatory mechanisms across varied genomic contexts.

In conclusion, this thesis delivers unified, scalable, and biologically informed computational methodologies for the functional analysis of small RNAs, establishing a coherent framework that connects discovery, prediction, and therapeutic evaluation.

Availability of Methodologies

Most computational tools and resources developed in this thesis are available to support transparency, reuse, and reproducibility. The source code for Chapter 4 (`mirLibSpark`) is publicly accessible at `https://github.com/bioinfoUQAM/mirlibspark` under the BSD 3-Clause License, and includes simulated positive and negative miRNA datasets.

The repositories for Chapter 6 (`siWalk`) and Chapter 7 (`ArcticFox`) are currently private but will be made

publicly available upon formal publication of their corresponding manuscripts. `siWalk`, hosted at `https://github.com/bioinfoUQAM/siWalk`, will be released under the MIT License and will include the full source code, curated precursor datasets, engineered feature files, and the trained machine learning model. `ArcticFox`, available at `https://github.com/bioinfoUQAM/arcticfox`, will be released under AGPLv3 License (GNU Affero General Public License Version 3) to support use in both academic and industrial environments. Each repository provides documentation and usage examples to support community adoption.

While no standalone package was released for the wheat iRNA study (Chapter 5), the analysis reuses modules from `mirLibSpark` and includes a Spark-parallelized prototype for rule-based siRNA classification. Due to the study's collaborative scope and reliance on third-party tools—such as psRNATarget and other external services—the workflow requires manual coordination between software components, making it unsuitable for distribution as a fully automated pipeline. Nonetheless, the full analysis framework is transparently documented in the manuscript, with detailed workflows and step-by-step procedures thoroughly described in the supplementary materials and figure legends, enabling reproducibility and independent reuse.

Funding This research was supported by a CGSD Doctoral Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and a B2X Doctoral Scholarship (DOI 10.69777/255823) from the Fonds de recherche du Québec – Nature et technologies (FRQNT).

The Centre de recherches mathématiques (CRM) nominated the author for an FRQNT International Internship award, which was granted but declined by the author due to pandemic-related travel restrictions. Additional support for conference participation was provided by the International Society for Computational Biology (ISCB) through a travel fellowship, in partnership with the University of Colorado Department of Biomedical Informatics. All fellowships and awards listed above were granted directly to the author in support of this doctoral research and do not include institutional or supervisor-held research funding.

APPENDIX A

MIRLIBSPARK

MirLibSpark: A Scalable NGS Plant MicroRNA Prediction Pipeline for Multi-Library Functional Annotation

Chao-Jung Wu
Amine M. Remita
Abdoulaye Baniré Diallo
wu.chao_jung@courrier.uqam.ca
remita.amine@courrier.uqam.ca
diallo.abdoulaye@uqam.ca
Department of Computer Science, Université du Québec à Montréal
Montréal, Québec

ABSTRACT

The emergence of the Next Generation Sequencing increases drastically the volume of transcriptomic data. Although many standalone algorithms and workflows for novel microRNA (miRNA) prediction have been proposed, few are designed for processing large volume of sequence data from large genomes, and even fewer further annotate functional miRNAs by analyzing multiple libraries. We propose an improved pipeline for a high volume data facility by implementing mirLibSpark based on the Apache Spark framework. This pipeline is the fastest actual method, and provides an accuracy improvement compared to the standard. In this paper, we deliver the first distributed functional miRNA predictor as a standalone and fully automated package. It is an efficient and accurate miRNA predictor with functional insight. Furthermore, it compiles with the gold-standard requirement on plant miRNA predictions.

CCS CONCEPTS

• **Computing methodologies** → **MapReduce algorithms**; • **Applied computing** → **Computational transcriptomics**; **Molecular sequence analysis**; **Bioinformatics**; *Agriculture*; • **Information systems** → *Clustering and classification.*

KEYWORDS

Distributed computing, Parallelization, Apache Spark, MicroRNA classification, Functional transcriptomic annotation

ACM Reference Format:
Chao-Jung Wu, Amine M. Remita, and Abdoulaye Baniré Diallo. 2019. MirLibSpark: A Scalable NGS Plant MicroRNA Prediction Pipeline for Multi-Library Functional Annotation. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3307339.3343463

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3343463

1 INTRODUCTION

The emergence of the Next Generation Sequencing (NGS) allows researchers to quickly and cheaply collect billions of expressed small RNAs (sRNAs) sequences according to a given condition so that the volume of small-RNAome data increases dramatically. However, the annotation of sRNAs has never been rapidly and systematically achieved. Parallelization approach is becmoing useful for computation-intensive tasks in biology [6]. In the era of big data, CPU speeds and storage capacity are greatly improving. Hard disk input/output (I/O) is also dealt with by Hadoop MapReduce. Furthermore, Resilient Distributed Dataset (RDD) abstraction is performed to imporve Hadoop's disk-based computing [29]. Scalable MapReduce methods are revolutionizing traditional bioinformatics tools in RNA research [9, 21, 24].

Nowadays, several well-known bioinformatics tools are dedicated to annotating sRNAs as miRNAs in plants. MiRDeep-P [28] employs a probabilistic scoring algorithm to predict novel miRNAs from precursor signatures and annotates directly known miRNA candidates as miRNAs without considering their dynamic expressions within the transcriptome. Currently, the best prediction is performed by ShortStack [3, 4], and the fastest prediction tool is miR-PREFeR [18]. Both ShortStack and miR-PREFeR incorporate miRNA annotation criteria [22] in their annotation algorithms and decrease false positive predictions so that they annotate miRNAs with low false positives, high precision and high accuracy [4]. Although ShortStack and miR-PREFeR take multiple pre-aligned sRNA-seq libraries as inputs, ShortStack is not scalable and has difficulty to deal with species having larger genomes such as maize that has 2 GB genome [18]. MiR-PREFeR spawns subprocesses to execute a function on distributed input data across multiple processors on a given machine. This approach improved the processing speed for maize sRNA-seq libraries [18]. However, the performance of any parallelization approach has not been assessed in species that have large genomes, such as wheat of 13 GB genome.

In our previous work [1], we established a standardized integrative procedure to annotate miRNAs. The procedure includes four major steps: miRNA prediction, target gene prediction, differential expression analysis and functional enrichment. This data-intensive annotation procedure reduces false positive predictions by taking into account experts criteria, and renders meaningful miRNA annotations due to the incorporation of functional analysis. However, none of the existing tools has worked towards scaling up the entire

procedure. To the best of our knowledge, there is only one miRNA predictor employing MapReduce [25] to perform a few steps of the procedure. Additionally, MR-microT server [15] uses MapReduce to annotate the downstream target genes of miRNAs.

In this work, we propose an improved miRNA prediction procedure that adds the miRNA annotation criteria for big data [4], incorporates functional and structural annotations and accelerates the processing time. We present mirLibSpark, a Spark-based miRNA annotation pipeline that executes the entire miRNA annotation procedure and is able to process multiple sRNA-seq libraries with high prediction performance and speed in both personal computer and high performance computing clusters.

2 MIRLIBSPARK ARCHITECTURE AND EXECUTION

2.1 MirLibSpark overview

Annotation modules. MirLibSpark contains modules dedicated to integrative procedure of functional miRNA annotation, as shown in Figure 1. It includes a module (M1) providing the reference genomes, Bowtie indices [17] and KEGG pathway functional annotations [14]. Its core module is the miRNA prediction module (M2). Additional modules are attached to help users to better annotate miRNAs (M3: target gene prediction, M6: KEGG pathway annotation of the target genes and M5: precursor visualization using VARNA [8]) and to aggregate multi-library data (M4: differential expression analysis and M7: functional enrichment analysis).

MiRNA prediction module (M2). The mirLibSpark pipeline implements the four major computational steps for miRNA prediction: filtering, alignment, folding and prediction. For each run, mirLibSpark pipeline is initiated with a tunable configuration, then 30 different operations are applied on the multiple RDD created objects to make resilient and distributed data. These operations apply rigorously the miRNA annotation criteria and procedures for plants defined in [4, 22]. The parameters were already optimized and were designed to be adjustable in the pipeline, because the experts' criteria may evolve in the future.

Additional functional annotation modules (M3-M7). Post-prediction processing among multiple libraries is performed in one run. We compute the statistics of differential expression to select functionally relevant miRNA candidates [5, 13]. Their target genes are subsequently analyzed for pathway enrichment [14, 27].

Implementation. Based on the computing load, computation steps within miRNA prediction module (shown in Figure 2), precursor visualization module and target gene prediction module are chained through Spark RDD operations [29]. In each step, an RDD object is generated to distribute the instances and operations across a cluster of computers. Each instance corresponds to a sRNA sequence and its related data. We used *map*() and *filter*() functions when it is required, where *map*() changes the content of an instance and *filter*() eliminates an instance based on a boolean statement.

We opt for the implementation of adapted functions to the Spark architecture for data and computing distribution instead of using extensively external programs. These programs should require less dependency and be currently maintained and open source to facilitate the pipeline redistribution. All the external programs are either distributed with the project or are automatically installed by mirLibSpark. Therefore, installation procedures for these programs are not required by the user.

Availability. MirLibSpark is available as an open source project in GitHub and as a docker image, along with its manual, in the following link https://github.com/bioinfoUQAM/mirLibSpark. We deliver mirLibSpark as a standalone package for two environments, (1) a personal computer or (2) a remote server with high computational power. For biologists, we recommend to use the docker image of mirLibSpark in their personal computers. For bioinformaticians, the GitHub project can be deployed in any server. Upon installation of mirLibSpark, it is ready to execute experiments for a selected genome in a personal computer or with minimal configuration in a high performance computing server.

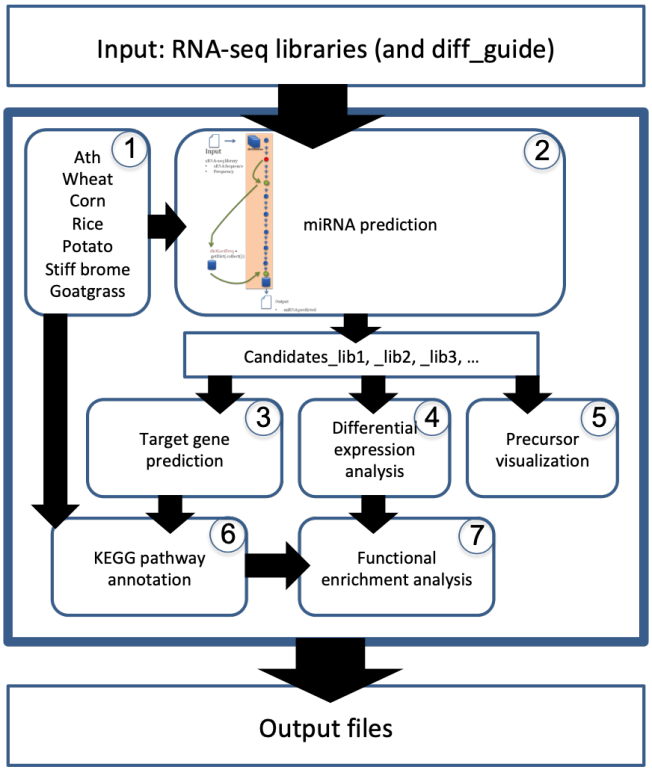


Figure 1: MirLibSpark overview. M1: genomic reference annotations. M2: miRNA prediction. M3: target gene prediction. M4: differential expression analysis. M5: precursor visualization. M6: KEGG pathway annotation. M7: functional enrichment analysis. M for module.

2.2 MiRNA prediction module (M2)

A simplified illustration of the RDD chaining in the prediction module is shown in Figure 2. A *SparkContext*() is initiated with adjustable parameters of Spark and mirLibSpark. Default values, which are optimized for plants, are used if not otherwise specified by users. In each submission request, the pipeline can handle one to several sRNA-seq libraries. The first RDD object is created by

reading the data of each input library. Thereafter, operations are applied on RDDs until the output is obtained and printed to a file. Several filters, including low complexity filter using dustmasker [23] and various miRNA and precursor length and abundance criteria [4, 22], are achieved by the $filter()$ operation. Only aligned sequences using Bowtie within a small number of genomic loci are retained. The RDD generated by alignment is reused as reference for the filter step of the miRNA candidate expression within precursor range. After mapping, only the sRNAs with length between 21 and 24 nucleotides (nt) are retained [4]. A set of known non-miRNA sequences according to the annotation of the genome are removed. Pre-miRNA candidates surrounding the sRNA genomic loci are extracted and their secondary structures are computed using RNAfold [20]. The RNA structures and the decisions of miRCheck [12, 22] are rendered by *map*() functions with specified rules as parameters. The precursors longer than 300 nt or containing a second loop that is larger than 5 nt are discarded [4]. Finally, a dominantly expressed potential miRNA within the pre-miRNA candidate sequence is considered a true miRNA and is subjected to subsequent functional analysis. Most of the parameters are easily adjustable as needed. Since the parameters are optimized for plants and can affect the efficiency and the accuracy of the pipeline, users do not need to change them unless suggested by the experts.

Several incorporated programs, such as Bowtie, already support parallelization, and indeed are not a time-consuming step. Therefore, the acceleration by our implementation is to parallelize and chain every single prediction step in one run. For example, RNAfold is parallelized within mirLibSpark, and giving mirCheck the resulting structure as input is also taken care of by mirLibSpark.

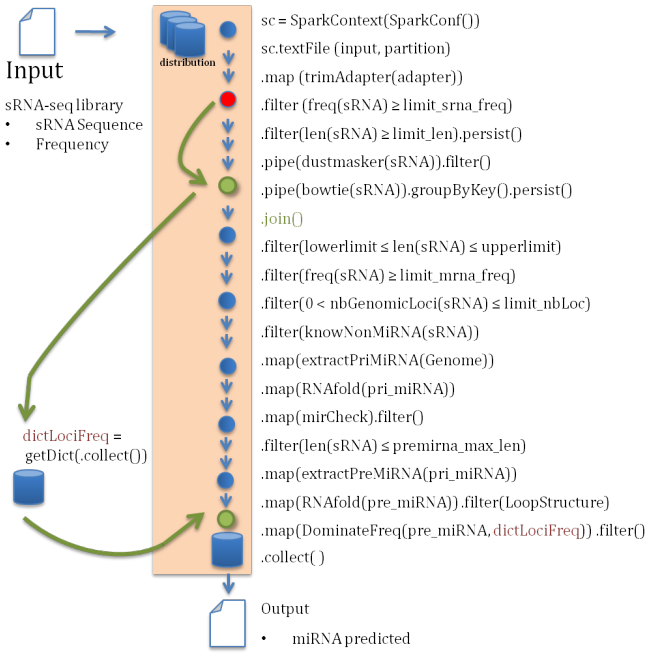


Figure 2: MirLibSpark miRNA prediction module.

2.3 Additional annotation modules

Genomic annotation (M1). To enhance user's experience, mirLibSpark already supports seven plant species of small and large genomes: *Arabidopsis thaliana* (ath), *Triticum aestivum* (wheat), *Zea mays* (corn), *Oryza sativa* (rice), *Solanum tuberosum* (potato), *Brachypodium distachyon* (stiff brome) and *Aegilops tauschii* (Tausch's goatgrass). The project includes a script for user to fetch the genomes and annotations from Ensembl plants (http://plants.ensembl.org/) [7] and to build automatically Bowtie indices for supported plant species. The genome-wide annotations of KEGG pathways are retrieved from KEGG database [14].

Target gene prediction (M3). The best 100 target genes are predicted by miRanda [11], parallelized in mirLibSpark. Gene variants are consolidated to a single identifier. Then the best 5 scored identifiers are subjected for functional analysis.

Differential expression analysis (M4). According to the guide file, comparisons are performed between libraries. Expression fold change for each miRNA candidate is calculated. Then Kal's test [13] is performed to examine the difference between samples by calculating a z-score and a p-value for each miRNA. The p-value is corrected by false discovery rate [5]. Only the miRNAs with fold changes greater than 2 or less than 0.5, and corrected p-values smaller than 0.05 are considered differentially expressed.

KEGG pathway annotation (M6) and functional enrichment analysis (M7). Target genes of differentially expressed miRNAs in the libraries are annotated with KEGG pathways by KEGG pathway annotation module (M6). Pathway enrichment is performed by hypergeometric analysis, similar to other miRNA functional analysis software mentioned in [27]. The chance of all the drawing combinations of the genes to obtain the sample frequency of a function based on the background frequency is the p-value. Functions with p-values smaller than 0.05 are considered enriched.

2.4 Execution

Configuration options. Tunable parameters and their default values are accessible in mirLibSpark, including the minimum frequency of sRNA and miRNA, minimum length of sRNA, length range of miRNA, maximum length of pre-miRNA, maximum number of genomic loci, precursor searching range, extra flank length to the validated precursor for folding visualization, temperature used for RNAfold, and parameter options for miRCheck, miRanda and Spark. Several parameters are related to the choice of species, such as genome, index, target file and pathway annotation. If a supported species is chosen, these related parameters are automatically set accordingly. In the case of custom species, users will provide the files and update the parameters.

Inputs. MirLibSpark takes as input one to several sRNA-seq libraries of the same input format. Four input formats are supported: tsv (readcount), reads, FASTA and FASTQ. A guide file (*diffguide_file*) defining the control-experiment pair is required if functional analysis modules are activated for a multi-library task. As an example, for a task analyzing two libraries (Lib1 and Lib2), if the user wishes to define that Lib1 is the control for Lib2, the guide file has to be provided by the user and formatted as follows.

```
Experiment->Control
Lib2->Lib1
```

Outputs. For any request, it would generate the first 10 of the output files as listed in the manual. For a request processing more than two libraries, if differential expression analysis parameter (DF) is set true, then two more outputs are produced; if KEGG pathway enrichment analysis parameter is set true, with the condition that DF is true, then one more output is produced.

3 PREPARATIONS FOR BENCHMARKING

3.1 Construction of simulated data sets

Positive data set. The positive data set (PD) is the entire high confidence *Arabidopsis thaliana* miRNAs of miRBase version 21 [16]. The same miRNA sequences may have different precursors and are registered as different miRNAs. Only unique miRNA sequences were collected for this study. In version 21, there are 349 unique miRNA registries for *Arabidopsis*, and a subset of them are high confidence (100).

Negative data set. The negative data set (ND) was constructed as follows. The universe of non-miRNA is presumably larger than true miRNA. Therefore, the number of ND sequences is 1,000 by applying an arbitrary factor of 10 on the number of PD. The ND sequences were generated by extracting sequences from *Arabidopsis* genome at random positions with lengths between 18 and 25 nt (inclusive). Sequences found in miRBase v21 *Arabidopsis* miRNAs were excluded in the ND collection. Sequences annotated as CDS, rRNA, snoRNA, snRNA and tRNA in TAIR10 were also excluded in the ND collection.

3.2 sRNA-seq libraries

We evaluated the behavior and the performance of mirLibSpark with real sRNA-seq libraries. GSE44622 series (27 libraries in total) [10], GSM518432, GSM738727 [18] and GSM1087974 *Arabidopsis* sRNA-seq libraries were obtained from NCBI Gene Expression Omnibus (GEO) repository. The *Triticum aestivum* sRNA-seq library, S002B35, was obtained from Li et al [19].

3.3 Implementation of a sequential version

The python script, *sequential.py*, is a prediction pipeline that executes in memory the major prediction steps as in the mirLibSpark but without Spark data management or parallelism.

3.4 Implementation of a miRDeep-P wrapper

The miRDeep-P predictor is semi-automated and requires users to execute several scripts and commands one after another. We implemented the python script, *run_mirdeep_p.py*, to execute the prediction of known and novel miRNAs automatically with the same parameters in mirLibSpark.

3.5 miRNA clustering

Similar sequences in the predicted miRNA set are clustered as a group based on the bitscore (greater than 20) given by blastn alignment in blastn-short task mode [2]. Bitscore indicates a log2-scaled estimation of the search space that is required to find a score as good as or better than this one by chance. When the bitscore is 20, we would have to search about 2^{20} = one million sequences to find a sequence similar to our query by chance.

4 BENCHMARKING AND DISCUSSIONS

4.1 Improved execution time

The performance of several plant miRNA predictors is compared with mirLibSpark (shown in Table 1). In the simulated *Arabidopsis* data sets, we incorporated 100 mirBase-v21 high confidence miRNAs as positive data set. We extracted 1,000 unannotated random short sequences from the genome of *Arabidopsis* to construct the negative data set. Execution time and prediction performance of mirLibSpark and the other tools were assessed by the processing of the simulated data sets with specified number of cores. For the real sRNA-seq NGS data, GSN518432 was used [18]. MirLibSpark incorporates several miRNA structure and expression criteria [4, 22]. Therefore, it has better F1 score and accuracy measure than ShortStack on the simulated data sets. Its processing of the *Arabidopsis* sRNA-seq library is 3.6 times faster than that of miR-PREFeR.

Table 1: Performance of Plant miRNA annotation software using (A) simulation data, (B) real world data adapted from [18]

(A)	mirLibSpark	miR-Deep-P	ShortStack
nbCore	32	1	1
Time (sec)	135	194	389
TP	84	69	22
FP	9	57	0
F1	0.87	0.61	0.36
MCC	0.86	0.57	0.45

(B)*	mirLibSpark	ShortStack	mirPrefer	mirDeepP
nbCore	32	1	32	1
Time (sec)	120	2191	577	74601
TP	39	31	56	47
TN	757514	757575	757247	757469
FP	54	1	304	91
FN	151	159	134	143
F1	0.276	0.279	0.204	0.287
MCC	0.293	0.398	0.214	0.29

nbCore: number of cores allocated. Time: execution time in seconds. TP: true positives. FP: false positives. F1: F-measure. MCC: Matthews correlation coefficient. *: Used the standard method to define positive and negative sets in the library, where negative set were the sequences in the library not registered in miRBase, and thus the FP predictions could be novel miRNAs.

4.2 Comparison of performance with other predictors

We ran the programs of mirLibSpark, mirDeep-P, ShortStack and mirPrefer to process the libraries of GSM518432 and GSM738727 and obtained the miRNA predictions of each program, respectively. Axtell and Meyer [4] used a statistic method that is different than ours for Table 1 for assessing the performance. In order to compare their assessments of plant annotation software with mirLibSpark, we adapted their statistic method. We thus defined the positive set as the miRBase-registered miRNAs (P). The true positive predictions were the predicted miRNAs (TP), while the false negative

predictions (FN) were the remaining instances in P. The negative set (N) was defined as the miRBase-registered miRNAs not expressed in the library, and thus the same number across all programs. The performance measures, including false positives, sensitivity, precision and accuracy, are shown in Figure 3. We also overlaid the corresponding data points of the assessments by [4] in our figures, labeled as *Lei1* and *Lei2*. Because of the differences of the parameters used in the programs as well as the versions of the genome, RNAfold and so on, our assessments were not the same as theirs.

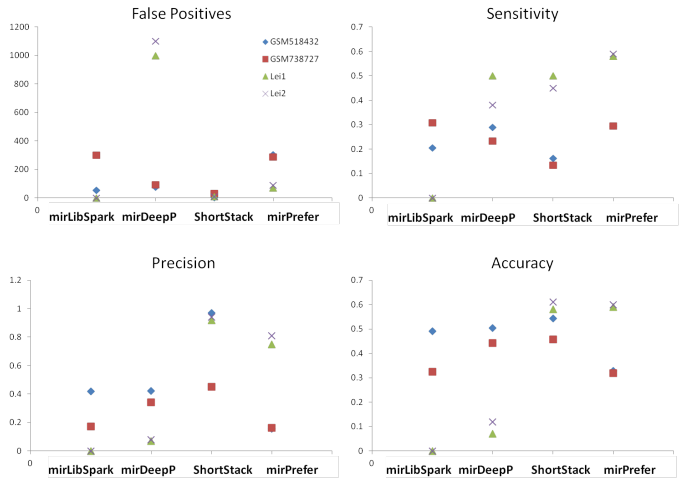


Figure 3: Comparative performance of miRNA annotation software from plants.

4.3 Scalability

MirLibSpark accelerates according to the number of core processors (nbCore). As the nbCore increases, the time of execution decreases and the demand for memory space remains the same for the same input. In addition, mirLibSpark is capable of processing large genome, such as wheat (see Table 2). It reduced the time of processing of a wheat library from more than 30 hours by the non-parallelized program *sequential.py* to 2.2 hours by mirLibSpark with 32 cores.

4.4 Case study

The miRNA prediction was reproduced for 27 *Arabidopsis* sRNA librareis in GEO Series GSE44622 using mirLibSpark. The computation of differential expressions of the predicted miRNAs and the enriched KEGG pathways of their target genes was guided for twenty experiment-control pairs as defined in a *diffguide_file*. The entire process, including both prediction and functional annotation, took about 2 hours on Compute Canada Graham computing cluster (https://docs.computecanada.ca/wiki/Graham). The miRNA prediction part alone took mirLibSpark 5,962 seconds, or 99 seconds per library on average. After clustering the predicted miRNAs based on sequence similarity determined by blastn, the clusters of mirLibSpark prediction were compared with those clusters of the original paper and the miRBase-v21 high/low confidence, using venn digram as shown in Figure 4.

Table 2: Scalability of mirLibSpark for sRNA-seq data from small and large genomes.

(A) ath sRNA-seq lib	nbCore	Time (sec)	MEM (GB)
sequential.py	1	12396	0.8
run_mirdeep_p.py	1	47752	0.7
mirLibSpark	1	2980	14.1
	2	1472	13.2
	3	1071	13.6
	4	837	12.0
	8	501	12.3
	16	270	11.4
	32	215	10.2
(B) wheat sRNA-seq lib	**nbCore**	**Time (sec)**	**MEM (GB)**
sequential.py	1	> 30 hrs *	15.3
mirLibSpark	8	23561	60.3
	16	13476	63.5
	32	8057	62.0

The input library for panel (A) is *Arabidopsis* GSM1087974 of 49 MB with genome 117 MB; for panel (B) is *T. aestivum* S002B35 of 2.1 GB with genome 13 GB. *: The program was allocated 30 hours due to limited resources and terminated before reaching the end of the program. The results of time and memory are the average of three executions of the same settings. Done in Graham computing cluster of Compute Canada with Intel E5-2683 V4 CPUs, running at 2.1 GHz and SATA SSD.

Our results show high degrees of overlapping predictions with Jeong et al [10] and the known miRNAs. Based on our simulation results, the non-overlapping clusters predicted by mirLibSpark (175 clusters) are possibly valid novel miRNAs. However, they used the old version of the prediction criteria [22] while we added the updated criteria [4]. They removed some predicted miRNAs manually according to additional criteria that were not performed by mirLibSpark. For example, if the sum of abundance in the 27 libraries is less than 400, or the genomic loci of the precursors are in siRNA-like regions, the miRNAs are removed [10]. Although we filtered siRNAs using predefined genomic annotations, we did not identify genomic regions that are likely siRNA precursors. Therefore, among our novel miRNAs collections, a fraction of functionally irrelevant predictions might not be excluded based on current criteria.

5 CONCLUSIONS AND FUTURE WORK

We facilitate the settings and annotations of several procedures, including the retrieval of the reference genome, building genomic indices, and installing dependency programs and libraries. MirLibSpark takes different input formats. Most of the validation parameters are flexible in order to meet species-specific preferences or criteria updates. This will allow a quick deployment and ready to use environment. The user manuals for Docker and computing cluster environments, respectively, are available in GitHub.

We introduced the simulated data sets to properly access the performance with statistical measures. We showed an improved accuracy on the prediction. It is scalable regarding multiple libraries and libraries from large genomes. The case study showed high degrees of overlapping with the prediction by others. One can easily

reproduce results among research teams applying the stand-alone, dependency-self-supporting, and fully automated mirLibSpark.

In the future, all the added dependencies will be updated monthly. The miRNAs predicted using mirLibSpark in wheat and other plant libraries will be collected in Wheat MicroRNA Portal [26]. We will also investigate how to standardize the exclusion procedures in real-world data, such as identifying the miRNA precursors residing in the putative siRNA genomic regions and the low occurrences among libraries.

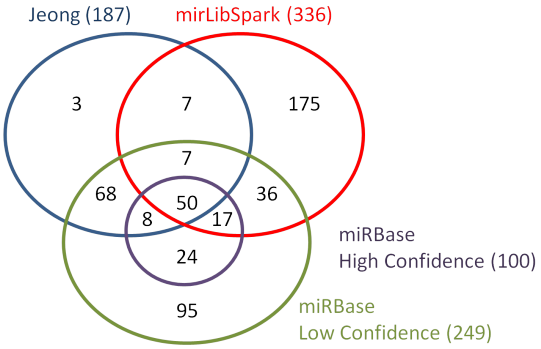


Figure 4: The venn diagram characterizing the predictions of GSE44622 by Jeong et al [10] and mirLibSpark, after a clustering step using blastn to group together similar sequences.

ACKNOWLEDGMENTS

This research was enabled in part by support provided by Calcul Québec and Compute Canada. It has also been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the The Fonds de recherche du Québec - Nature et technologies (FRQNT) for the grants to ABD. CJW and AMR are NSERC and FRQNT fellows. The authors would like to thank Hayda Almeida for testing mirLibSpark and providing useful feedback.

REFERENCES

[1] Zahra Agharbaoui, Mickael Leclercq, Mohamed Amine Remita, Mohamed A Badawi, Etienne Lord, Mario Houde, Jean Danyluk, Abdoulaye Baniré Diallo, and Fathey Sarhan. 2015. An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC genomics* 16, 1 (2015), 339.

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, 17 (1997), 3389–3402.

[3] Michael J Axtell. 2013. ShortStack: comprehensive annotation and quantification of small RNA genes. *Rna* (2013).

[4] Michael J Axtell and Blake C Meyers. 2018. Revisiting Criteria for Plant MicroRNA Annotation in the Era of Big Data. *The Plant Cell* 30, 2 (2018), 272–284.

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[6] Andreas Biegert, Christian Mayer, Michael Remmert, Johannes Söding, and Andrei N Lupas. 2006. The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic acids research* 34, suppl_2 (2006), W335–W339.

[7] Dan M Bolser, Daniel M Staines, Emily Perry, and Paul J Kersey. 2017. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In *Plant Genomics Databases*. Springer, 1–31.

[8] Kévin Darty, Alain Denise, and Yann Ponty. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 15 (2009), 1974.

[9] Marcelo Rodrigo de Castro, Catherine dos Santos Tostes, Alberto MR Dávila, Hermes Senger, and Fabricio AB da Silva. 2017. SparkBLAST: scalable BLAST processing using in-memory operations. *BMC bioinformatics* 18, 1 (2017), 318.

[10] Dong-Hoon Jeong, Shawn R Thatcher, Rebecca SH Brown, Jixian Zhai, Sunhee Park, Linda A Rymarquis, Blake C Meyers, and Pamela J Green. 2013. Comprehensive investigation of microRNAs enhanced by analysis of sequence variants, expression patterns, ARGONAUTE loading, and target cleavage. *Plant physiology* 162, 3 (2013), 1225–1245.

[11] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. 2004. Human microRNA targets. *PLoS biology* 2, 11 (2004), e363.

[12] Matthew W Jones-Rhoades and David P Bartel. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular cell* 14, 6 (2004), 787–799.

[13] Arnoud J Kal, Anton Jan van Zonneveld, Vladimir Benes, Marlene van den Berg, Marian Groot Koerkamp, Kaj Albermann, Normann Strack, Jan M Ruijter, Alexandra Richter, Bernard Dujon, et al. 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular biology of the cell* 10, 6 (1999), 1859–1872.

[14] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45, D1 (2017), D353–D361.

[15] Ilias Kanellos, Thanasis Vergoulis, Dimitris Sacharidis, Theodore Dalamagas, Artemis Hatzigeorgiou, Stelios Sartzetakis, and Timos Sellis. 2014. MR-microT: a MapReduce-based MicroRNA target prediction method. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*. ACM, 47.

[16] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. 2018. miRBase: from microRNA sequences to function. *Nucleic acids research* 47, D1 (2018), D155–D162.

[17] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10, 3 (2009), R25.

[18] Jikai Lei and Yanni Sun. 2014. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics* 30, 19 (2014), 2837–2839.

[19] Qiang Li, Brook Byrns, Mohamed A Badawi, Abdoulaye Banire Diallo, Jean Danyluk, Fathey Sarhan, Debbie Laudencia-Chingcuanco, Jitao Zou, and D Brian Fowler. 2018. Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant physiology* 176, 3 (2018), 2376–2394.

[20] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 6, 1 (24 Nov 2011), 26.

[21] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* (2010).

[22] Blake C Meyers, Michael J Axtell, Bonnie Bartel, David P Bartel, David Baulcombe, John L Bowman, Xiaofeng Cao, James C Carrington, Xuemei Chen, Pamela J Green, et al. 2008. Criteria for annotation of plant MicroRNAs. *The Plant Cell* 20, 12 (2008), 3186–3190.

[23] Aleksandr Morgulis, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology* 13, 5 (2006), 1028–1040.

[24] Abhinav Nellore, Leonardo Collado-Torres, Andrew E Jaffe, José Alquicira-Hernández, Christopher Wilks, Jacob Pritt, James Morton, Jeffrey T Leek, and Ben Langmead. 2016. Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* (2016), btw575.

[25] Ren-Hao Pan, Lin-Yu Tseng, I-En Liao, Chien-Lung Chan, K Robert Lai, and Kai-Biao Lin. 2015. Design of an NGS MicroRNA predictor using multilayer hierarchical MapReduce framework. In *Data Science and Advanced Analytics (DSAA). 36678. IEEE International Conference on*. IEEE, 1–8.

[26] Mohamed Amine Remita, Etienne Lord, Zahra Agharbaoui, Mickael Leclercq, Mohamed A Badawi, Fathey Sarhan, and Abdoulaye Baniré Diallo. 2016. A novel comprehensive wheat miRNA database, including related bioinformatics software. *Current Plant Biology* 7 (2016), 31–33.

[27] Ioannis S Vlachos, Konstantinos Zagganas, Maria D Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. 2015. DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic acids research* (2015), gkv403.

[28] Xiaozeng Yang and Lei Li. 2011. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 27, 18 (2011), 2614–2615.

[29] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2–2.

A.2 MirLibSpark manual

microRNA validation using Spark framework

BSD-3-Clause license

3 stars 0 forks 2 watching Branches Activity Custom properties
Tags

Public repository

master 4 Branches 2 Tags Go to file t Go to file Add file Code

JulieCJWu Merge remote-tracking branch 'juliecjwumirlibspark/master' 5ea8c82 · 4 years ago

dbs	mirdup as an option (#95)	7 years ago
input_samples	Docker (#98)	6 years ago
lib	mirdup as an option (#95)	7 years ago
src	resolved nbLoc issue (#10)	4 years ago
workdir	remove the submission file that builds spark cluter	5 years ago
.gitattributes	Minor modifs (#34)	8 years ago
.gitignore	resolved nbLoc issue (#10)	4 years ago
LICENSE	Create LICENSE	8 years ago
README.md	update branches purpose	5 years ago

mirLibSpark

A plant microRNA prediction and validation software using Spark framework.

Its prediction procedures and parameters are optimized for plants and comply with the gold-standard requirement on miRNA predictions.

It is a standalone and fully automated package with zero-dependency. **Fully automated**: one command line to analyze your data, and annotate miRNAs and their functions in one run. **Zero-dependency**: dependencies are automatically installed.

Citation:

Please cite the paper for your work using this tool.

Title: MirLibSpark: A Scalable NGS Plant MicroRNA Prediction Pipeline for Multi-Library Functional Annotation

Cite: Chao-Jung Wu, Amine M. Remita and Abdoulaye Baniré Diallo. 2019, September. MirLibSpark: A Scalable NGS Plant MicroRNA Prediction Pipeline for Multi-Library Functional Annotation. ParBio'19-8th Workshop on Parallel and Cloud-based Bioinformatics and Biomedicine. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Pages 669-674. ACM.

BibTex:

```
@article{wu2019mirlibspark,
  title={MirLibSpark: A Scalable NGS Plant MicroRNA Prediction Pipeline for Multi-Library Functional Annotation},
```

```
  author={Wu, Chao-Jung and Remita, Amine M. and Diallo, Abdoulaye Banir{\'e}},
  journal={ParBio'19-8th Workshop on Parallel and Cloud-based Bioinformatics and Biomedicine. In Proceedings of the 10th ACM
International Conference on Bioinformatics, Computational Biology and Health Informatics},
  pages={669-674},
  year={2019},
  publisher={ACM}
}
```

Link to download the paper: https://dl.acm.org/citation.cfm?id=3343463

Branches:

`master` : the version used in ACMBCB-ParBiopaper `meyers18criteria` : pipeline fullfill meyers18 criteria `reportRemoved` : based on meyers18criteria branch, report all filters rejcting reasons step by step. `checkreason` : based on meyers18criteria branch, report MIRcheck rejecting reasons, and MIRcheck not filtering

Dependencies:

Languages: `python2` (scipy, statsmodels), `perl` , `java`

Packages: `pyspark` , `duskmasker` , `bowtie` , `RNAfold`

Distributed with mirLibSpark: `miRanda` , `VARNA` , `miRdup` , `miRCheck`

Environments: Users run `mirLibSpark` in one of the following environments: `Docker` in a personal computer; or `Servers` ; or `Linux` in a personal computer (In linux option, user has to install dependencies).

Docker mode requres some beginner to intermediate bioinformatics skills, and is recommended for biologists. Advanced users with high-throughput computation skills are encouraged to use server mode for an enhanced computational power.

Two manuals are provided as follows, one for *docker mode*, one for *server mode*.

A. Manual for Docker

Users need to have installed `Docker Desktop` in your computer or any suitable machines. Please refer to https://www.docker.com/products/docker-desktop.

Then follow this manual to pull `mirLibSpark` docker image and perform runs.

Further editing of the options is NOT required but permitted. For example, pipeline parameters are already optimized for plants but users can modify them if the experts' criteria evolve in the future.

All dependencies will be installed automatically upon the installation of mirLibSpark docker image.

Genomic annotation files for *Arabidopsis* are included with installation. Files for other supported or custom species can be installed and indexed using a simple command line in `mirLibSpark` (see this manual).

`Pull and run mirLibSpark image.`

The following command lines will download mirLibSpark image from Docker Hub, prepare the environment, and initiate an interactive mirLibSpark docker session.

```
mkdir input output
docker pull juliewu/mirlibsparkdocker:latest
docker run -v abs_path/to/output:/mirLibSpark/output -v abs_path/to/input:/mirLibSpark/input -it
juliewu/mirlibsparkdocker:latest
```

Once you are inside the docker image, you are in the `src` folder of the `mirLibSpark project` .

Users can analyze the miRNAs using: (A1) `prediction` pipeline, or (A2) `prediction + differential analysis` pipeline, or (A3) `prediction + differential analysis + enrichment analysis` pipeline. The entire analysis procedure is `A3` , KEGG pathway enrichment analysis pipeline, while A1 and A2 perform part of the analysis procedure and are shown for pedagogic purpose in this manual.

A1. prediction pipeline

Using one to several *Arabidopsis* library. Step 1: put some RNA-seq libraries in input folder. Use a small demo file for a quick test; or use an *Arabidopsis* library `GSM1087974` (100.txt) as an example. Make sure input is empty before copying.

```
rm -fr ../input/*
cp ../input_samples/library_a_len1.txt ../input
```

or

```
cp ../input_samples/100.txt ../input
```

Step 2: verify mirLibSpark program parameters in `stdout` with `--dummy`. See the parameter options with `--help`.

```
spark-submit mirLibPipeline.py --dummy
```

Step 3: execute mirLibSpark program from `src` folder. Modify the parameters as needed. Note that the run takes minutes to a few hours, depending on available resources in the hardware.

```
spark-submit mirLibPipeline.py --sc_heartbeat 300
```

Step 5: run descriptions are shown in `stdout`. When the run is done, find the reports in `output` folder. The name of the report folder is the same as the job id created by linux time. The description of the report files is listed in the end of this manual.

A2. Differential analysis pipeline.

This analysis requires users to define the `Experiment-Control` pairs.

Step 1: put two or more files in `input` folder. Make sure that no other file is located in `input` folder. Assume you are in the `src` folder of the `project`. Use demo files as an example.

```
rm -fr ../input/*
cp ../input_samples/library_a_len1.txt ../input
cp ../input_samples/library_b_len3.txt ../input
```

Step 2: edit the *diffguide* file `diffguide_ath.txt` as needed, in `src` folder. It looks like this:

> Experiment->Control

> library_b_len3->library_a_len1

Users use *diffguide* file to tell the program your experimental design. In this example, one experiment-control pair is defined. Users need to make sure the mentioned libraries are provided in `input` folder.

Because users can not edit files inside the docker, users can do the following steps to edit a file:

```
cp diffguide_ath.txt ../input
```

Then users open and edit the file from your local computer, outside the Docker machine. Once the editing is done, do the following:

```
mv ../input/diffguide_ath.txt .
```

Step 3: execute mirLibSpark program from `src` folder.

```
spark-submit mirLibPipeline.py --perform_differential_analysis --diffguide_file diffguide_ath.txt --sc_heartbeat 300
```

A3. KEGG pathway enrichment analysis pipeline.

Execute mirLibSpark program from `src` folder

```
spark-submit mirLibPipeline.py --perform_differential_analysis --diffguide_file diffguide_ath.txt --
perform_KEGGpathways_enrichment_analysis --sc_heartbeat 300
```

Because this analysis depends on the activation of differential analysis, users may notice that in the command line above, both *perform_differential_analysis* and *perform_KEGGpathways_enrichment_analysis* are flagged.

A4. Build supported species dbs

Execute one of the following commands from `src` folder. Please select ONE of the command lines that corresponds to your preferred species.

```
python init_dbs_ensembl40_v2.py wheat 2 curl-build
python init_dbs_ensembl40_v2.py corn 2 curl-build
python init_dbs_ensembl40_v2.py rice 1 curl-build
python init_dbs_ensembl40_v2.py potato 1 curl-build
python init_dbs_ensembl40_v2.py brome 1 curl-build
```

A5. Build custom species dbs

We have supported the indexing for several popular plant species (ath, wheat, corn, rice, potato, brome). If your preferred species has not yet been supported, we provide the command line to index your custom species.

Step 1: put your custom genome fasta file in `input` folder. The genome file may contain one to several sequences in fasta format.

step 2: execute mirLibSpark program from `src` folder.

```
python init_dbs_customGenome.py ../input/speciesname.fasta
```

Resulting dbs files can to be copied outside of image for future reuse, because closing the docker image will remove all custom changes.

B. Manual for Servers

Follow this manual to download `mirLibSpark project` from GitHub and perform runs in a remote server.

Users can choose to use Compute Canada servers or any other servers. In order to run `mirLibSpark` in a server, users only need to edit the submission file.

The submission files are server-specific, and therefore we provide a template for Graham server in Compute Canada. Compute Canada users only need to edit their user credentials in the submission file template.

For other servers, such as Microsoft Azure, please refer to the service provider for the instruction of the task submission procedures.

Further editing of the options is NOT required but permitted. For example, pipeline parameters are already optimized for plants but users can modify them if the experts' criteria evolve in the future.

All dependencies will be loaded automatically upon task execution. Docker is not used in this mode.

Supporting files for *Arabidopsis* are included with installation. Files for other supported or custom species can be installed and indexed using a simple command line in `mirLibSpark` .

If users are not familiar with the submission procedures in remote servers, users can choose to use Docker in your computer (See section *A. Manual for Docker*).

`git clone` mirLibSpark project.

Install the `mirLibSpark project` in your Compute Canada account.

```
git clone git@github.com:JulieCJWu/mirLibSpark.git
```

or download the latest release of source codes https://github.com/bioinfoUQAM/mirLibSpark/releases

B1. Pipeline usage

Step 1: put some RNA-seq libraries in `input` folder. Use a small demo file for a quick test; or use an *Arabidopsis* library GSM1087974 (100.txt) as an example.

```
cd mirLibSpark
mkdir input
cp ../input_samples/library_a_len1.txt ../input
```

Step 2: edit the submission file `mirlibspark_submission.sh` Use your favorite editor to open and edit the file. Execute `mirLibSpark` program from `workdir` folder.

```
cd workdir
vim mirlibspark_submission.sh
```

(1) General settings Set the following parameters

> #SBATCH --account=yourSponsorAccount

> #SBATCH --time=hh:mm:ss

> #SBATCH --mail-user=yourname@email.com

(2) Execution (around line 63)

Use the following command directly or modify it as needed.

> spark-submit --master ${MASTER_URL} --executor-memory ${SLURM_MEM_PER_NODE}M ../src/mirLibPipeline.py

Use `--dummy` to test your settings.

> spark-submit --master ${MASTER_URL} --executor-memory ${SLURM_MEM_PER_NODE}M ../src/mirLibPipeline.py --dummy

Note that users can analyze the miRNAs using: (1) `prediction` pipeline, or (2) `prediction + differential analysis` pipeline, or (3) `prediction + differential analysis + enrichment analysis` pipeline. The entire analysis procedure is `(3)`, while (1) and (2) perform part of the analysis procedure are shown for pedagogic purpose in this manual. For details of how to set up the command lines, please see in this manual section A1, A2, and A3, respectievly.

Step 3: submit the submission file in the job queue

```
sbatch mirlibspark_submission.sh
```

B2. Build supported species dbs

This section presents how to build dbs in a remote server. Alternatively, users can prepare the dbs files using Docker mode in your local computer, and then copy the resulting files to your remote server.

Because in submission runs, there might not be internet access, the build is done in two steps: first download the files from internet, then use submission file to build the dbs index.

Step1: execute one of the following commands from `src` folder. Please select the ONE of the following command lines that corresponds to your preferred species.

```
python init_dbs_ensembl40_v2.py wheat 2 curl
python init_dbs_ensembl40_v2.py corn 2 curl
python init_dbs_ensembl40_v2.py rice 1 curl
python init_dbs_ensembl40_v2.py potato 1 curl
python init_dbs_ensembl40_v2.py brome 1 curl
python init_dbs_ensembl40_v2.py wheatD 1 curl
```

Step 2: edit the submission file `submit_init_dbs_ensembl40.sh` and then submit the task.

```
vim submit_init_dbs_ensembl40.sh
```


Remove ONE of the `#` to activate the desired species. Submit the submission file in the job queue.

```
sbatch submit_init_dbs_ensembl40.sh
```

Users can build custom species dbs index as described in a previous section, *A5. Build custom species dbs*.

README BSD-3-Clause license

Name of output	Description
(1) parameters_and_time.txt	the register of user parameters and the execution time records.
(2-4) miRNAprediction_Lib1.txt; _Lib2.txt; and _Lib3.txt	the details of predicted miRNAs that pass the miRNA criteria in each library.
(5) summaryBinary.txt	a tabulated table indicating the presence of predicted miRNAs in each library.
(6) summaryFreq.txt	a tabulated table indicating the expression abundance of predicted miRNAs in each library.
(7) precursorindex.txt	the details of all predicted miRNAs in terms of unique genome coordinates.
(8) precursorindex.html	an HTML table displaying the 2D structures of each precursor drawn by VARNA software.
(9) mirna_and_targets.txt	the targets of each miRNA predicted by miRanda software.
(10) targetsKEGGpathway.txt	the KEGG pathways of each target genes.
(11-12) differential_Lib2vsLib1.txt; and _Lib3vsLib1.txt	the statistics report of differential expressed miRNAs in Lib2 or Lib3 using Lib1 as baseline.
(13) enrichment_pval_upper.csv	the statistics report as a table listing enriched KEGG pathways in each library.

Releases 2

mirLibSpark1.00 Latest
on Oct 15, 2019

+ 1 release

Packages

No packages published
Publish your first package

Contributors 2

JulieCJWu Chao-Jung Wu

maremita Amine M. Remita

Languages

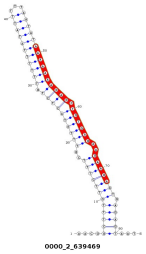
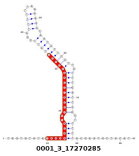
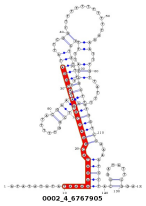
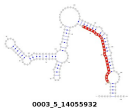
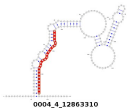
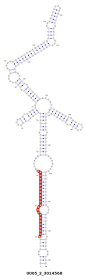
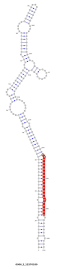
Java 70.1% Python 21.4% Perl 8.2% Shell 0.3%

Suggested workflows

Based on your tech stack

miRNAs and their genomic precursors

A.3 Example html output

Serial	pre-miRNA.Structure.Visualization	miRNA.Sequence & Coordination / pre-miRNA.Sequence / pre-miRNA.Structure
0000		AAAAAAGAGCTGTGAGAAGAAGAA Chr 2:639469 [+] AACGGCTTTGTTCTTTCGTTTATACTCTGTTTTTTATATTTTTAATATAAAAAAAGAGCTGTGAGAAGAAGA ATGAGAGTAAT ....(((((((((((..((((((((((. ((((((((((...))))))))))))))))).)))))).)))))))).)))))...
0001		AAAAAGATACGGATCGAAGCGGAT Chr 3:17270285 [-] TAATATATATAAAAAGATACGGATCGAAGCGGATATTCGATTTTTTAAAAAACAATATTCGTTATTTGATTC GGTTTTTACGAATACTACAT .............((((..((((((((((((((((... ((((....)))).)))))))))..)))))))))..)))).............
0002		AAAAAGATACGGATCGAAGCGGAT Chr 4:6767905 [-] ATATATATATAAAAAGATACGGATCGAAGCGGATATTCGGTTTTTTTTTTTTTAAAATTTGTTATTTGATTCG GTTTTTACGAATACTACATTTTCCTATTTGCTTTGGTCCAAAATCTTACGGATATTCGGA .............(((((..((((((((((((((...((..............((..(((..... (((((.......)))))...)))..)))))))))))))))))))...))))).((.....))..
0003		AAAACAACCAAAGGACATGACGAC Chr 5:14055932 [-] TAAACTTTTCAAAACAACCAAAGGACATGACGACAATGTTCGAAGAAAGTCCTTAGTAATCTCGAGATGAGA TATTCCATGCTTTACTCAAAAAAGGTAATAATCTCATATCATCGAACTCCTTCAATTTTTCATGTTGTTTTT GTTTCTTCGGTTGTAGATAGAAAGATG ....(((((....(((((((.(((((((.(((((((..(((((((.((.....(((((....)).))) (((((((((.... ((((((........))))))))..)))))))..)).))))))..............)))))))..)).)))))..)))))))...
0004		AAAACAGATCTCGCCATACTGCAA Chr 4:12863310 [-] TACCTTATACTTGTAGTATGAAATGCTATCTTTTGCTAGTTATCTTAGTTGTTTTCCTCCTCCTTTAGCTTG TTTGTCATCTACTCTTGTTCTTGACGAATATTTATTTTTGTATTAGTGTTAAGCTTTGTTTCTAAGATGCCA AAACAGATCTCGCCATACTGCAAACTTGTTAAA ..........((((((((((...((..(((((((((...((.(((((((.................. (((((((((((... ((....))...)))))))))..................))))).....))))))))))))))..)))))..)).))))))))
0005		AAAAGGGCATTGAGAAAAGGGCA Chr 2:3014568 [-] GTAGAGTTGCAAAAGGGCATTGAGAAAAGGGCATTAGAGTGAATTATTCATAATAATACCACAGTTTTATAT AAAATCTATTCAACCAAAATAAATTAGAAATAAATTAAAAATATTATATTTGGGTTATCTATTTCTAATTTA GGGTATACATTTTAACTATTTAAAATAAATTTTGATTCATTTTAATATGGGAGATTTGCAAAACTACCCTTT TTTTACTACCCTTTTCCAACAATAC (((..((((.(((((((...((((((((((.......(((((((.((((((((((.((....)).)))))). ((((((.(((....((....(((((((((((((((.. ((((...))))))..))))...)))))))))))))))..)))..))))))).....(((((((. (((....))).)))))))))))))).)))))))..............)))))))...)))))..)))
0006		AAAATATAGAGTTGGGCATTT Chr 2:12270189 [-] AGACCATCTCCAATGCGCAACTCTATATTTCCTCTATTTTACAGTTGACTCTATAAAATATAGTTGATTTTA CTGCATTGTATGCCTCTAAAATAGAGGAAAAATGTAGAGATTGCTCTTTTATTAGCTCAACTATATATTCTC TAAACTCTAAAATAGAGGAAAATATAGAGTTGGGCATTTGAGATGCTCT (((.((((((.(((((.(((((((((((((((((((((((((.((((.....(((...(((((((((.(((. (((((((.....((((((...))))))...)))))))))).. (((........)))))))))))))))))).....)))).)))))))))))).)))))))))))).))))).)))))).)))
0007		AAAATATAGAGTTGGGCATTTG Chr 2:12270188 [-] AAGACCATCTCCAATGCGCAACTCTATATTTCCTCTATTTTACAGTTGACTCTATAAAATATAGTTGATTTT ACTGCATTGTATGCCTCTAAAATAGAGGAAAAATGTAGAGATTGCTCTTTTATTAGCTCAACTATATATTCT CTAAACTCTAAAATAGAGGAAAATATAGAGTTGGGCATTTGAGATGCTCTA

APPENDIX B

WHEAT PROJECT

Prediction and Evolutionary Dynamics of Wheat miRNAs and siRNAs During Field Cold Acclimation

Chao-Jung Wu[1], Farzaneh Rahmdani[1], Zahra Agharbaoui[2], Mohamed Badawi[2], Jean Danyluk[2], Qiang Li[3,4], David Brian Fowler[3], Fathey Sarhan[2], and Abdoulaye Banire Diallo[1]

[1]*Département d'informatique, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada*
[2]*Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada*
[3]*National Research Council Canada, Saskatoon, Saskatchewan S7N 0W9, Canada*
[4]*Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5A8, Canada*

Abstract

MicroRNAs (miRNAs) and small-interfering RNAs (siRNAs) are key regulators of plant stress responses. While siRNAs share biogenesis and silencing mechanisms with miRNAs, their role in stress adaptation remains less explored. Here, we introduce a unified bioinformatics workflow to simultaneously analyze miRNAs and siRNAs involved in wheat cold acclimation. We use *interfering RNAs (iRNAs)* to collectively refer to miRNAs and siRNAs. Using a large sRNA-seq dataset from field-grown wheat cultivars with varying *VRN-A1* expression and freezing tolerance, we identified 339 iRNAs, including 96 novel iRNAs, 33 trans-acting siRNAs, 90 dual-identity iRNAs, and 66 cold response iRNAs (58 miRNAs and 9 siRNAs). To enhance prediction confidence, miRNAs and siRNAs were validated using several bioinformatics approaches, including miRBase, literature on wheat miRNA, ShortStack, siRNA collections, and BLASTN. Evolutionary tracing suggested that wheat miRNAs and dual-identity iRNAs likely originate from ancestral subgenomes, while siRNAs and novel iRNAs emerged more recently. siRNAs primarily derive from coding regions, while those originating from transposable elements may be rapidly purged. The lack of ancestral enrichment in cold-response iRNAs suggests that recent selective pressures shaped their role in cold adaptation. Functional analysis showed that miRNAs and siRNAs jointly regulate cold adaptation by targeting transcription factors, kinases, and hormone signaling pathways, and may contribute to adaptation through *Vrn-A1*-dependent vernalization. Our findings provide insights into the genomic origins and functional diversification of cold-responsive iRNAs, revealing a coordinated regulatory network governing wheat stress adaptation. This study underscores the necessity of a unified approach to small RNA analysis and offers a comprehensive framework for investigating regulatory RNA evolution and function in plants.

Keywords: cold acclimation, wheat small RNA prediction, small RNAs in plant stress response, small RNA evolution, bioinformatics pipeline for small RNAs

1 introduction

1.1 Justification for Simultaneous Prediction of miRNAs and siRNAs

Among the RNAs detectable in next-generation small RNA sequencing (NGS sRNA-seq), miRNAs and siRNAs are two major classes of RNA effectors involved in RNA interference (RNAi) [43, 28]. miRNAs and siRNAs are functionally and evolutionarily intertwined, sharing biogenesis, mode of action, and regulatory mechanisms. Both are processed by Dicer-like enzymes into small RNA duplexes, loaded onto Argonaute proteins, and guide RNAi to regulate gene expression through mRNA degradation, translational repression, or chromatin modification [43, 28]. In addition to gene silencing, RNAi can activate

target transcription [15, 80]. Their evolutionary connection is evident as siRNAs originally served as an ancient defense mechanism against transposable elements (TEs) and viruses, while primitive miRNAs likely evolved from siRNA-hosting intermediates precursors to provide more precise regulatory control and complexity in plants [58]. Certain TE-related precursors can generate both miRNAs and siRNAs [79, 58], and the miRNA families MIR159 and MIR319 may have originated from a phased stem-loop precursor [57]. Some miRNAs target the 3'-UTR and lead to mRNA up-regulation, which was thought to be a functional characteristic of siRNAs [94]. These findings blur the boundary between these RNA classes. Furthermore, both small RNA types are involved in plant stress responses, highlighting their overlapping roles in gene regulation [10, 36, 72].

Given these shared features, we adopt the term **interfering RNAs (iRNAs)** as a broad category encompassing both miRNAs and siRNAs. We argue that miRNAs and siRNAs should be analyzed as a unified regulatory system to provide deeper insights into small RNA-mediated regulation, rather than treating them as separate entities.

1.2 Categorizing Interfering RNAs

Canonical miRNA precursors are single-stranded RNA molecules transcribed from the genome, typically forming hairpin structures and measuring less than 300 nucleotides in length [8]. In contrast, siRNAs arise from double-stranded RNAs, which can form through bidirectional transcription, complementary strand pairing, or hairpin structures similar to miRNA precursors [6]. Various **siRNA subclasses** exist, arising from coding and non-coding transcripts and undergoing distinct processing mechanisms. These include natural antisense siRNAs (natsiRNAs), derived from overlapping transcripts; trans-acting siRNAs (tasiRNAs), derived from TAS genes and targeting mRNAs at different loci [40]; heterochromatic siRNAs (hc-siRNAs), involved in heterochromatin formation [96]; piwi-interacting RNAs (piRNAs), expressed in animal reproductive cells and associated with specific proteins [32]; mirtrons, processed from spliced introns [85]; and phased siRNAs (PHAS), which generate regularly spaced small RNAs [45].

Although miRNAs and siRNA subtypes have diverse origins, they are all processed by Dicer-like proteins, making their identification feasible using bioinformatics approaches [6]. Understanding these subclasses not only enhances small RNA prediction and annotation but also refines our ability to systematically analyze their regulatory roles in plant stress adaptation.

1.3 Current Studies on miRNAs and siRNAs in Plants

Over the past two decades, **miRNAs** have been extensively studied as key post-transcriptional regulators in plant development, stress responses, and genome maintenance [69, 103, 63, 53, 3, 1, 14]. Numerous studies have demonstrated their crucial roles in plant responses to abiotic stresses such as drought, heat, cold, and salinity [59, 48, 92, 99, 89, 31, 1, 68, 88]. These regulatory networks are finely tuned to specific species, tissue types, and environmental conditions [105, 38, 1, 88]. The availability of curated databases like miRBase [52] has further facilitated systematic studies of miRNA-mediated gene regulation, enabling comprehensive comparative analyses across plant species.

In contrast, **siRNAs** remain comparatively underexplored, largely due to the lack of widely adopted computational tools and dedicated resources similar to those for miRNAs [64]. While challenges persist,

several studies have sought to elucidate the genomic origins and regulatory roles of siRNAs in plants. In *Arabidopsis*, tasiRNAs [17, 40] and natsiRNAs [116] have been shown to mediate transcriptional and post-transcriptional silencing through highly coordinated regulatory mechanisms. The discovery of 24-nucleotide (nt) phased siRNAs in maize anthers [113] highlighted the presence of tissue-specific PHAS-24 siRNA networks that play essential roles in reproductive development. Interestingly, this strategy is similar to piRNAs in animal reproductive cells, where they associate with Argonaute/Piwi proteins to mediate mRNA cleavage [32].

Although several specialized siRNA databases exist, no centralized repository provides systematic and continuously updated siRNA annotations akin to miRBase for miRNAs [52]. The MPSS database archives plant PHAS loci from *Arabidopsis*, maize, and rice [74, 45], while Penn State University's plant small RNA genes repository (PSU Catalogue) compiles siRNA predictions across 47 plant species [65, 7]. However, these collections are fragmented, and comprehensive *in silico* predictions for many plant genomes, including wheat, remain largely incomplete.

The absence of a standardized siRNA reference hinders large-scale comparative analyses and complicates the development and validation of computational models for siRNA classification. As a result, small RNA classes are often analyzed in isolation rather than as interconnected components of a broader regulatory network.

1.4 Wheat as a Model for iRNA Evolution and Stress Responses

As miRNAs and siRNAs play key roles in plant stress responses [51], their thorough identification and target analysis are essential for uncovering the intricate regulatory networks that govern physiological processes. Wheat (*Triticum aestivum* L.), a major global staple crop, is an ideal model for studying iRNA-mediated stress responses due to its polyploid genome and the variation in freezing tolerance among cultivars. As a hexaploid species (2n=6x=42, AABBDD) resulting from the divergence and hybridization of three progenitor genomes, wheat offers a unique system for exploring small RNA evolution. Its complex genome facilitates comparative analyses of regulatory conservation, divergence, and functional innovation across homologous gene networks [70].

Although siRNAs have been less studied, recent research highlights their regulatory roles in wheat stress responses and genome stability. For instance, Yao et al. [109] demonstrated that non-coding RNAs, excluding miRNAs, are upregulated under abiotic stress, guiding mRNA degradation to modulate stress-responsive pathways. Tang et al. [92] extended this finding by identifying a complex network of siRNAs involved in wheat's response to cold stress, refining the differential expression of genes essential for stress tolerance. Sun et al. [90] reported abundant TE-associated siRNAs in developing wheat grain, suggesting their involvement in gene regulation and genome stability maintenance. Kenan-Eichler et al. [50] found that hybridization and polyploidization events lead to siRNA deregulation, influencing gene expression patterns that affect wheat's environmental adaptability.

Moreover, the availability of cold-acclimating and non-acclimating cultivars enables the dissection of small RNA-mediated regulation in stress adaptation, further establishing wheat as an exceptional model for investigating the role of small RNAs in both adaptive responses and evolutionary dynamics.

1.5 Computational Challenges in Small RNA Prediction in Wheat

Despite advancements in small RNA research, predicting miRNAs and siRNAs in wheat remains difficult due to the genome's large size (13 GB), high redundancy, and polyploid complexity [5]. Effective small RNA classification requires predictive models that integrate knowledge of their biogenesis, functional roles, and regulatory mechanisms and are computationally scalable for large genomic and transcriptomic datasets.

While miRNA annotation benefits from well-defined structural criteria [8], siRNA prediction remains constrained by the structural diversity of siRNA precursors and the lack of dedicated tools. To address both miRNA and siRNA prediction constraints, we integrate phase scoring, which detects siRNAs generated through consecutive Dicer digestion cycles [40], and an established miRNA predictor [104], into a scalable computational framework [111] to ensure efficient and accurate small RNA identification.

Capturing the full complexity of small RNA dynamics in wheat requires a biologically meaningful dataset that accounts for genetic variation, temporal changes, and real-world environmental conditions. We leverage a large sRNA-seq dataset from wheat cultivars with varying freezing tolerance, sampled across warm and freezing months. By analyzing field-grown plants, we ensure our findings reflect miRNA and siRNA regulation under natural stress challenges.

1.6 Major Contributions

In this study, we present an integrated bioinformatics and statistical workflow for predicting miRNAs and siRNAs, analyzing their differential expression, and characterizing their functional roles in wheat cold acclimation. This workflow addresses key limitations in small RNA analysis by incorporating a scalable miRNA prediction tool for wheat's large genome, integrating a well-accepted scoring method for siRNA precursor identification, and unifying the analysis of miRNAs and siRNAs, which are often studied separately regardless of their shared biogenesis and functional interplay. The workflow was deployed on a large sRNA-seq dataset of wheat grown under natural field conditions. We performed a statistical functional analysis from late summer to the freezing season on winter and spring cultivars that differ in their *VRN-A1* gene expression levels and their ability to develop freezing tolerance, using the same wheat samples for which we already understand their cold response at the mRNA level [56]. The primary contributions of this study include:

1. **Identification of Dual-Identity iRNAs** – Our dual-predictor approach revealed 90 iRNAs with both miRNA and siRNA characteristics, suggesting an evolutionary and functional link between the two RNA classes.

2. **Discovery of Novel iRNAs** – We identified 96 previously unreported miRNAs and siRNAs, expanding the known small RNA repertoire in wheat. Some novel iRNAs regulate cold responses, underscoring their potential role in stress adaptation.

3. **Genomic Origins of iRNAs** – We traced the genomic loci that generated miRNAs and siRNAs into distinct categories, including ancestral genomes, coding regions, TAS homologs, and transpos-

able elements. Their distribution patterns provide insights into their evolutionary trajectories and regulatory roles.

4. **Identification of cold-response iRNAs** – We identified 66 cold-response iRNAs with distinct expression patterns across wheat cultivars, revealing their involvement in stress adaptation. These iRNAs regulate key genes associated with cold tolerance, including kinases and transcription factors, and exhibit cultivar-specific regulatory roles.

5. **A Unified Small RNA Analysis Framework** – We developed a comprehensive bioinformatics workflow that integrates miRNA and siRNA predictions, differential expression analysis, and functional annotation.

This work serves as both a case study and a methodological framework for future small RNA research in complex plant genomes. By demonstrating the feasibility and necessity of integrating miRNA and siRNA analysis, we highlight the evolutionary significance and functional interplay of these small RNAs, thereby providing a more systematic methodology for investigating small RNA-mediated stress responses in plants.

2 Results

2.1 Integrated miRNA-siRNA Prediction

RNA interference pathways are crucial for gene regulation, but genome-wide studies have mainly focused on miRNAs, neglecting siRNAs. To bridge this gap, we developed a computational framework that predicts both miRNAs and siRNAs from wheat small RNA sequencing data, and accounts for differential expression, regulation, and functional analyses (Figure 1). Predicting miRNA and siRNA subtypes simultaneously from the same datasets allows direct comparisons of their genomic origins, expression patterns, and regulatory roles. We incorporated a high accuracy, scalable miRNA predictor adhering to gold-standard plant miRNA annotation criteria [104], to address computational challenges in large-genome species like wheat. We also created a scalable, Python-based in-house classifier that uses phasing scores and precursor structure constraints to detect phased siRNA loci [40]. By consolidating miRNA and siRNA predictions into a single analytical pipeline, we expanded the classification scope and supported the assumption that they share common biogenesis pathways and regulatory functions. Analyzing multiple libraries allowed for the removal of false iRNA families by retaining only those with recurrent detection across samples and sufficient read support, leading to the identification of 339 iRNAs, including 215 miRNAs and 127 siRNAs (Figure 2, Table S19).

miRNAs and siRNAs are usually categorized based on their genomic origin, alongside the structural and functional characteristics of their precursors [95], which reflect their biogenesis pathways and evolutionary trajectories. Through comparative analysis, miRNAs and siRNAs can also be classified as broadly conserved or lineage-specific, showing their conservation, diversification, and regulatory roles. We categorized the 339 iRNAs into three groups based on their precursor loci: ancestral genomes, coding regions, and transposable elements. An iRNA was classified as ancestral if it originated from the A and B progenitor genomes [70]. Accordingly, **105 ancestral iRNAs** were identified. An iRNA was categorized

as originating from coding regions if its genomic locus overlapped with the mRNAs previously detected in the same wheat datasets [56]. A total of **190 iRNAs** originated from precursors overlapping with expressed **coding transcripts** on either chromosome strand. An iRNA was TE-derived if it originated from transposable elements or long terminal repeat (LTR) regions of the wheat genome, as annotated by IWGSC [5] (see M&M section 4.4 for details). A total of **45 iRNAs** were classified as **TE-derived**.

We further validated our predictions by multiple strategies and classified them into conserved and novel iRNAs. We also noted discrepancies in annotation, with some predicted miRNAs identified in databases as siRNAs and vice versa. **First**, to identify miRNAs and siRNAs conserved in wheat and other plant species, we compared the mature sequences of our predictions with miRBase [52], allowing two mismatches using miRFinder [84]. Of the 339 iRNAs, 95 matched entries in miRBase (46 from wheat *T. aestivum*, 42 from other monocots, and 7 from additional plants). By further incorporating matches with conserved wheat miRNAs from literature, including IWGSC-annotated miRNAs [5] and our previous studies [1, 83], we identified 144 iRNAs as conserved wheat miRNAs. However, some conserved wheat miRNAs were classified as siRNAs in our analysis. **Second**, we cross-referenced our predictions with other phase-score-based annotations, including the PSU plant small RNA catalog (a collection based on ShortStack) [65], and analyzed our 80 wheat libraries using ShortStack [7], allowing two mismatches. Among the 339 sequences, 203 iRNAs were detected, but the classifications of some miRNAs and siRNAs were swapped. **Third**, by comparing precursor sequences with the NCBI Nucleotide database [13], we found that 58 iRNAs originated from known miRNA precursors, while **33 originated from recognized TAS** precursors. Finally, there were **96 novel iRNAs** that had not been reported elsewhere, comprising 42 miRNAs and 55 siRNAs. Among the 339 iRNAs, **90 dual-identity iRNAs** were identified, with each mature sequence classified as both a miRNA and a siRNA according to any combination of the referenced databases. See Figure 2 for the summary.

2.2 Discovery of Dual-Identity Small RNAs

A key discovery of our dual-predictor approach was the identification of **90 dual-identity iRNAs**, which exhibited characteristics of both miRNAs and siRNAs. They were predicted or annotated inconsistently across databases, likely due to their dual characteristics as both miRNAs and siRNAs. These molecules may undergo flexible processing pathways, suggesting a potential evolutionary link between miRNA and siRNA biogenesis. Among them, three iRNAs were independently predicted by both our miRNA and siRNA classifiers. Cross-referencing with external databases (Figure 4) revealed that 80 predicted miRNAs had been previously classified as siRNAs, while seven siRNAs in our dataset were previously annotated as miRNAs. Transposable elements are known for their potential to generate small RNAs with both miRNA and siRNA characteristics through dual-processing mechanisms in plants [79]. However, in our genomic context analysis, dual-identity iRNAs were **not** predominantly associated with transposable element loci ($15.6\% \approx 13.5\%$, Table 1), suggesting that additional regulatory mechanisms influence their biogenesis. Instead, these dual-identity iRNAs were overrepresented in ancestral loci, indicating their evolutionary conservation and possible functional importance ($40.0\% > 31.0\%$, Table 1).

2.3 Novel iRNA Discovery

In this study, we identified **96 novel iRNAs** (Figure 2, Table S19). These iRNAs have no prior annotations in public databases or predictive models. To better understand these novel iRNAs, we examined the enrichment of their genomic origins and their potential functional roles (Table 1).

Ancestral iRNAs, which make up 31.0% of all iRNAs, were enriched among conserved miRNAs (36.2%) but were underrepresented in novel siRNAs (17.7%). This suggests that miRNAs are more closely tied to ancient regulatory functions, while novel siRNAs may be more dynamically influenced by recent genomic changes and species-specific adaptations. Coding regions, which account for 56.1% of all iRNAs, contributed to 60.9% of known iRNAs but were less frequent in novel iRNAs (43.8%). This suggests that novel iRNAs originated from a broader range of genomic regions. Notably, 37 novel iRNAs were identified as cultivar-specific (Table S16), suggesting that certain newly emerged novel iRNAs may contribute to genotype-dependent regulatory networks.

Among the novel iRNAs, 16 were identified as cold-response (Table S19), with **14 classified as high-confidence cold-responsive iRNAs** due to their targeting of known acclimation-regulated mRNAs previously detected in the same wheat datasets [56]. These high-confidence iRNAs are more likely to have functional relevance in cold adaptation, as their predicted targets are directly involved in stress-responsive pathways (Table 3). These newly identified iRNAs may mediate the cold response by regulating mRNAs involved in signal transduction, structural integrity, metabolism, and gene expression, thereby influencing kinase signaling, cell wall remodeling, osmoprotection, protein stabilization, and transcriptional regulation. Several iRNAs, such as S-029 and S-348, targeted receptor-like kinases that mediate stress perception and activate downstream cold-response pathways. Others, including M-025 and M-313, regulated proteins involved in cell wall stabilization, enhancing freezing tolerance by reinforcing structural integrity. Metabolic adjustments could be facilitated by iRNAs like M-169, which regulated fructan metabolism for osmoprotection, and M-266, which targeted heat shock proteins to prevent cold-induced protein misfolding. Transcriptional control could be influenced by M-239 and M-320, which regulated stress-responsive transcription factors and serine/threonine kinases involved in gene activation under cold conditions.

The identification and confirmation of 96 new iRNAs represent a significant expansion in the catalog of small RNAs in wheat. By distinguishing between ancestral and newly emerged iRNAs, the study revealed that novel iRNAs might arise from a wider array of genomic regions. Novel iRNAs could emerge in a lineage-specific manner, potentially influencing phenotypic variation among wheat cultivars. Moreover, the identification of high-confidence cold-responsive iRNAs underscores their functional relevance in stress adaptation. These findings highlight the dynamic evolution and potential regulatory roles of novel iRNAs, uncovering a range of previously overlooked candidates for probing stress adaptation mechanisms and serving as molecular markers in the development of cold-resistant wheat cultivars.

2.4 Identification of Wheat siRNA-Producing Genes

We explored precursor identity as an alternative validation approach for our predicted siRNAs. Given the short length of siRNAs, their precursors provide a more stable reference for homology-based identification. Using BLASTN searches against the Nucleotide database [13], we identified **eight** precursors with significant sequence similarity to known **TAS genes** (Table 6). Among them, six were homologous

to wheat *TAS3b* genes, while the other two resembled *TAS3a* genes from different plant species, suggesting that they represent **novel TAS genes in wheat**. From these precursors, we determined that **33** predicted siRNAs were **tasiRNAs**. By homology searches, a straightforward and effective method, we validated siRNA precursors and expanded the known tasiRNA repertoire in wheat. However, similarity searches did not detect other potential siRNA-producing genes, such as the PHAS and hc-siRNA genes. This absence may result from insufficient siRNA annotations for these subtypes in database entries or the genetic specificity of certain tissues or species.

Difficulties in Detecting TAS Using Phase Score Method. The phase score evaluates siRNA arrangement along a precursor by integrating contributions from the primary effector siRNA and the sibling siRNAs generated through multiple Dicer digestion cycles [40]. To assess whether the eight TAS genes exhibited phasing characteristics, we compared their phasing scores with the actual sRNA alignment patterns along their genomic contigs (Figure S17). Our analysis showed that the 21-nt sRNA species maintained a consistent 21-nt interval, with most TAS precursors undergoing at least seven phasing cycles, which indicates strong phasing signals.

High levels of effector siRNA expression indicate strong phasing activity; however, weak or absent sibling siRNAs may underestimate phasing signals and potentially exclude true siRNA loci, even if they share homology with known TAS genes. For instance, precursor 2A__598975000 exhibited seven phasing cycles, but its siRNA frequency peaked at merely 140 reads, with some sibling phasing positions supported by as few as five reads. This limitation highlights the inherent challenge of detecting siRNA loci using the phase score method. Enhancing computational strategies and increasing sequencing depth may mitigate these challenges and improve detection accuracy.

miRNA Trigger for siRNA Biogenesis? miRNAs, especially miR390, have been hypothesized to be essential triggers for siRNA biogenesis through targeted precursor cleavage [9]. However, it is still uncertain whether miRNA cleavage in wheat is an absolute requirement. To explore this, we employed psRNATarget [23] to predict cleavage sites on the validated TAS precursors. Our analysis identified two miR390 cleavage sites aligning with phasing borders in six of the eight wheat *TAS3a* and *TAS3b* precursors. Despite an exhaustive search, two TAS precursors contained only a single miR390 cleavage site, implying that tasiRNA production may not always depend on this miRNA trigger.

Several factors could explain the absence of detectable cleavage sites. Genetic variation between the Canadian wheat cultivars analyzed in this study and the *Chinese Spring* reference genome [5] may have altered miRNA recognition motifs, obscuring expected cleavage patterns. Additionally, some tasiRNA precursors may initiate biogenesis through alternative pathways, relying on a single cleavage or miRNAs other than miR390 [25, 33, 37]. These observations suggest that, while miRNA cleavage is a well-established trigger for tasiRNA production, it may not be the only mechanism, allowing for alternative pathways contributing to tasiRNA biogenesis.

2.5 Evolutionary Landscape of Wheat iRNAs

The evolutionary dynamics of small RNAs are shaped by their genomic origins and selective pressures. By analyzing chromosomal distribution, ancestral inheritance, and genomic context, we uncovered patterns that revealed how miRNAs, siRNAs, and dual-identity iRNAs emerged from distinct genomic landscapes

under different constraints and followed unique evolutionary trajectories in wheat.

The chromosomal distribution of iRNA precursors in the wheat genome highlights key differences between miRNAs and siRNAs. While miRNAs were broadly distributed across all seven chromosomes and conserved across the three progenitor genomes (A, B, and D), siRNAs tended to accumulate in specific genomic regions and were notably absent in others (Figure 3). This disparity suggests that miRNAs undergo strong evolutionary conservation, whereas siRNAs emerge from specific genomic hotspots. The spatial organization of siRNAs might reflect functional compartmentalization, where siRNAs were preferentially associated with specific genomic regions. The absence of siRNAs in some regions suggests that these areas lack the necessary triggers for production.

The genomic origins of iRNAs further illuminate their evolutionary divergence. Among the 339 identified iRNAs, **105** originated from ancestral genomes, **190** from coding regions, and **45** from transposable elements (Figure 2; Table S16, sheet 1 *all*, see columns H: *isAncestral*, AL: *XLOCprecursor*, BI: *isTEorLTR*). These distinct origins suggest multiple evolutionary pathways, where some iRNAs are conserved while others evolve novel functions. To investigate these evolutionary trajectories, we conducted a statistical enrichment analysis to examine the distribution of iRNA categories across ancestral genomes, coding regions, and transposable elements (Table 1). Our findings, summarized below, highlight distinct mechanisms shaping the biogenesis of miRNAs, siRNAs, and dual-identity iRNAs in wheat.

- **Ancestral iRNAs** (backgound = 31.0%) were significantly enriched in miRNAs (37.2%), known iRNAs (36.2%), and dual-identity iRNAs (40.0%). In contrast, ancestral iRNAs were underrepresented in siRNAs (20.5%) and novel iRNAs (17.7%). This distribution suggests that miRNAs and dual-identity iRNAs are more evolutionarily conserved for their regulatory functions, whereas siRNAs and novel iRNAs may have emerged more recently and are subjected to stronger selective pressures and rapid turnover.

- **Coding region-derived iRNAs** (background = 56.1%) were depleted by miRNAs (50.2%), while significantly enriched by known iRNAs (60.9%), siRNAs (65.4%), and tasiRNAs (78.8%). This observation suggests that siRNAs (and their subtypes) are more likely to originate from protein-coding genes and are directly associated with gene expression, whereas miRNAs emerging from intergenic regions exhibit greater stability.

- **TE (transposable element)-derived iRNAs** (background = 13.3%) were significantly overrepresented in miRNAs (18.6%) and depleted in siRNAs (3.9%). The low abundance of siRNAs originating from transposable elements implies that these loci have shorter lifespans, likely due to rapid purging or their transient role in gene regulation. Over time, some TE-derived sequences may undergo structural stabilization and acquire functional relevance, leading to their retention as miRNAs with more conserved and stable biogenesis pathways.

2.6 Identification of cold response iRNAs

With 339 iRNAs predicted from wheat cultivars varying in freezing tolerance, grown from late summer to freezing, the dataset provided a unique opportunity to investigate small RNA contributions to cold

adaptation. By analyzing temporal expression dynamics, we identified miRNAs and siRNAs involved in the cold response through a series of analytical procedures. Temporal expression profiles were assessed across five time points (T1–T5), with pairwise comparisons between T1 and subsequent time points. We examined genetic and phenotypic differences between winter and spring cultivars using *encoded vector clustering* (see M&M section 4.7 for details). This method aimed to identify iRNAs regulated **coordinately** up- or down-regulated between spring and winter cultivars, isolate cold-response iRNAs from confounding environmental factors, and distinguish direct cold responses from *Vrn-A1*-dependent growth effects. In contrast, **non-coordinated** iRNAs, even if differentially expressed, were classified as constitutive iRNAs (Ctsv) and not considered cold-response. We identified **66 cold-response (CR) iRNAs** from 32 encoded profile clusters among 204 differentially expressed (DE) iRNAs (Figure 2, Table S8 sheet 4, M&M section 4.7). The remaining 138 DE iRNAs and 135 non-DE iRNAs were classified as constitutive iRNAs, indicating they do not respond to cold stress. CR iRNAs were annotated as up-regulated (CRup), down-regulated (CRdw), or mixed (CRupCRdw) based on winter-spring coordination. Many iRNAs showed significant differential expression over time (Figure 5), including several cold-responsive iRNAs from established families, along with others that denote new regulatory elements crucial for cold adaptation.

This set of 66 cold-response iRNAs includes **58 CR miRNAs** and **9 CR siRNAs**, with one exhibiting dual identity. Notably, 25 of these iRNAs belong to 19 conserved miRBase families (Table 2), while 11 are conserved within wheat but are not cataloged in miRBase. Additionally, 13 were identified by ShortStack, 1 was classified as a tasiRNA, and 16 were determined to be novel (Table S19). Notably, 14 of the 16 novel cold-response iRNAs regulated validated acclimation-related mRNA targets (Table 3). We found that cold-response iRNAs were not significantly enriched in ancestral genomes (33.3% $\approx$ 31.0%, Table 1), suggesting that recent selective pressures have shaped their role in cold adaptation. This implies that wheat's small RNA-mediated cold response is primarily driven by newly evolved genomic elements rather than ancient regulatory pathways.

2.7 Functions of Cold-Response iRNAs in Acclimation

Many conserved cold-response miRNA families have been well-documented for their roles in helping plants adapt to cold stress (Table 2). To further expand this understanding and uncover additional iRNAs involved in cold acclimation, we integrated previously characterized mRNA datasets from the same wheat samples [56]. The functional roles of the 66 cold-response miRNAs and siRNAs in cellular components, molecular functions, and biological pathways can be inferred through comprehensive functional analysis [20]. This analysis involved predicting target genes, annotating them with Gene Ontology (GO) terms and KEGG pathways using UniProt IDs, and performing statistical tests (see M&M for details). This approach allowed us to classify several novel cold-responsive iRNAs as high-confidence candidates, including three siRNAs (S-029, S-348, and S-454) and several miRNAs not previously linked to cold response (Table 3). These siRNAs specifically targeted acclimation-related mRNAs identified by Li et al. [56], providing further support for their regulatory roles in cold stress adaptation.

Building on these findings, we examined the broader functional contributions of cold-response iRNAs to wheat acclimation. These regulatory molecules participated in multiple biological processes essential

for freezing tolerance, including kinase signaling, transcriptional control, stress adaptation, metabolic homeostasis, and hormone signaling. The following functional categories are summarized based on Figure 7, Table 2, and Table 3.

Kinase Signaling and Stress Perception. Kinase-mediated signaling acts as molecular sensors that trigger downstream defense responses upon detecting environmental stress [18]. We found that cold-response iRNAs collectively targeted a substantial number of kinases (63 unique UniProt IDs) and transcription factors (90 unique UniProt IDs) (Table S19, UniProt IDs in Table S16). While some kinase targets were putative, others had well-documented roles in plant temperature responses. Notable examples included Casein kinase 1 [81], Wall-associated receptor kinase [100], Coatomer [117], NB-ARC domain kinase [118], Leucine-rich repeat receptor-like protein kinase (LRR-RLK) [108], and GsSRK kinase [86]. To explore the regulatory interactions between CR iRNAs and their kinase targets, we analyzed their expression patterns, identifying pairs with strong and statistically significant correlations (Table 4). Selected iRNA-kinase expression patterns across cultivars and time points were illustrated in Figure 8. These correlations showed that certain iRNAs inhibited kinase expression, whereas others might engage in feedback loops that increase their target's expression. iRNAs regulated these kinases, showing a precise and dynamic response to cold stress that helped plants effectively activate protective mechanisms.

Transcriptional Regulation. Transcription factors (TFs) and DNA-binding proteins coordinate the changes in gene expression necessary for cold acclimation by regulating stress-responsive pathways and developmental transitions. Our analysis showed that many cold-response iRNAs specifically targeted TF families that have been well characterized in cold adaptation, including SQUAMOSA promoter-binding proteins (SBP) [112], Calmodulin-binding transcription activator (CAMTA) [98], NAC domain proteins [67], GATA-type transcription factors [114], and MBD domain-containing proteins [42]. These TFs contributed to diverse regulatory networks, modulating plant development, stress-responsive gene expression, and hormonal signaling. Interestingly, transcription factors were preferentially regulated by miRNAs rather than siRNAs, suggesting that miRNAs play a specialized role in tuning transcriptional responses to environmental cues. Cold-response iRNAs also influenced post-transcriptional regulatory mechanisms. Some iRNAs target genes involved in mRNA splicing, RNA modification, and DNA replication. These functions helped to maintain efficient gene expression and genome stability, which are critical for sustaining cellular functions under stress. To further investigate these regulatory interactions, we identified miRNA and its target that show strong and significant expression correlations (Table 5). The temporal expression patterns revealed how cold-responsive iRNAs dynamically adjusted their gene expression programs in response to low temperatures across cold-acclimation time points (Figure 9).

Homeostasis and Stress Adaptation. Maintaining cellular homeostasis is essential for cold acclimation. Our analysis showed that cold-response iRNAs regulated genes involved in heat shock proteins, fatty acid β-oxidation, fructan metabolism, and ubiquitin-mediated protein degradation. These pathways helped stabilize cellular energy levels and ensured the efficient removal of damaged or misfolded proteins. In addition, hormonal signaling modulated stress responses and growth under unfavorable conditions. Cold-response iRNAs influenced hormonal signaling, including abscisic acid-mediated signaling, auxin-dependent growth regulation, and cytokinin-controlled cell division. By regulating homeostasis and hormonal signaling, iRNAs enabled wheat to maintain cellular integrity while balancing survival strategies

with continued development in response to cold stress.

2.7.1 Expression Clustering of cold-response miRNAs and siRNAs

Through functional enrichment analysis (Figure 7), we found that these 66 CR iRNAs exhibit distinct functional associations. *CRup* iRNAs were primarily enriched in mRNA splicing, plant organ development, RNA modification, and response to salt stress, while *CRdw* iRNAs were linked to DNA replication, cytokinin-activated signaling, fatty acid beta-oxidation, and the negative regulation of hypoxia-induced intrinsic apoptotic signaling pathways.

To investigate the dynamic interaction between miRNAs and siRNAs, we applied hierarchical clustering on paired winter-spring expression profiles. This process organized CR iRNAs into eight temporal clusters (C1–C8) according to their similar expression patterns, using unsupervised learning in DEGreport [77] (Figure 6). Each cluster displayed a unique temporal pattern of regulation, suggesting a chronological activation of stress-response genes, outlined as follows.

- C1 (11 miRNAs, 4 siRNAs): Fatty acid beta-oxidation, DNA replication.
- C2 (10 miRNAs): Ribosomal small subunit assembly, cell wall biogenesis, response to abscisic acid.
- C3 (8 miRNAs, 2 siRNAs): Photoreactive repair, UV protection, syncytium formation, glucose import.
- C4 (25 miRNAs, 2 siRNAs): Response to auxin, long-chain fatty-acyl-CoA metabolic process, proton export across the plasma membrane.
- C5–C8 (4 miRNAs, 1 siRNAs): Transmembrane transport, RNA modification, auxin-activated signaling pathways, and protein ubiquitination.

Notably, miRNAs and siRNAs were frequently co-expressed within the same clusters (Table S16, sheet *Cold response*, see column *Cluster*), suggesting that cold-response siRNAs functioned in tandem with miRNAs in modulating key stress regulatory pathways.

2.8 iRNAs in *Vrn-A1*-Mediated Vernalization

The *Vrn-A1* gene regulates flowering time in wheat by determining whether a cultivar requires prolonged cold exposure (vernalization) to transition from vegetative to reproductive growth. Norstar (*vrn-A1*) is a winter wheat cultivar with exceptional cold tolerance, while Manitou (*Vrn-A1*) is a spring wheat suited to shorter growing seasons. To dissect the genetic control of vernalization and wheat adaptation, we analyzed these cultivars along with their isogenic derivatives, Iso8S (spring-habit Norstar) and Iso11W (winter-habit Manitou), which differ only in their *Vrn-A1* genotype. Iso8S (*Vrn-A1*) converts Norstar into spring wheat, whereas Iso11W (*vrn-A1*) makes Manitou behave as a winter wheat. These isogenic lines provide an ideal system for investigating how iRNAs regulate vernalization, cold response, and flowering in wheat.

Among the 204 differentially expressed (DE) iRNAs, 121 exhibited cultivar-specific differential expression. In particular, we identified 11 Iso11W-specific, 23 Iso8S-specific, 27 Manitou-specific, 18 Norstar-specific, 15 spring-habit specific (Manitou and Iso8S), 5 winter-habit specific (Norstar and Iso11W), 6 Manitou-background specific (Manitou and Iso11W), and 16 Norstar-background specific iRNAs (Norstar and Iso8S) (Table S16, several sheets).

2.8.1 Differential iRNA Networks Under *Vrn-A1* for Growth Strategies

Winter-habit cultivars (Norstar and Iso11W) exhibited distinct iRNA expression profiles associated with cold survival, dormancy, and vernalization-controlled flowering. We found that cold-response miR9674 and miR5062 members were specifically differentially regulated in these cultivars (Table 7). Functional enrichment analysis (Figure SF3) showed that these iRNAs modulated meiosis, histone post-translational modifications (methylation, phosphorylation, acetylation), flowering processes (microgametogenesis, megagametogenesis), negative regulation of embryonic development, gene silencing, and vacuole-related transport.

Conversely, **spring-habit cultivars** (Manitou and Iso8S) expressed different sets of iRNAs, including miR166, miR167, miR319, miR393, miR396, and miR9778 (Table 7, functional enrichment in Figure SF3). Some miRNA families were known for their cold response (Table 2). These iRNAs were enriched in pathways related to plant development, gibberellic acid homeostasis, chromatin assembly, DNA replication, and responses to oxidative stress. The involvement of these iRNAs in bilateral symmetry, meristem growth, and hormone metabolism suggests their role in rapid development and cold-independent flowering.

In conclusion, our findings suggest that *Vrn-A1*-dependent iRNA expression patterns contribute to the distinct developmental strategies of winter and spring habit wheat, emphasizing the importance of iRNA networks in shaping wheat adaptation to seasonal growth conditions.

2.8.2 iRNA Isoforms in Cultivar Adaptation

Cultivar specificity can sometimes arise from the differential expression of **isoforms** within a single miRNA or siRNA family. For example, among the 66 **cultivar-specific miRNAs** (including one dual-identity miRNA), the **miR159** family exhibited distinct isoform expression patterns that varied between cultivars. Similarly, within the 56 **cultivar-specific siRNAs**, members of the **TAS families** predominantly expressed different siRNA siblings derived from the same TAS precursors across cultivars (Table 8). These results showed that small RNA isoforms were not produced randomly but were tightly regulated by the cultivar's genetic background. This genetically programmed regulation probably affected small RNA-mediated pathways, including those governing vernalization (*Vrn-A1*) and cold adaptation, thereby shaping cultivar-specific responses and promoting adaptation.

3 Discussion

In this study, we used freezing-tolerant wheat cultivars as a case study to demonstrate the power of an integrated framework for small RNA research. By jointly analyzing different classes of small RNAs, we

identified 339 iRNAs (215 miRNAs and 127 siRNAs), including 96 novel iRNAs, 33 tasiRNAs, 90 dual-identity iRNAs, and 66 cold-response iRNAs. These findings significantly expand the known repertoire of small RNAs in wheat.

We uncovered key regulatory elements in cold stress adaptation by integrating miRNA and siRNA prediction within a unified framework. The presence of 90 dual-identity iRNAs suggests an evolutionary and functional link between miRNAs and siRNAs, blurring the boundaries between these two classes. Identifying siRNA-generating genes in wheat further expands our understanding of the siRNA landscape. Notably, the discovery of TAS precursors lacking miR390 cleavage sites suggests that tasiRNA production may occur through non-canonical mechanisms, such as alternative miRNA triggers. Additionally, 66 cold-response iRNAs were identified, many targeting key regulatory genes involved in cold adaptation.

To ensure the reliability of our predictions, we employed multiple validation strategies, including cross-referencing with established databases such as miRBase, ShortStack predictions, literature-reported wheat miRNAs, and phase-score-based annotations. Furthermore, their precursors were validated through BLASTN analysis, which indicated the presence of TAS homologs and further substantiated their classification as functional small RNAs. By integrating multiple validation strategies, we boosted confidence in small RNA identification and functional annotation, demonstrating the robustness of our bioinformatics approach.

The genomic mapping of iRNAs to ancestral loci, coding regions, and transposable elements indicates that small RNAs have evolved through various pathways, with selective pressures shaping their functional diversification. miRNAs and dual-identity iRNAs are more evolutionarily conserved, while siRNAs and novel iRNAs tend to emerge from recent genomic adaptations. The discovery of dual-identity iRNAs suggests an intermediate state in the evolutionary continuum between miRNAs and siRNAs. Cold-response iRNAs, in particular, seem to be a recent evolutionary adaptation rather than a product of conserved regulatory mechanisms. This finding further supports the hypothesis that small RNA-mediated regulation is an adaptive mechanism that enhances stress resilience in wheat.

Our analysis further revealed distinct roles of cold-response iRNAs in stress adaptation. These iRNAs target key regulatory genes, including kinases, transcription factors, and metabolic enzymes. The clustering of cold-response iRNAs (C1–C8) revealed distinct expression patterns, suggesting multiple regulatory layers in wheat's response to cold stress. Some clusters were enriched in auxin-related genes, while others were associated with metabolic adjustments and transcriptional regulation. The temporal co-expression of certain miRNAs and siRNAs within these clusters suggests a coordinated regulatory mechanism that orchestrates stress responses during the seasonal transition from warm to cold at different developmental and environmental stages. Moreover, cultivar-specific iRNAs highlight the genetic variability in wheat's response to environmental stimuli. Adjusting cultivar-specific cold-response iRNAs may subsequently regulate flowering time and stress responses, offering a potential avenue for targeted breeding and genetic modifications to develop wheat varieties optimized for diverse environmental conditions.

Despite these advances, several challenges remain in identifying promising small RNA markers. The complexity and large size of the wheat genome pose challenges for small RNA prediction and classification. Computational predictors still struggle to accurately distinguish siRNA subtypes because of the absence of a dedicated reference database. Additionally, curation efforts and annotation criteria for siRNAs remain

insufficient, leading to fragmented knowledge and limiting comparative analyses. Resource constraints further impact the development of bioinformatics tools for small RNA studies in wheat and other complex plant genomes.

Nevertheless, our study demonstrates the feasibility and importance of an integrated approach to small RNA analysis. Addressing current limitations by standardizing prediction criteria, curating small RNA classes and their precursors, and expanding the capacity of bioinformatics resources will enhance the reliability and applicability of future research. Ultimately, these efforts will support bioinformatics-driven breeding strategies, providing opportunities to enhance crop resilience in response to environmental stressors.

4 Materials and methods

4.1 Small RNA preparation from wheat field samples

Four wheat cultivars that differ in their freezing tolerance and vernalization responses – winter-habit 'Norstar' (vrn-A1/vrn-A1), spring-habit 'Manitou' (Vrn-A1/Vrn-A1) and their reciprocal near isogenic lines (NIL), spring-habit Norstar 'Iso8S' (Vrn-A1/Vrn-A1) and winter-habit Manitou 'Iso11W' (vrn-A1/vrn-A1) – were grown under field conditions in two different years and were sampled from autumn to just before the freezing season. All four wheat cultivars were planted in late summer. Winter-habit wheat would undergo vernalization, a cold promoted flowering process, and spring-habit wheat would eventually die from cold over winter Li et al. [56]. Air and soil temperature data are presented in the supplemental data section (See S1). Crown tissues were sampled and at least five crowns were used for each RNA extraction. Samples were grounded to a powder in dry ice and total RNA was extracted using the mirVana™ miRNA Isolation Kit (Life Technologies). RNA quantity and quality were analyzed using Nanodrop spectrophotometer and Agilent 2100 bioanalyzer system, respectively. Small RNA libraries were constructed using the TruSeq Small RNA Sample Preparation Kit (Illumina) and the sequencing was performed on Illumina HiSeq 2000 Platform. Each cultivar had 20 libraries, and a total of 80 sequencing libraries were constructed, representing sRNAs of two biological replicates at five time points between the beginnings of September and November from each of the four cultivars grown in two years, 2010 and 2013 in Saskatoon (see Li et al. [56] for details of cultivar genotypes, field conditions and sample collection).

4.2 Bioinformatics approach for the identification of miRNAs and siRNAs in wheat cold response

We investigated wheat cold acclimation in the field by integrating several bioinformatic and statistical methods in the analysis workflow (Figure 1). We will describe the workflow briefly here as well as in detail for each incorporated method in the next sections. Sequencing data were generated by our collaborators. We performed quality control, prediction of interfering miRNAs and siRNAs from expressed sequences, differential expression, differential regulation and functional analysis among time series and cultivars. We incorporated the methods that we developed, including miRNA and siRNA predictions, removing biases, identification of cold response iRNAs, and we annotated the genes expressed in the same libraries for target gene analysis. Using two de novo prediction pipelines, 215 miRNAs and 127 siRNAs from the

wheat field libraries were reported in this study. We used a large dataset consisting of 80 libraries which enabled us to predict iRNA candidates and eliminate noise and batch effects. To predict miRNA, we used miRLibSpark that embedded the gold-standard requirement for plant miRNA prediction. To predict a siRNA, we incorporated the H-score [40] to search sustained phasing potentials in the genomic landscape. To understand the functions of the predicted miRNA and siRNA from wheat field samples, we studied their expressions. We devised an encoding scheme to translate differential expression profiles of spring and winter cultivars into cold response regulation. Differential regulation analysis among time series and cultivars showed 66 cold response iRNAs.

4.3 Identification of miRNA and siRNA candidates from sequenced small RNAs

These analyses were performed on the 80 sequencing libraries. Each library was checked for quality using FastQC [4]. The adapters were removed using cutadapt 0.9 [71]. Afterwards, putative iRNAs were predicted using miRNA prediction and siRNA prediction procedures illustrated in Figure 1 and described in the following paragraphs.

4.3.1 MiRNA prediction

MiRNA precursors are single-stranded RNA molecules expressed from the genome and exhibiting several structural features. They are cleaved by Dicer on the double-stranded hairpin into miRNA/miRNA* duplexes. The consensus determining features of a valid miRNA are specified recently [73, 8]. However, limited choices of the existing miRNA predictors [7, 55] are capable of predicting novel and conserved miRNAs derived from organisms of big genomes using reasonable time and computing resources, such as from wheat having 13 Giga base pairs of the genome. We designed and implemented mirLibSpark [104], a scalable NGS plant miRNA prediction pipeline. Its miRNA prediction module compiles with the gold-standard requirement on plant miRNA predictions [73, 8]. The computational complexity has been optimized in mirLibSpark so that it takes only 2 to 3 hours to predict miRNAs from an average wheat sRNA library. The operations in mirLibSpark for this study are as follows. Low complexity sRNAs were removed using dustmasker [75]. Reads with size between 21 and 24 nts [8] were retained and mapped to the reference genome using Bowtie allowing zero mismatch and up to 15 mapped genomic loci. A set of known non-miRNA sequences (tRNA, rRNA, snoRNA) according to the annotation of the genome were removed. Small RNAs with at least 100 reads were chosen for folding predictions. For each sRNA, we considered two sequences that could include a potential pre-miRNA candidate as follows: 200 nts before the start and 500 nts after the end of the mapping; 500 nts before the start and 200 nts after the end of the mapping on the genome. The secondary structures of the extracted sequences (pri-miRNA, about 700 nts) were folded with RNAfold from the ViennaRNA v2.4.11 package [61] to identify those having a hairpin-like shape, one of the fundamental characteristics of pre-miRNAs, using miRCheck [43] and modified parameters complying with Meyers et al. [73]. Then, predicted pre-miRNA sequences were excised from the pri-miRNAs and resubmitted to RNAfold with an optimized hairpin and repredicted again using miRCheck. Precursors longer than 300 nts or containing a second loop that was larger than 5 nts were discarded [8]. Finally, a dominantly expressed potential miRNA within the pre-miRNA candidate sequence was considered a true miRNA [8]. This setting showed an improved accuracy for miRBase Arabidopsis

annotations (87%) [52, 104] and could identify the majority of the IWGSC-annotated miRNAs (52.5%) [5, 83]. IWGSC stands for the International Wheat Genome Sequencing Consortium.

4.3.2 SiRNA prediction

Endogenous siRNAs are the major component of sRNA populations in plants [60]. SiRNAs are generated from Dicer's cutting of dsRNA precursors and the dsRNA region can come from either a single RNA forming hairpin, also known as a miRNA precursor or a dsRNA duplex, also known as a siRNA precursor [40, 16, 46]. Since Dicer dices along the dsRNA substrate one cycle after another until riding off the substrate, one to several cycles of phased siRNAs can be generated and a cluster of siRNAs is thus observed as regularly-spaced 21-nt or 24-nt duplex populations aligned on the precursor genome [40, 17]. The following characteristics [7] of a valid siRNA precursor have been included in our analyses: expression of the precursor transcript; specific sizes of siRNAs originating from the precursor; strength of the phasing score from the precursor. We adopted a generalized phase score Howell et al. [40] to calculate the potential of a selected-size sRNA (21 nts or 24 nts) being a siRNA at every sRNA-aligned position, considering both sRNA abundance and number of phased positions on the precursor. The phase score H-score Howell et al. [40] to quantify the phasing potential of a given genomic position based on sRNA expressions was defined as follows. Given l, $\mathbf{G}$ and $\mathbf{L}$. l the PHAS size, $l \in \{18, 19, ..., 30\} \in \mathbb{N}^+$. $\mathbf{G}$ the genome, $\mathbf{G} = \{\text{A, C, T, G}\}^r$, and $r \in \{10^6, \ldots, 10^{10}\} \in \mathbb{N}^+$. $\mathbf{L}$ the ensemble of S the sRNA sequences in a library, $\mathbf{L} = \{S_1, S_2, \ldots, S_n\}$, where $S = (\alpha, q)$, and $\alpha = \{\text{A, C, T, G}\}^r$, $r < 50 \in \mathbb{N}^+$, $q \in \mathbb{N}^+$. Phase Score P_j for locus j on $\mathbf{G}$:

$$P_j = ln\left(1 + \sum_{i=1}^{8} K_{j+(i-1)\times l}\right)^{n-2}, P \geq 0$$

where $K_a \in \{\mathbb{N}^+, 0\}$ is the sum of the frequency q of any $S \in \mathbf{L}$ that is aligned to locus a on positive strand or locus $(a-2)$ on negative strand, and $S.r = l$; n the number of $K > 0$, $n \leq 8 \in \mathbb{N}^+$. Because canonical miRNA precursors, by definition, host only one pair of miRNA duplex, we set the parameters in H-score to predict siRNA precursors that at least 3 cycles of Dicer cuts are observed within 8 cycles. We thoroughly surveyed the architecture of the genome by developing a chromosome-walk algorithm. The operations of siRNA prediction procedure for this study are as follows. Low complexity sRNAs were removed using dustmasker [75]. Reads were retained if they mapped to the reference genome using Bowtie [54] allowing zero mismatch and up to 15 mapped genomic loci. After alignment, each sRNA was assigned to a CONTIG. The genomic CONTIGs were indexed from the whole genome, each spanning 5250 nts with 250 nts overlapping the CONTIGs before and after itself. For example, 6D__285740000_285745250 (or 6D__285740000 for short) is a CONTIG on chromosome 6D between positions 285740000 and 285745250. A CONTIG was retained if at least 3 sRNA species and at least 100 of total reads were aligned on the CONTIG to verify if the precursor was expressed, and at least one position had a phasing potential H-score greater than zero for the selected size of siRNAs. A putative precursor sequence was excised from the positive strand of the double-stranded CONTIG between the first and last Dicer cuts with additional flanking nucleotides. Random degradation may appear and such precursor candidates and the siRNAs derived from them were detected and excluded manually. Finally, all the sRNAs that had a phasing potential on the predicted precursors were considered functional siRNAs.

4.4 Sequence analyses for precursor and mature iRNAs

Each precursor is part of a CONTIG in our genomic indexing system based on genomic coordinates. The mature iRNAs were clustered together if they came from the same precursor CONTIG. Nucleotide sequence similarity analysis for the precursors was performed using bowtie [54] against the mRNA transcripts collected from the same field samples for sRNA analysis, designated XLOC database [56], as well as using BLASTN against the NCBI Nucleotide database with default parameters [13]. The annotations of the hits are the potential functions of precursors.

MiRBase family analysis for the mature sequences was performed using the miRFinder tool on Wheat MicroRNA Portal (with parameters as follows: Min sequence length 18, Min identity percentage 80, Min query coverage 0.9, Min subject coverage 0.9, Max of gaps 0) [84] against miRBase v22 database [52]. Maximum of 2 mismatches were allowed. Hits were retained if they matched with the miRNAs of the 6 monocot species wheat *Triticum aestivum* or plants *Oryza sativa*, *Hordeum vulgare*, *Arabidopsis thaliana*, *Brachypodium distachyon* and *Zea mays*. A local version of microRNA Finder [84] was used to search against additional wheat databases that were not available on the portal, such as literature collections [1, 83], IWGSC-annotated miRNAs [5], collections based on ShortStack (PSU catalogue [65]) and the predictions using ShortStack [7] on our 80 wheat libraries. Wheat is a hexaploid having 3 sets of genomes, denoted as A, B and D genomes. An iRNA was defined as an ancestral iRNA if it appeared in hexaploid (AABBDD) or in tetraploid (AABB) genomes because it suggests a more ancient origin [70]. Sequences originating from wheat coding regions were defined as mapping to the positive or negative strand of the expressed XLOC sequences. Sequences originating from wheat transposable elements and long terminal repeats loci were defined as mapping to the positive or negative strand of IWGSC-annotated coordinates, respectively. Considering both strands was to harmonize the siRNAs originating from both strands.

4.5 Expression statistical analyses of iRNAs

To perform differential expression analysis on the expressed iRNAs in 80 wheat libraries, we used DESeq2 [62, 39] because it provides a quantitative analysis for studies based on small replicate numbers and large dynamic ranges. It incorporates Bayes-based shrinkage estimation for dispersions and fold changes to improve the stability and interpretability of estimates. Read frequencies of the expressed iRNAs prior to normalization and their metadata (cultivar, year, time point, replicate and soil temperature) were used as input for DESeq2. It normalized read frequencies based on the input read frequency population. We considered time point 1 of each experiment (for each year and each cultivar) as the baseline, and compared time points 2, 3, 4 and 5 with time point 1 of the same experiment. The significant difference in the normalized read frequencies between time point 1 and the four remaining time points in each cultivar was computed. An adjustment for multiple comparisons based on the false discovery rate (FDR) [11] was performed and iRNAs with FDR $\leq 5\%$ were kept. To identify read frequencies that were inconsistent between the two biological replicates (r1 and r2), we used the gauge formula: [if r_1 and r_2 do not fulfill $abs(\frac{r1-r2}{2}) \leq (\frac{r1+r2}{2}) \times 20\%$, they are inconsistent], where abs is the absolute value. Most inconsistent entries were related to libraries Iso11W-2010-t3 and Iso11W-2013-t2. Finally, iRNAs with fold change (FC) lower than 0.5 or higher than 2.0 (ie. abs(log2FC $\geq$ 1)) based on consistent replicate entries were

considered differentially expressed.

4.6 Assigning regulated iRNAs in genetic backgrounds

The differentially expressed iRNAs were grouped into cultivars Iso11w, Norstar, Iso8s and Manitou. Each group contains 20 libraries of the same cultivar collected at five time points in 2010 and 2013, respectively. If the expression of a RNA species was differential in only a cultivar, it was regulated specific to that cultivar. If a RNA species was differentially expressed in both spring habit cultivars (Manitou and Iso8S) but not in either winter habit cultivars (Norstar or Iso11W), it was spring-habit specific; and the other way around was winter-habit specific [76]. Similarly, we defined a RNA species as Manitou-background specific if it was commonly differentially expressed in both Manitou-background cultivars (Manitou and Iso11W) but not in either Norstar-background cultivars (Norstar or Iso8S); and the other way around was Norstar-background specific.

4.7 Differential regulation analysis based on iRNA expression patterns during field cold acclimation

We classified each iRNA by assessing its expression profile with an encoded vector clustering approach built on the time course expression in paired winter and spring cultivars. To do so, we first compared the expression of an iRNA in a cultivar at time points 2, 3, 4 and 5 with time point 1 of each year, and noted up- (u), down- (d) and not-regulated (0) according to fold change ≥ 2, ≤ 0.5 and others, respectively. For example, the pattern for an RNA X in a winter cultivar in a year through time points 2 to 5 was noted as [00uu]. The pattern for X in the counterpart spring cultivar in the same year was noted as [uuuu]. We then paired the winter and spring cultivars in the same year to create an encoded expression profile vector, and thus X had an expression profile of [0u,0u,uu,uu]. This analysis allowed us to identify the differentially expressed RNAs as cold-responsive if they expressed at least partially coordinately between winter and spring cultivars (eg. up profile [0u,0u,uu,uu] or down profile [00,0d,dd,d0]); as unresponsive or constitutively expressed if they were not regulated at the same time points between the paired cultivars (eg. profiles [00,0u,0u,u0], [00,00,00,00]). If winter and spring cultivars have the coordinated trend for an iRNA, it shows that despite the difference in their supposed phenotypes, the iRNA behaves similarly in winter and spring cultivars. It must be responding directly to the common environmental stimuli, cold, rather than the consequential regulation of growth and development. If not coordinated, even if it is differentially expressed, it is not responding to cold, and therefore constitutive. We concluded the regulation of each RNA based on its profile and quoted up-, down-regulated and unresponsive in cold response as CRup, CRdw and Cstv, respectively. See Table S8 for profile and regulation conversions. Using DEGreport [77] to show the coordinated expression changes of cold response iRNAs, samples were grouped by cultivars for calculating the mean sequence abundance as the representation at each time point. It then performs pairwise correlations between iRNAs to hierarchically cluster them.

4.8 Functional annotations of XLOC database

The wheat mRNA transcripts – 151690 XLOCs – collected from the same field samples for sRNA analysis [56] were used to annotate precursors and interfering targets in this study. Functional annotations of the XLOCs were based on Uniprot sequence similarity [44]. The Uniprot database [21, 22] – providing high-quality sets of protein identities, cross-reference IDs and functional annotations – was accessed in October 2020. The best 20 Uniprot hits for each XLOC sequence were identified using the local blastx program [13, 19] against Uniprot's Ensembl Plants collection, which also hosts the RefSeq v1.0 assembly from the IWGSC [110, 5]. Peptide hits with more than 50% identity or e-value smaller than 0.0001 were retained [12, 19]. To benefit from annotations in other species while maintaining relevance to wheat, hits were restricted to monocots. The performance of annotation for chosen filtering parameters was assessed in Figures SF6 and SF7. Protein IDs, GO terms and KEGG pathways [20, 47] were retrieved from the best Uniprot hits. To address the gaps in annotations resulting from the Uniprot update, we also incorporated the XLOC annotations provided by Li et al. [56].

4.9 Functional enrichments of iRNAs through target genes

MiRNA and siRNA target genes were identified using the 2017 schema of psRNATarget, with a stringent Expectation threshold of 2.0 [24], against the XLOC database. psRNATarget also reports cleavage sites. psRNATarget also reports cleavage sites. We annotated the XLOC database with KEGG pathways and GO terms including GOBP, GOMF and GOCC. The overall enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the hypergeometric distribution test (similar to [97]). The wheat genome GO term (or KEGG pathway) background was constructed by taking into account all the terms covered by the expressed protein repertoire. The terms covered by the target genes of differentially regulated iRNAs during the time course were compared against the background. Terms supported by at least two iRNAs and p-value ≤ 0.01 were considered enriched. The same procedure was also applied for the targets of specific types of small RNAs, such as miRNAs and siRNAs, involved in cold response (CR, CRup, CRdw) and originating from ancestral genomes.

5 Figures and Tables

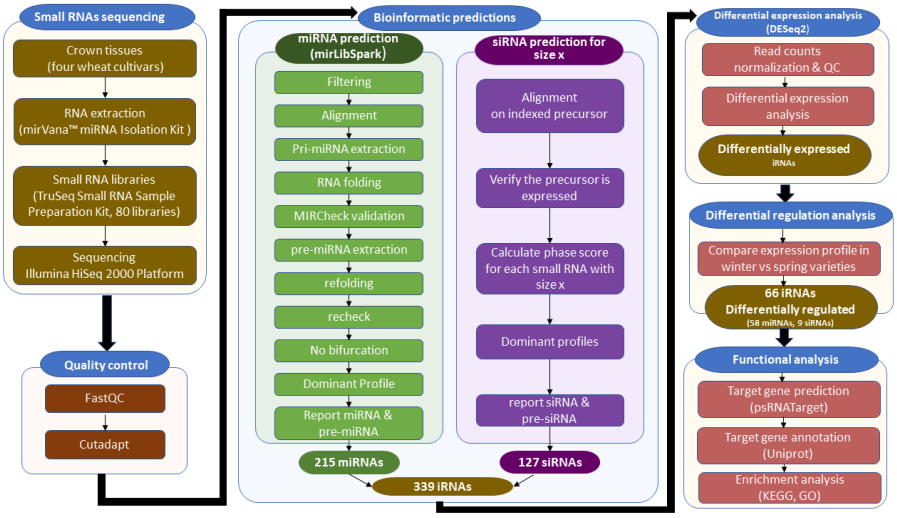


Figure 1: Bioinformatics analysis workflow of wheat miRNA and siRNA prediction and characterization. The workflow is divided into six steps: Small RNA sequencing, Quality control, Bioinformatics predictions, Differential expression analysis, Differential regulation analysis, and Functional analysis. The bioinformatics predictions contain miRNA predictions made by mirLibSpark and siRNA predictions made using in-house Python scripts. The numbers of predicted iRNAs were obtained from analyzing 80 wheat libraries.

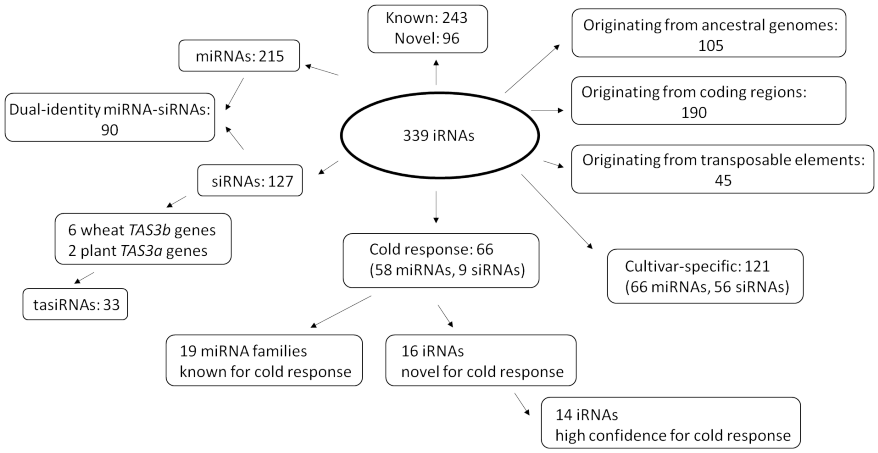


Figure 2: Classification of predicted iRNAs. A total of 339 iRNAs were identified from 80 wheat sRNA-seq libraries, including 96 novel siRNAs, 33 trans-acting siRNAs (tasiRNAs), and 90 dual-identity iRNAs. Expression profiling revealed 58 miRNAs and 9 siRNAs associated with cold responses; they belong to 19 miRNA families known to be involved in the cold response, and 14 iRNAs are novel but highly confident in cold response. We categorized the 339 iRNAs based on the genomic context of their precursor loci into three groups: originating from ancestral genomes (105), coding regions (190), and transposable elements (45). There are 121 cultivar-specific iRNAs. Note that some iRNAs are dual identities and could be counted twice in subcategories.

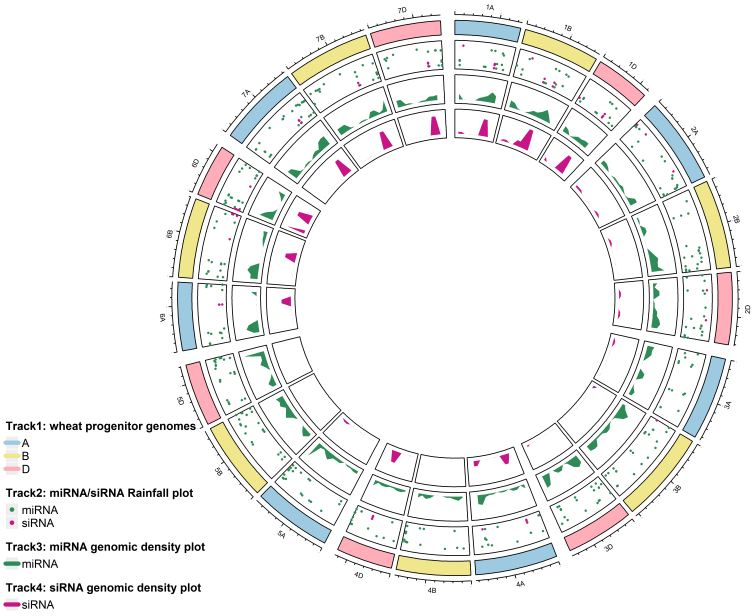


Figure 3: Chromosomal distribution of miRNAs and siRNAs in the wheat genome. This Circos plot displays miRNA and siRNA precursors' genomic locations and density across the three wheat progenitor genomes (A, B, and D). The outermost ring represents wheat chromosomes, with miRNA-originating loci mapped in the second ring and siRNA-originating loci in the third ring. The innermost heatmap tracks the genomic density of these small RNAs, revealing distinct clustering patterns. miRNAs are more evenly distributed, while siRNAs tend to form dense clusters, particularly in transposable element-rich regions. The plot highlights genomic hotspots with abundant phased siRNAs, suggesting potential regulatory roles. This figure provides an integrated view of small RNA chromosomal organization, helping to identify evolutionary trends and functional loci for further study.

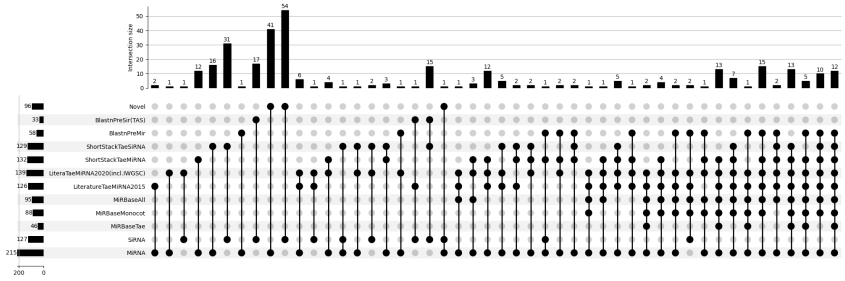


Figure 4: Cross-database comparison of predicted miRNAs and siRNAs. This UpSet plot illustrates the intersection of our predicted miRNAs and siRNAs with independent small RNA datasets and predictors from miRBase, literature sources, ShortStack predictions, and BLASTN-annotated precursors. (Left panel) The horizontal bar graph of total number of sRNA species in this study seen in different database - MiRBaseTae/Monocot/All as the sRNA being annotated as wheat, monocot ot plant miRNA in MiRBase v22.0 using miRFinder; LiteratureTaeMiRNA2015/2020 as the sRNA being annotated as wheat miRNA in literature (including IWGSC); BlastnPreMir/Sir as their precursors being annotated as miRNA or siRNA precursor in BLASTN similarity search (BLASTN only found similarity with TAS for siRNA precursors); and ShortStackMiRNA/SiRNA as the sRNA being predicted as miRNA or siRNA from our libraries using ShortStack, or annotated in PSU wheat sRNA database using miRFinder. Novel collects the sRNAs in this study not found in any databases above. (Right panel) The dots and lines on the bottom show which intersection is represented by the vertical bar graph above.

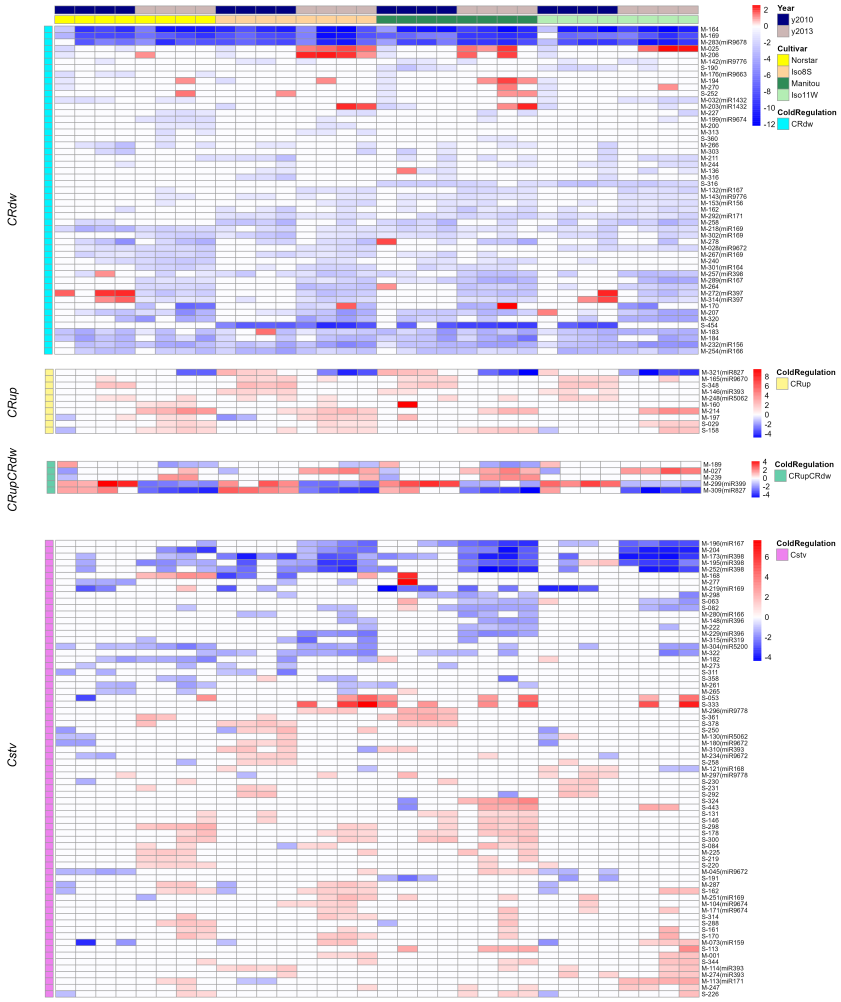


Figure 5: Temporal differential expression of miRNAs and siRNAs across four wheat cultivars in 2010 and 2013. iRNAs were classified into cold response categories (CRup, CRdw, CRupCRdw) using encoded vector clustering, which grouped them based on coordinated expression trends between winter and spring cultivars before plotting. Each panel separately displays these categories, revealing distinct regulatory responses and temporal dynamics. Each row represents the expression pattern of an iRNA across cultivars (Manitou, Iso11W, Norstar, and Iso8S) and time points in log2 fold change. Red, blue and white cells indicate expressions were up, down, or no change, respectively, as compared to their time point 1. All differentially expressed miRNAs and siRNAs are shown, except in the Cstv category that selected ones are shown by filtering changes in fewer than (<) 5 samples. MiR family of a miRNA, if available, is shown next to the ID.

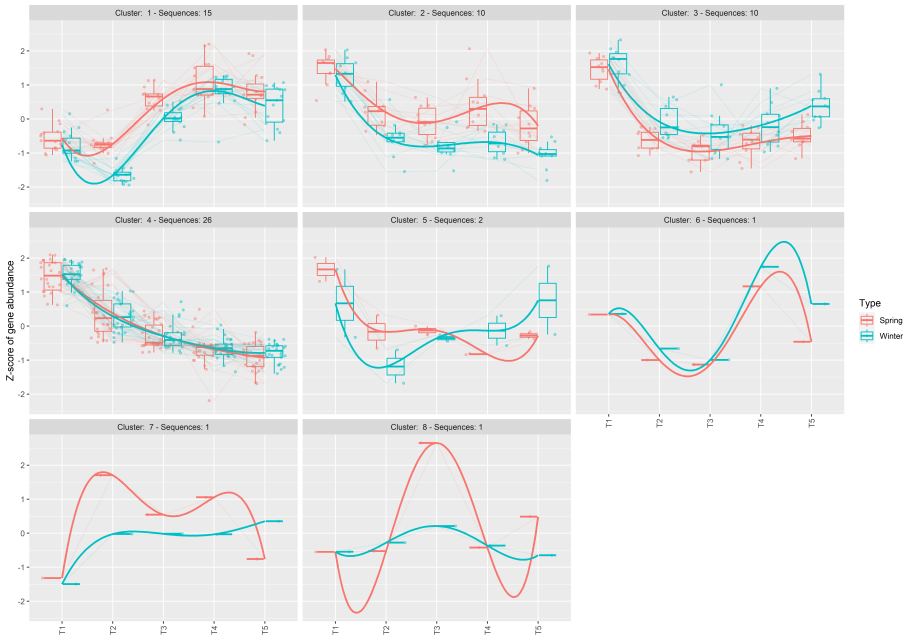


Figure 6: Winter-spring paired temporal expression profiles of cold response iRNAs. Identified 66 cold response iRNAs are clustered into 8 groups based on their differential expressions considering sampling time from warm to cold and cold hardiness between winter and spring cultivars.

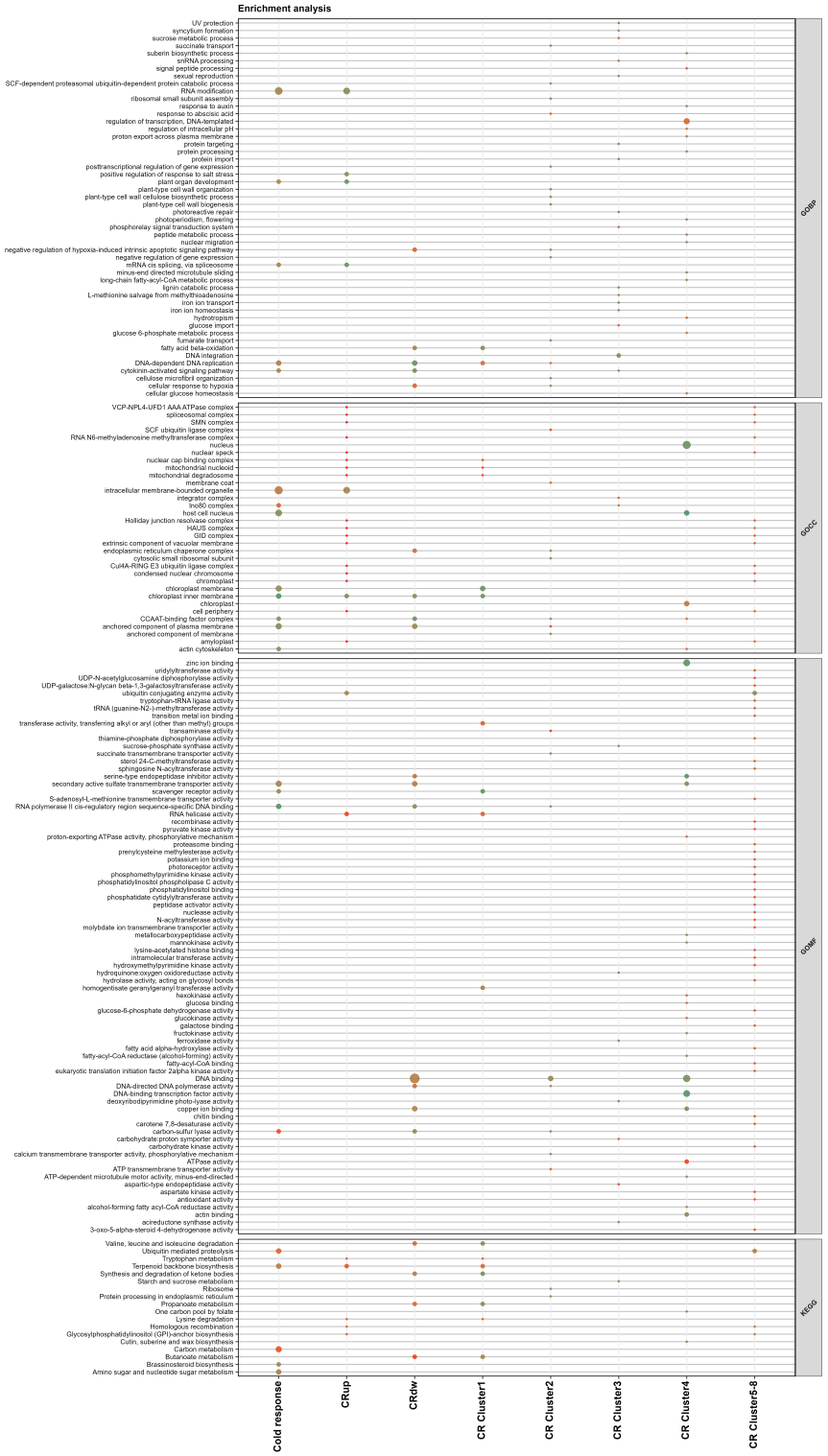

Enrichment analysis
GOBP
GOCC
GOMF
KEGG
Count
2
4
6
8
10
13
Pvalue
0.04
0.03
0.02
0.01
0.00
Cold response
CRup
CRdw
CR Cluster1
CR Cluster2
CR Cluster3
CR Cluster4
CR Cluster5-8

Figure 7: **Functional enrichment of target genes regulated by cold-response iRNAs.** Enrichment analysis was performed for different groups of cold-response iRNAs, including all cold-response iRNAs (**CR**), upregulated (**CRup**), downregulated (**CRdw**), and eight expression clusters (**C1**–**C8**). **CR iRNAs** are enriched in kinase signaling, transcriptional regulation, and cold stress response. **CRup iRNAs** are associated with mRNA splicing, plant organ development, RNA modification, and response to salt stress. **CRdw iRNAs** are linked to DNA replication, cytokinin-activated signaling, fatty acid beta-oxidation, and the negative regulation of hypoxia-induced intrinsic apoptotic signaling. Among the expression clusters, **C1** is enriched in fatty acid metabolism and DNA replication; **C2** in ribosomal assembly, cell wall biogenesis, and abscisic acid response; **C3** in photoreactive repair, UV protection, and glucose import; **C4** in auxin response, fatty acid metabolism, and proton export; and **C5**–**C8** in auxin-activated signaling pathways, transmembrane transport, RNA modification, and protein ubiquitination. Enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the standard hypergeometric distribution test for the terms covered by the target genes of each RNA group. Terms with FDR ≤ 0.05 were considered significantly enriched.

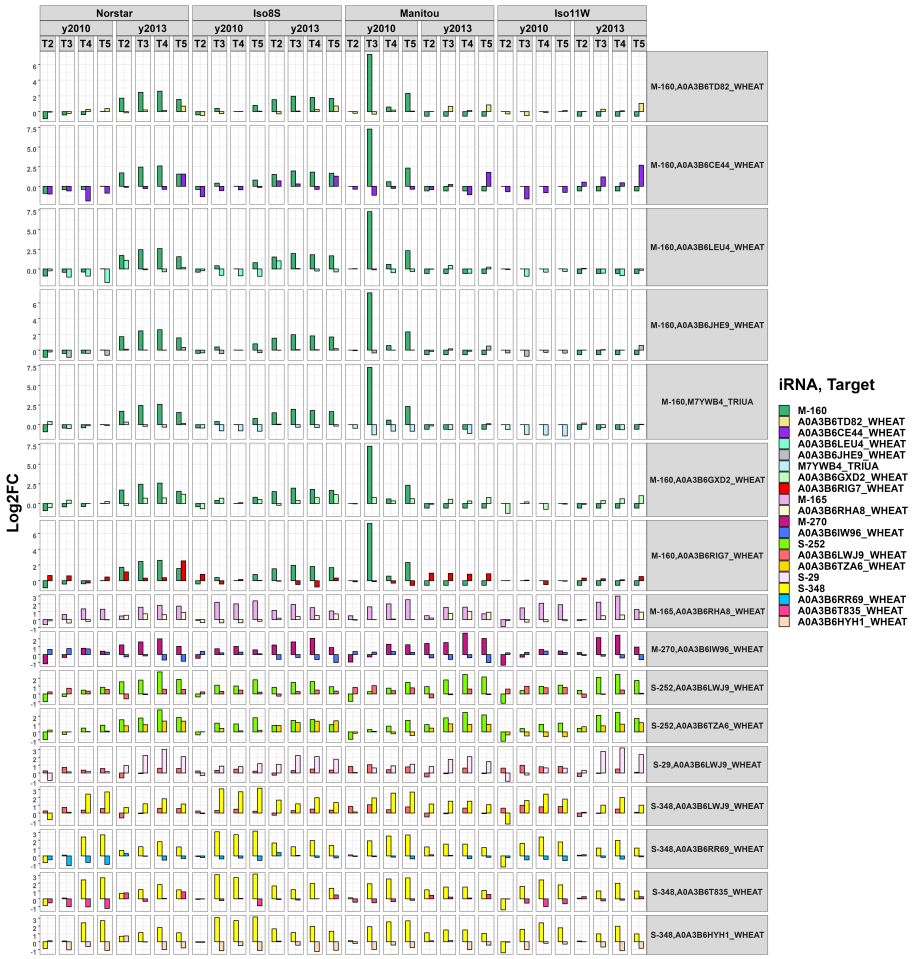


Figure 8: Gene expression patterns of cold response mature sequences and their kinase targets. Gene expression patterns of representative pairs of mature sequences, precursor, and downstream target kinase genes in 80 wheat libraries for selected cold response iRNAs. Log2 fold changes in the y-axis, and time point series in the x-axis.

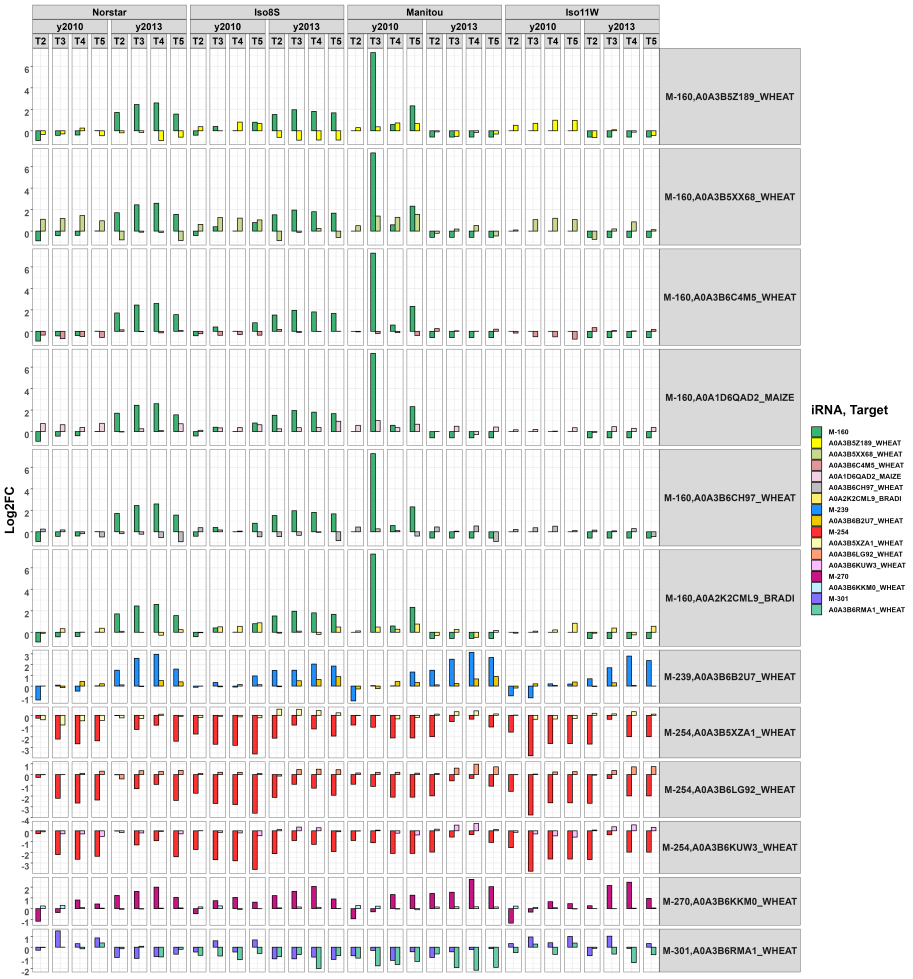


Figure 9: Gene expression patterns of cold response mature sequences and their transcription factor targets. Gene expression patterns of representative pairs of mature sequences, precursor, and downstream target kinase genes in 80 wheat libraries for selected cold response iRNAs. Log2 fold changes in the y-axis, and time point series in the x-axis.

Table 1: Overview of miRNA and siRNA prediction showing various categories of predicted iRNAs originating from (A) ancestral genomes, (B) coding regions, (C) transposable elements.

(A)	Total number	Number of iRNAs originating from ancestral genomes	%	Comment
Novel iRNA	96	17	17.7	Underrepresented (p=0.00051)
Known iRNA	243	88	36.2	Overrepresented (p=0.00051)
miRNA	215	80	37.2	Overrepresented (p=0.00068)
siRNA	127	26	20.5	Underrepresented (p=0.00078)
Dual identity iRNA	90	36	40.0	Overrepresented (p=0.0223)
iRNA originating from coding regions	190	81	42.6	Overrepresented (p<0.00001)
Cold response iRNA	66	22	33.3	Similar to background (p=0.37303)
All iRNA (background)	339	105	31.0	Background

(B)	Total number	Number of iRNAs originating from coding regions	%	Comment
Known iRNA	243	148	60.9	Overrepresented (p=0.00307)
Novel iRNA	96	42	43.8	Underrepresented (p=0.00307)
miRNA	215	108	50.2	Underrepresented (p=0.00306)
siRNA	127	83	65.4	Overrepresented (p=0.00508)
tasiRNA	33	26	78.8	Overrepresented (p=0.00392)
All iRNA (background)	339	190	56.1	Background

(C)	Total number	Number of iRNAs originating from transposable elements [1]	%	Comment
miRNA	215	40	18.6	Overrepresented (p=0.00005)
siRNA	127	5	3.9	Underrepresented (p=0.00003)
Dual identity iRNA	90	14	15.6	Similar to background (p=0.28205)
All iRNA (background)	339	45	13.3	Background

Over- and underrepresented: $p < 0.05$ by upper or lower side of the hypergeometric distribution test. 1: including transposable elements (TE) and long terminal repeats (LRT) loci.

Table 2: Known cold-response miRNA families identified in this study. Among the 66 cold-response iRNAs identified, several belong to well-characterized miRBase families with established roles in plant cold acclimation. These miRNAs regulate key adaptive processes, including transcriptional control, signal transduction, stress response pathways, and structural stability. This table provides supporting literature for their involvement in cold response. CRdw: cold response down-regulated; CRup: cold response up-regulated; na: not available.

miR family	NewID	Cold response	Cold response in literature	Other abiotic stresses in literature	Target description
miR1432	M-032,M-203	CRdw	na	water [34]	DNA-dependent DNA replication, cell wall biogenesis
miR156	M-232,M-153	CRdw	[92, 88, 91, 106]	UVB [35]	Squamosa promoter-binding-like proteins
miR164	M-301	CRdw	[66, 106, 92, 88]	salt, mechanical [35]	PIF3, Circadian Clock, flower development [56]
miR166	M-254	CRdw	[66, 92]	UVB [35]	Homeobox-leucine zipper protein HOX33
miR167	M-289,M-132	CRdw	[106, 66, 92, 88]	UVB [35]	Cytochrome P450
miR169	M-218,M-267,M-302	CRdw	[106, 92, 88]	UVB, salt [35]	Nuclear transcription factor Y subunit A-5
miR171	M-292	CRdw	[88, 91, 106]	UVB [35]	GRAS transcription factor
miR393	M-146	CRup	[106, 92, 88, 91]	UVB, abscisic acid, hypoxia [35]	Transport inhibitor response 1-like protein Os05g0150500
miR397	M-272,M-314	CRdw	[106, 88, 115, 91, 101]	water, salt [35]	Laccase-12, VRN-A1 allelic effect (both MA nad NO backgrounds)
miR398	M-257	CRdw	[88, 66, 99]	UVB, copper [35]	E3 ubiquitin-protein ligase listerin, snRNA processing
miR399	M-299	CRupCRdw	[78]	salt [35]	ubiquitin-conjugating enzyme
miR5062	M-248	CRup	[88]	na	Protein argonaute MEL1, Potassium transporter, Reverse transcriptase, Carboxypeptidase
miR827	M-321,M-309	CRup	[88]	na	SPX domain-containing membrane protein Os02g45520, disease resistance protein RGA3
miR9663	M-176	CRdw	[88]	na	Polyglutamine-binding protein 1
miR9670	M-165	CRup	[88]	na	G-type lectin S-receptor-like serine/threonine-protein kinase SD2-5
miR9672	M-028	CRdw	[88]	na	F-box only protein 7
miR9674	M-199	CRdw	[88]	Al [1]	phosphatase, NB-ARC, PPR_long, ANK_REP_REGION, and F-box domains-containing protein, mitochondrial protein Rf1
miR9678	M-283	CRdw	na	na	Serine protease SPPA
miR9776	M-142,M-143	CRdw	na	drought [82]	na

Table 3: High-confidence novel cold-response iRNAs with known functional targets identified as differentially expressed genes in Li et al. [56]. These novel cold-response iRNAs regulate stress perception, gene silencing, structural integrity, metabolic adaptation, protein stability, and transcriptional control.

NewID	Class	miR family	Cold response	Target Uniprot ID [1]	mRNA type [56]	Description
S-029	siRNA	noHit	CRup	A0A3B6BX82_WHEAT	NO backgrounds	Cysteine-rich receptor-like protein kinase 26
S-348	siRNA	noHit	CRup	A0A3B5Z507_WHEAT	NO backgrounds	G-type lectin S-receptor-like serine/threonine-protein kinase SRK
S-454	siRNA	noHit	CRdw	A0A3B6TTW8_WHEAT	NO backgrounds	Argonaute 1D
M-025	miRNA	noHit	CRdw	A0A080YUK3_WHEAT	MA backgrounds	Wall-associated receptor kinase 5, VRN-A1 allelic effect (both MA nad NO backgrounds [56])
M-136	miRNA	noHit	CRdw	A0A1D5UIN7_WHEAT	Up-regulated	Disease resistance RPP8-like protein 2
M-143	miRNA	noHit	CRdw	W5D700_WHEAT	Up-regulated	Subtilisin-chymotrypsin inhibitor-2A, VRN-A1 allelic effect (NO background)
M-160	miRNA	noHit	CRup	A0A3B6B273_WHEAT	Down-regulated	VRN-A1 allelic effect (NO background), ABA-dependent pathway [56]
M-169	miRNA	noHit	CRdw	A0A3B6NIM8_WHEAT	NO backgrounds	Fructan 1-exohydrolase w1
M-200	miRNA	noHit	CRdw	A0A3B6NIR3_WHEAT	NO backgrounds	Mitochondrial protein Rf1
M-239	miRNA	noHit	CRupCRdw	A0A3B6B2U7_WHEAT	Up-regulated	GATA transcription factor 5
M-266	miRNA	noHit	CRdw	A0A3B6NSR2_WHEAT	NO backgrounds	Heat shock 70 kDa protein 17
M-270	miRNA	noHit	CRdw	A0A3B6PNX4_WHEAT	MA backgrounds	Acetylornithine deacetylase
M-313	miRNA	noHit	CRdw	A0A3B6LTE8_WHEAT	Down-regulated	COBRA-like protein 3
M-320	miRNA	noHit	CRdw	A0A3B6NQ17_WHEAT	NO backgrounds	Serine/threonine-protein kinase DDB_G0276461

1: Uniprot ID of the target with the highest psRNAtarget score

Table 4: The cold response mature sequences and their kinase targets. Cold-response miRNAs and siRNAs' expression-correlated targets that were annotated as kinase are listed. We set the correlation strength to four levels from high to low by their spearman correlation coefficient's absolute values: strong, moderate, fair and weak [2]. The p value of spearman smaller than 0.05 was labeled with a star (*).

iRNA	Target Uniprot ID	Description	kinase known in literature for cold response	Spearman correlation coefficient	pvalue	Correlation strength
M-160	A0A3B6TD82_WHEAT	Wall-associated receptor kinase 3	[26, 49, 100]	0.25	0.02773	fair*
M-160	A0A3B6CE44_WHEAT	LRR receptor-like serine/threonine-protein kinase FLS2	[87, 118, 108, 49]	0.26	0.02087	fair*
M-160	A0A3B6LEU4_WHEAT	Protein kinase domain-containing protein	[107]	-0.52	0	moderate*
M-160	A0A3B6JHE9_WHEAT	Coatomer subunit beta'-3	[117]	-0.23	0.04097	fair*
M-160	M7YWB4_TRIUA	Casein kinase I isoform beta	[29, 81]	-0.42	0.0001	fair*
M-160	A0A3B6GXD2_WHEAT	Protein kinase domain-containing protein	[107]	0.3	0.00613	fair*
M-160	A0A3B6RIG7_WHEAT	Protein kinase domain-containing protein	[107]	-0.5	0	moderate*
M-165	A0A3B6RHA8_WHEAT	Receptor-like serine/threonine-protein kinase (EC 2.7.11.1)	[18]	-0.22	0.05296	fair
M-270	A0A3B6IW96_WHEAT	TPX2 domain-containing protein;Os03g0212600 protein (Targeting protein for Xklp2 containing protein, expressed)	na	0.35	0.00161	fair*
S-252	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	[107]	0.39	0.0004	fair*
S-252	A0A3B6TZA6_WHEAT	Protein kinase domain-containing protein	[107]	0.24	0.03276	fair*
S-29	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	[107]	-0.21	0.06668	fair
S-348	A0A3B6LWJ9_WHEAT	Protein kinase domain-containing protein	[107]	0.26	0.02102	fair*
S-348	A0A3B6RR69_WHEAT	Protein kinase domain-containing protein;NB-ARC domain-containing protein	[118]	-0.29	0.00932	fair*
S-348	A0A3B6T835_WHEAT	G-type lectin S-receptor-like serine/threonine-protein kinase SRK	[18]	-0.31	0.00585	fair*
S-348	A0A3B6HYH1_WHEAT	Wall-associated receptor kinase 5	[26, 49, 100]	-0.21	0.06585	fair

Table 5: The cold response mature sequences and their DNA-binding protein targets. Cold-response miRNAs and siRNAs' expression-correlated targets that were annotated as a DNA-binding or transcription factor are listed. We set the correlation strength to four levels from high to low by their spearman correlation coefficient's absolute values: strong, moderate, fair and weak [2]. The p value of spearman smaller than 0.05 was labeled with a star (*).

iRNA	Target Uniprot ID	Description	TF/DNA-binding known in literature for cold response	Spearman correlation coefficient	pvalue	Correlation strength
M-153	A0A3B6SFZ3_WHEAT	GW8-B1;SBP-type domain-containing protein;Squamosa promoter-binding-like protein 16	[102, 112]	-0.3	0.006	fair*
M-153	A0A3B6RIB5_WHEAT	SBP-type domain-containing protein;GW8-B1	na	0.3	0.007	fair*
M-160	A0A3B5Z189_WHEAT	Polyubiquitin	[30]	-0.64	0.000	moderate*
M-160	A0A3B5XX68_WHEAT	Helicase ATP-binding domain-containing protein;DNA helicase (EC 3.6.4.12)	[93]	0.6	0.000	moderate*
M-160	A0A3B6C4M5_WHEAT	ANK_REP_REGION domain-containing protein;Calmodulin-binding transcription activator 5	[98]	-0.48	0.000	fair*
M-160	A0A1D6QAD2_MAIZE	HAT transposon superfamily protein;rRNA N-glycosidase (EC 3.2.2.22)	na	-0.46	0.000	fair*
M-160	A0A3B6CH97_WHEAT	ERCC4 domain-containing protein	na	0.46	0.000	fair*
M-160	A0A2K2CML9_BRADI	DNA repair protein RAD51 homolog	na	0.38	0.001	fair*
M-160	A0A3B6C8P2_WHEAT	NAC domain-containing protein 8	[41, 27]	-0.24	0.034	fair*
M-218	T1MVL7_TRIUA	CCAAT-binding transcription factor B (CCAAT-binding transcription factor family protein)	na	-0.28	0.014	fair*
M-218	A0A3B6C2Y1_WHEAT	CCAAT-binding transcription factor B (CCAAT-binding transcription factor family protein)	na	-0.25	0.028	fair*
M-239	A0A3B6B2U7_WHEAT	GATA-type domain-containing protein;Os04g0539500 protein	[114]	0.25	0.025	fair*
M-254	A0A3B5XZA1_WHEAT	Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9)	na	-0.33	0.003	fair*
M-254	A0A3B6LG92_WHEAT	Homeobox-leucine zipper protein HOX33	na	0.29	0.008	fair*
M-254	A0A3B6KUW3_WHEAT	START domain-containing protein;Homeobox-leucine zipper protein HOX10	na	-0.29	0.008	fair*
M-270	A0A3B6KKM0_WHEAT	Methyl-CpG-binding domain-containing protein 2;CW-type domain-containing protein;MBD domain-containing protein	[42]	0.29	0.009	fair*
M-301	A0A3B6RMA1_WHEAT	NAC domain-containing protein	[67]	0.22	0.050	fair*

Table 6: Homology-Based Identification and Classification of TAS Precursors.

Classification (*BLASTN*)	**TAS precursor** (CONTIG)
TAS3b in wheat	2A__598975000, 2D__450395000, 6A__406360000, 6B__446715000, 6B__446720000, 6D__285740000
TAS3a in other plants	4A__158175000, 4D__316535000

Table 7: Comparison of *Vrn-A1*-Dependent iRNA Functions in Winter- and Spring-Habit Wheat.

Allele **Cultivar** **Phenotype**	**Recessive *vrn-A1*** Requires vernalization **Winter-Habit** (Norstar ∩ Iso11W) Cold-tolerant Delayed flowering	**Dominant *Vrn-A1*** No vernalization required **Spring-Habit** (Manitou ∩ Iso8S) Rapid development Cold-independent flowering
Cultivar-specific iRNA (CR in bold face)	***miR9674***, ***miR5062***	***miR166***, ***miR167***, ***miR319***, ***miR393***, *miR396*, *miR9778*
Target	MEL1 (Argonaute; regulates meiosis and gene silencing) Histone H3-K9 (chromatin modification)	GID1 (gibberellin receptor) GRF (growth-regulating factor)
Enriched Biological Processes	Meiosis Histone modification (methylation, phosphorylation, acetylation) Gene silencing Dormancy Flowering delay (Microgametogenesis, Megagametogenesis) Vacuole transport Negative regulation of embryonic development	Gibberellic acid homeostasis Chromatin assembly DNA replication Oxidative stress response Bilateral symmetry Meristem growth Hormone metabolism Fatty acid degradation

Table 8: Cultivar-specific iRNA families.

miRNA and tasiRNA family	**Manitou**	**Iso11W**	**Norstar**	**Iso8S**
miR159	S-217	M-159	M-073	M-073, M-238
Chr 4A (158175000)	S-378, S-361	-	S-274	S-221, S-361
Chr 6A (406360000)	-	-	S-265	S-447
Chr 6B (446715000)	S-129	S-129	S-358	S-358

Table footnotes:

- miR159 is a miRBase-defined miRNA family.
- tasiRNA families are defined by their origin from the same genomic segment that is homologous to known TAS genes. "Chr 4A (158175000)" denotes a precursor tag for a genomic segment on chromosome 4A, starting at position 158,175,000 and spanning 5,250 nucleotides. Other tags follow the same convention.

Acknowledgments

This research has been supported by funding from the Natural Sciences and Engineering Research Council of Canada (CGSD research grant to CJW, operational grants to ABD, and discovery grant to FS), the Fonds de recherche du Quebec - Nature et technologies (B2X research grant to CJW and operational grants to ABD), Genome Canada/Genome Prairie (DBF), Ducks Unlimited Canada (DBF), and the Saskatchewan Ministry of Agriculture and the Canada-Saskatchewan Growing Forward 2 bilateral agreement (DBF). This research was enabled in part by support provided by The Digital Research Alliance of Canada. The authors would like to thank Golrokh Kiani and Amine M. Remita for their participation in the methodology revision of the manuscript.

References

[1] Zahra Agharbaoui, Mickael Leclercq, Mohamed Amine Remita, Mohamed A Badawi, Etienne Lord, Mario Houde, Jean Danyluk, Abdoulaye Baniré Diallo, and Fathey Sarhan. An integrative approach to identify hexaploid wheat mirnaome associated with development and tolerance to abiotic stress. *BMC genomics*, 16(1):339, 2015.

[2] Haldun Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18 (3):91–93, 2018.

[3] Bala Ani Akpinar, Melda Kantar, and Hikmet Budak. Root precursors of micrornas in wild emmer and modern wheats show major differences in response to drought stress. *Functional & integrative genomics*, 15(5):587–598, 2015.

[4] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data. *Bioinformatics, Babraham Institute (Cambridge, UK)*, 2010.

[5] Rudi Appels, Kellye Eversole, Catherine Feuillet, Beat Keller, Jane Rogers, Nils Stein, Curtis J Pozniak, Frédéric Choulet, Assaf Distelfeld, Jesse Poland, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361(6403):eaar7191, 2018.

[6] Michael J Axtell. Classification and comparison of small rnas from plants. *Annual review of plant biology*, 64:137–159, 2013.

[7] Michael J Axtell. Shortstack: comprehensive annotation and quantification of small rna genes. *Rna*, 19(6):740–751, 2013.

[8] Michael J Axtell and Blake C Meyers. Revisiting criteria for plant microrna annotation in the era of big data. *The Plant Cell*, 30(2):272–284, 2018.

[9] Michael J Axtell, Calvin Jan, Ramya Rajagopalan, and David P Bartel. A two-hit trigger for sirna biogenesis in plants. *Cell*, 127(3):565–577, 2006.

[10] Patricia Baldrich, Klementina Kakar, Christelle Siré, Ana Beatriz Moreno, Angélique Berger, Meritxell García-Chapa, Juan José López-Moya, José Luis Riechmann, and Blanca San Segundo. Small rna profiling reveals regulation of arabidopsis mir168 and heterochromatic sirna415 in response to fungal elicitors. *BMC genomics*, 15:1–16, 2014.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[12] Nicholas H Bergman. *Comparative genomics, Chapter 9 BLAST QuickStart*. Springer Science & Business Media, 2007.

[13] Grzegorz M Boratyn, Christiam Camacho, Peter S Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L Madden, Wayne T Matten, Scott D McGinnis, Yuri Merezhuk, et al. Blast: a more efficient report with usability improvements. *Nucleic acids research*, 41(W1):W29–W33, 2013.

[14] Hikmet Budak and B Ani Akpinar. Plant mirnas: biogenesis, organization and origins. *Functional & integrative genomics*, 15(5):523–531, 2015.

[15] Caterina Catalanotto, Carlo Cogoni, and Giuseppe Zardo. Microrna in control of gene expression: an overview of nuclear functions. *International journal of molecular sciences*, 17(10):1712, 2016.

[16] Srinivas Chakravarthy, Samuel H Sternberg, Colleen A Kellenberger, and Jennifer A Doudna. Substrate-specific kinetics of dicer-catalyzed rna processing. *Journal of molecular biology*, 404(3): 392–402, 2010.

[17] Ho-Ming Chen, Yi-Hang Li, and Shu-Hsing Wu. Bioinformatic prediction and experimental validation of a microrna-directed tandem trans-acting sirna cascade in arabidopsis. *Proceedings of the National Academy of Sciences*, 104(9):3318–3323, 2007.

[18] Xuexue Chen, Yanglin Ding, Yongqing Yang, Chunpeng Song, Baoshan Wang, Shuhua Yang, Yan Guo, and Zhizhong Gong. Protein kinases in plant responses to drought, salt, and cold stress. *Journal of integrative plant biology*, 63(1):53–78, 2021.

[19] Ana Conesa and Stefan Götz. Blast2go: a comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, 2008, 2008.

[20] Gene Ontology Consortium et al. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261, 2004.

[21] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.

[22] UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49(D1):D480–D489, 2021.

[23] Xinbin Dai and Patrick Xuechun Zhao. psrnatarget: a plant small rna target analysis server. *Nucleic acids research*, 39(suppl 2):W155–W159, 2011.

[24] Xinbin Dai, Zhaohong Zhuang, and Patrick Xuechun Zhao. psrnatarget: a plant small rna target analysis server (2017 release). *Nucleic acids research*, 46(W1):W49–W54, 2018.

[25] Felipe Fenselau de Felippes, Antonin Marchais, Alexis Sarazin, Stefan Oberlin, and Olivier Voinnet. A single mir390 targeting event is sufficient for triggering tas3-tasirna biogenesis in arabidopsis. *Nucleic acids research*, 45(9):5539–5554, 2017.

[26] Luiz Felipe Valter de Oliveira, Ana Paula Christoff, Júlio Cesar de Lima, Bruno Comparsi Feijó de Ross, Gilberto Sachetto-Martins, Marcia Margis-Pinheiro, and Rogerio Margis. The wall-associated kinase gene family in rice genomes. *Plant Science*, 229:181–192, 2014.

[27] Pengfei Diao, Chong Chen, Yuzhen Zhang, Qingwei Meng, Wei Lv, and Nana Ma. The role of nac transcription factor in plant cold response. *Plant Signaling & Behavior*, 15(9):1785668, 2020.

[28] John G Doench, Christian P Petersen, and Phillip A Sharp. sirnas can function as mirnas. *Genes & development*, 17(4):438–442, 2003.

[29] Chun-Hai Dong and Haixia Pei. Over-expression of mir397 improves plant tolerance to cold stress in arabidopsis thaliana. *Journal of Plant Biology*, 57(4):209–217, 2014.

[30] Chun-Hai Dong, Manu Agarwal, Yiyue Zhang, Qi Xie, and Jian-Kang Zhu. The negative regulator of plant cold responses, hos1, is a ring e3 ligase that mediates the ubiquitination and degradation of ice1. *Proceedings of the National Academy of Sciences*, 103(21):8281–8286, 2006.

[31] Yujie Fang, Kabin Xie, and Lizhong Xiong. Conserved mir164-targeted nac genes negatively regulate drought resistance in rice. *Journal of experimental botany*, 65(8):2119–2135, 2014.

[32] Thalia A Farazi, Stefan A Juranek, and Thomas Tuschl. The growing catalog of small rnas and their association with distinct argonaute/piwi family members. *Development*, 135(7):1201–1214, 2008.

[33] Qili Fei, Rui Xia, and Blake C Meyers. Phased, secondary, small interfering rnas in posttranscriptional regulatory networks. *The Plant Cell*, 25(7):2400–2415, 2013.

[34] Lorenzo Giusti, Erica Mica, Edoardo Bertolini, Anna Maria De Leonardis, Primetta Faccioli, Luigi Cattivelli, and Cristina Crosatti. microrna differentially modulated in response to heat and drought stress in durum wheat cultivars with contrasting water use efficiency. *Functional & integrative genomics*, 17(2-3):293–309, 2017.

[35] Praveen Guleria, Monika Mahajan, Jyoti Bhardwaj, and Sudesh Kumar Yadav. Plant small rnas: biogenesis, mode of action and their roles in abiotic stresses. *Genomics, proteomics & bioinformatics*, 9(6):183–199, 2011.

[36] Cheng Guo, Li Li, Xifeng Wang, and Chun Liang. Alterations in sirna and mirna expression profiles detected by deep sequencing of transgenic rice with sirna-mediated viral resistance. *PLoS One*, 10 (1):e0116175, 2015.

[37] Guanghui Guo, Xinye Liu, Fenglong Sun, Jie Cao, Na Huo, Bala Wuda, Mingming Xin, Zhaorong Hu, Jinkun Du, Rui Xia, et al. Wheat mir9678 affects seed germination by generating phased sirnas and modulating abscisic acid/gibberellin signaling. *The Plant Cell*, 30(4):796–814, 2018.

[38] Ran Han, Chao Jian, Jinyang Lv, Yan Yan, Qing Chi, Zhanjie Li, Qian Wang, Jin Zhang, Xiangli Liu, and Huixian Zhao. Identification and characterization of microrna s in the flag leaf and developing seed of wheat (triticum aestivum l.). *BMC genomics*, 15(1):289, 2014.

[39] Norikazu Hara, Masataka Kikuchi, Akinori Miyashita, Hiroyuki Hatsuta, Yuko Saito, Kensaku Kasuga, Shigeo Murayama, Takeshi Ikeuchi, and Ryozo Kuwano. Serum microrna mir-501-3p as a potential biomarker related to the progression of alzheimer's disease. *Acta neuropathologica communications*, 5:1–9, 2017.

[40] Miya D Howell, Noah Fahlgren, Elisabeth J Chapman, Jason S Cumbie, Christopher M Sullivan, Scott A Givan, Kristin D Kasschau, and James C Carrington. Genome-wide analysis of the rna-dependent rna polymerase6/dicer-like4 pathway in arabidopsis reveals dependency on mirna-and tasirna-directed targeting. *The Plant Cell*, 19(3):926–942, 2007.

[41] Ruibo Hu, Guang Qi, Yingzhen Kong, Dejing Kong, Qian Gao, and Gongke Zhou. Comprehensive analysis of nac domain transcription factor gene family in populus trichocarpa. *BMC plant biology*, 10(1):1–23, 2010.

[42] Zhaorong Hu, Ying Yu, Rui Wang, Yingyin Yao, Huiru Peng, Zhongfu Ni, and Qixin Sun. Expression divergence of tambd2 homoeologous genes encoding methyl cpg-binding domain proteins in wheat (triticum aestivum l.). *Gene*, 471(1-2):13–18, 2011.

[43] Matthew W Jones-Rhoades and David P Bartel. Computational identification of plant microrna s and their targets, including a stress-induced mirna. *Molecular cell*, 14(6):787–799, 2004.

[44] Trupti Joshi and Dong Xu. Quantitative assessment of relationship between sequence similarity and function similarity. *BMC genomics*, 8(1):1–10, 2007.

[45] Atul Kakrana, Pingchuan Li, Parth Patel, Reza Hammond, Deepti Anand, Sandra Mathioni, and Blake Meyers. Phasis: A computational suite for de novo discovery and characterization of phased, sirna-generating loci and their mirna triggers. *BioRxiv*, page 158832, 2017.

[46] Atul Kakrana, Sandra M Mathioni, Kun Huang, Reza Hammond, Lee Vandivier, Parth Patel, Siwaret Arikit, Olga Shevchenko, Alex E Harkess, Bruce Kingham, et al. Plant 24-nt reproductive phasirnas from intramolecular duplex mrnas in diverse monocots. *Genome research*, 28(9):1333–1344, 2018.

[47] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[48] Melda Kantar, Stuart J Lucas, and Hikmet Budak. mirna expression patterns of triticum dicoccoides in response to shock drought stress. *Planta*, 233(3):471–484, 2011.

[49] Seyyedeh-Sanam Kazemi-Shahandashti and Reza Maali-Amiri. Global insights of protein responses to cold stress in plants: Signaling, defence, and degradation. *Journal of plant physiology*, 226: 123–135, 2018.

[50] Michal Kenan-Eichler, Dena Leshkowitz, Lior Tal, Elad Noor, Cathy Melamed-Bessudo, Moshe Feldman, and Avraham A Levy. Wheat hybridization and polyploidization results in deregulation of small rnas. *Genetics*, 188(2):263–272, 2011.

[51] Basel Khraiwesh, Jian-Kang Zhu, and Jianhua Zhu. Role of mirnas and sirnas in biotic and abiotic stress responses of plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(2):137–148, 2012.

[52] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2018.

[53] Kuaybe Yucebilgili Kurtoglu, Melda Kantar, Stuart J Lucas, and Hikmet Budak. Unique and conserved micrornas in wheat chromosome 5d revealed by next-generation sequencing. *PLoS One*, 8(7):e69801, 2013.

[54] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[55] Jikai Lei and Yanni Sun. mir-prefer: an accurate, fast and easy-to-use plant mirna prediction tool using small rna-seq data. *Bioinformatics*, 30(19):2837–2839, 2014.

[56] Qiang Li, Brook Byrns, Mohamed A Badawi, Abdoulaye Banire Diallo, Jean Danyluk, Fathey Sarhan, Debbie Laudencia-Chingcuanco, Jitao Zou, and D Brian Fowler. Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant physiology*, 176 (3):2376–2394, 2018.

[57] Yang Li, Chaoqun Li, Guohui Ding, and Youxin Jin. Evolution of mir159/319 microrna genes and their post-transcriptional regulatory link to sirna pathways. *BMC evolutionary biology*, 11(1):1–18, 2011.

[58] Yang Li, Chaoqun Li, Jie Xia, and Youxin Jin. Domestication of transposable elements into microrna genes in plants. *Plos one*, 6(5):e19212, 2011.

[59] Han-Hua Liu, Xin Tian, Yan-Jie Li, Chang-Ai Wu, and Cheng-Chao Zheng. Microarray-based analysis of stress-regulated micrornas in arabidopsis thaliana. *Rna*, 14(5):836–843, 2008.

[60] Yong-Xin Liu, Meng Wang, and Xiu-Jie Wang. Endogenous small rna clusters in plants. *Genomics, Proteomics & Bioinformatics*, 12(2):64–71, 2014.

[61] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, Nov 2011.

[62] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[63] Stuart J Lucas and Hikmet Budak. Sorting the wheat from the chaff: identifying mirnas in genomic survey sequences of triticum aestivum chromosome 1al. *PLoS One*, 7(7):e40859, 2012.

[64] Anna Lukasik, Maciej Wójcikowski, and Piotr Zielenkiewicz. Tools4mirs–one place to gather all the tools for mirna analysis. *Bioinformatics*, 32(17):2722–2724, 2016.

[65] Alice Lunardon, Nathan R Johnson, Emily Hagerott, Tamia Phifer, Seth Polydore, Ceyda Coruh, and Michael J Axtell. Integrated annotations and analyses of small rna–producing loci from 47 diverse plants. *Genome research*, 30(3):497–513, 2020.

[66] De-Kang Lv, Xi Bai, Yong Li, Xiao-Dong Ding, Ying Ge, Hua Cai, Wei Ji, Nalahu Wu, and Yan-Ming Zhu. Profiling of cold-stress-responsive mirnas in rice by microarrays. *Gene*, 459(1-2):39–47, 2010.

[67] Xiaolong Lv, Shanrong Lan, Kateta Malangisha Guy, Jinghua Yang, Mingfang Zhang, and Zhongyuan Hu. Global expressions landscape of nac transcription factor family and their responses to abiotic stresses in citrullus lanatus. *Scientific Reports*, 6(1):1–14, 2016.

[68] Xingli Ma, Zeyu Xin, Zhiqiang Wang, Qinghua Yang, Shulei Guo, Xiaoyang Guo, Liru Cao, and Tongbao Lin. Identification and comparative analysis of differentially expressed mirnas in leaves of two wheat (triticum aestivum l.) genotypes during dehydration stress. *BMC Plant Biology*, 15(1): 21, 2015.

[69] Allison C Mallory and Hervé Vaucheret. Functions of micrornas and related small rnas in plants. *Nature genetics*, 38(6):S31–S36, 2006.

[70] Thomas Marcussen, Simen R Sandve, Lise Heier, Manuel Spannagl, Matthias Pfeifer, Kjetill S Jakobsen, Brande BH Wulff, Burkhard Steuernagel, Klaus FX Mayer, Odd-Arne Olsen, et al. Ancient hybridizations among the ancestral genomes of bread wheat. *science*, 345(6194), 2014.

[71] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[72] Anikó Meijer, Tim De Meyer, Klaas Vandepoele, and Tina Kyndt. Spatiotemporal expression profile of novel and known small rnas throughout rice plant development focussing on seed tissues. *BMC genomics*, 23(1):1–17, 2022.

[73] Blake C Meyers, Michael J Axtell, Bonnie Bartel, David P Bartel, David Baulcombe, John L Bowman, Xiaofeng Cao, James C Carrington, Xuemei Chen, Pamela J Green, et al. Criteria for annotation of plant micrornas. *The Plant Cell*, 20(12):3186–3190, 2008.

[74] Donald Danforth Plant Science Center Meyers Lab. MPSS database: List of genome-specific PHAS loci. `https://mpss.danforthcenter.org/phas_index.php`, 2023. Accessed: 2023-04-13.

[75] Aleksandr Morgulis, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. A fast and symmetric dust implementation to mask low-complexity dna sequences. *Journal of Computational Biology*, 13(5):1028–1040, 2006.

[76] Youlian Pan, Yifeng Li, Ziying Liu, Jitao Zou, and Qiang Li. Computational genomics insights into cold acclimation in wheat. *Frontiers in Genetics*, 13, 2022.

[77] L Pantano. Degreport: Report of deg analysis. r package version 1.34.0. `http://lpantano.github.io/DEGreport/`, 2022.

[78] Kankan Peng, Yu Tian, Xianze Sun, Chunhua Song, Zhipeng Ren, Yuzhuo Bao, Jinpu Xing, Yuanshan Li, Qinghua Xu, Jing Yu, et al. tae-mir399-ubc24 module enhances freezing tolerance in winter wheat via a cbf signaling pathway. *Journal of Agricultural and Food Chemistry*, 69(45):13398–13415, 2021.

[79] Jittima Piriyapongsa and I King Jordan. Dual coding of sirnas and mirnas by plant transposable elements. *Rna*, 14(5):814–821, 2008.

[80] Robert F Place, Long-Cheng Li, Deepa Pookot, Emily J Noonan, and Rajvir Dahiya. Microrna-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences*, 105(5):1608–1613, 2008.

[81] Myrthe Praat, Ive De Smet, and Martijn van Zanten. Protein kinase and phosphatase control of plant temperature responses. *Journal of Experimental Botany*, 72(21):7459–7473, 2021.

[82] Amna Abdul Rahim, Muhammad Uzair, Nazia Rehman, Obaid Ur Rehman, Nageen Zahra, and Muhammad Ramzan Khan. Genome-wide identification and characterization of receptor-like protein kinase 1 (rpk1) gene family in triticum aestivum under drought stress. *Frontiers in Genetics*, 13, 2022.

[83] Farzaneh Rahmdani, Amine M Remita, Chao-Jung Wu, and Abdoulaye Banire Diallo. Assessing the evidence of published mirnas in wheat (triticum aestivum). In *The 24th Annual International Conference on Research in Computational Molecular Biology, June 22-25, 2020, Virtual*, page poster. RECOMB, 2020.

[84] Mohamed Amine Remita, Etienne Lord, Zahra Agharbaoui, Mickael Leclercq, Mohamed A Badawi, Fathey Sarhan, and Abdoulaye Baniré Diallo. A novel comprehensive wheat mirna database, including related bioinformatics software. *Current Plant Biology*, 7:31–33, 2016.

[85] Grzegorz Rorbach, Olgierd Unold, and Bogumil M Konopka. Distinguishing mirtrons from canonical mirnas with data exploration and machine learning methods. *Scientific reports*, 8(1):7560, 2018.

[86] Mélodie Sawicki, Marine Rondeau, Barbara Courteaux, Fanja Rabenoelina, Gea Guerriero, Eric Gomès, Ludivine Soubigou-Taconnat, Sandrine Balzergue, Christophe Clément, Essaïd Ait Barka, et al. On a cold night: transcriptomics of grapevine flower unveils signal transduction and impacted metabolism. *International journal of molecular sciences*, 20(5):1130, 2019.

[87] Amolkumar U Solanke and Arun K Sharma. Signal transduction during cold stress in plants. *Physiology and Molecular biology of plants*, 14(1):69–79, 2008.

[88] Guoqi Song, Rongzhi Zhang, Shujuan Zhang, Yulian Li, Jie Gao, Xiaodong Han, Mingli Chen, Jiao Wang, Wei Li, and Genying Li. Response of micrornas to cold treatment in the young spikes of common wheat. *BMC genomics*, 18(1):212, 2017.

[89] Anna Stief, Krzysztof Brzezinka, Jörn Lämke, and Isabel Bäurle. Epigenetic responses to heat stress at different time scales and the involvement of small rnas. *Plant Signaling & Behavior*, 9(10): e970430, 2014.

[90] Fenglong Sun, Weiwei Guo, Jinkun Du, Zhongfu Ni, Qixin Sun, and Yingyin Yao. Widespread, abundant, and diverse te-associated sirnas in developing wheat grain. *Gene*, 522(1):1–7, 2013.

[91] Ramanjulu Sunkar, Yong-Fang Li, and Guru Jagadeeswaran. Functions of micrornas in plant stress responses. *Trends in plant science*, 17(4):196–203, 2012.

[92] Zhonghui Tang, Liping Zhang, Chenguang Xu, Shaohua Yuan, Fengting Zhang, Yonglian Zheng, and Changping Zhao. Uncovering small rna-mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing. *Plant physiology*, 159(2):721–738, 2012.

[93] Narendra Tuteja, Sarvajeet Singh, and Renu Tuteja. Helicases in improving abiotic stress tolerance in crop plants. *Improving crop resistance to abiotic stress*, pages 435–449, 2012.

[94] Ayla Valinezhad Orang, Reza Safaralizadeh, Mina Kazemzadeh-Bavili, et al. Mechanisms of mirna-mediated gene regulation from common downregulation to mrna-specific upregulation. *International journal of genomics*, 2014, 2014.

[95] Hervé Vaucheret and Olivier Voinnet. The plant sirna landscape. *The Plant Cell*, 36(2):246–275, 2024.

[96] André Verdel, Songtao Jia, Scott Gerber, Tomoyasu Sugiyama, Steven Gygi, Shiv IS Grewal, and Danesh Moazed. Rnai-mediated targeting of heterochromatin by the rits complex. *Science*, 303 (5658):672–676, 2004.

[97] Ioannis S Vlachos, Konstantinos Zagganas, Maria D Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G Hatzigeorgiou. Diana-mirpath v3. 0: deciphering microrna function with experimental support. *Nucleic acids research*, 43(W1):W460–W466, 2015.

[98] Kieu Thi Xuan Vo, Chi-Yeol Kim, Anil Kumar Nalini Chandran, Ki-Hong Jung, Gynheung An, and Jong-Seong Jeon. Molecular insights into the function of ankyrin proteins in plants. *Journal of plant biology*, 58(5):271–284, 2015.

[99] Bing Wang, Yan-fei Sun, Na Song, Jin-ping Wei, Xiao-jie Wang, Hao Feng, Zhi-yuan Yin, and Zhen-sheng Kang. Micrornas involving in cold, wounding and salt stresses in triticum aestivum l. *Plant Physiology and Biochemistry*, 80:90–96, 2014.

[100] Hu Wang, Huanhuan Niu, Minmin Liang, Yufei Zhai, Wei Huang, Qin Ding, Yu Du, and Minghui Lu. A wall-associated kinase gene cawakl20 from pepper negatively modulates plant thermotolerance by reducing the expression of aba-responsive genes. *Frontiers in plant science*, page 591, 2019.

[101] Qiongshan Wang, Nian Liu, Xiyan Yang, Lili Tu, and Xianlong Zhang. Small rna-mediated responses to low-and high-temperature stresses in cotton. *Scientific reports*, 6:35558, 2016.

[102] Yi Wang, Zongli Hu, Yuxin Yang, Xuqing Chen, and Guoping Chen. Function annotation of an sbp-box gene in arabidopsis based on analysis of co-expression networks and promoters. *International journal of molecular sciences*, 10(1):116–132, 2009.

[103] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. Long noncoding rnas: functional surprises from the rna world. *Genes & development*, 23(13):1494–1504, 2009.

[104] Chao-Jung Wu, Amine M Remita, and Abdoulaye Baniré Diallo. Mirlibspark: A scalable ngs plant microrna prediction pipeline for multi-library functional annotation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 669–674. ACM, 2019.

[105] Fuliang Xie, Charles Neal Stewart Jr, Faten A Taki, Qiuling He, Huawei Liu, and Baohong Zhang. High-throughput deep sequencing shows that micro rna s play important roles in switchgrass responses to drought and salinity stress. *Plant Biotechnology Journal*, 12(3):354–366, 2014.

[106] Shengchun Xu, Na Liu, Weihua Mao, Qizan Hu, Guofu Wang, and Yaming Gong. Identification of chilling-responsive micrornas and their targets in vegetable soybean (glycine max l.). *Scientific reports*, 6:26619, 2016.

[107] Sudesh Kumar Yadav. Cold stress tolerance mechanisms in plants. a review. *Agronomy for sustainable development*, 30(3):515–527, 2010.

[108] Liang Yang, Kangcheng Wu, Peng Gao, Xiaojuan Liu, Guangpu Li, and Zujian Wu. Gslrpk, a novel cold-activated leucine-rich repeat receptor-like protein kinase from glycine soja, is a positive regulator to cold stress tolerance. *Plant Science*, 215:19–28, 2014.

[109] Yingyin Yao, Zhongfu Ni, Huiru Peng, Fenglong Sun, Mingming Xin, Ramanjulu Sunkar, Jian-Kang Zhu, and Qixin Sun. Non-coding small rnas responsive to abiotic stress in wheat (triticum aestivum l.). *Functional & integrative genomics*, 10:187–190, 2010.

[110] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[111] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

[112] Xiucun Zeng, Yaozhao Xu, Jinjin Jiang, Fenqin Zhang, Li Ma, Dewei Wu, Youping Wang, and Wancang Sun. Identification of cold stress responsive micrornas in two winter turnip rape (brassica rapa l.) by high throughput sequencing. *BMC Plant Biology*, 18:1–13, 2018.

[113] Jixian Zhai, Han Zhang, Siwaret Arikit, Kun Huang, Guo-Ling Nan, Virginia Walbot, and Blake C Meyers. Spatiotemporally dynamic, cell-type–dependent premeiotic and meiotic phasirnas in maize anthers. *Proceedings of the National Academy of Sciences*, 112(10):3146–3151, 2015.

[114] Hongjia Zhang, Tao Wu, Zhao Li, Kai Huang, Na-Eun Kim, Ziming Ma, Soon-Wook Kwon, Wenzhu Jiang, and Xinglin Du. Osgata16, a gata transcription factor, confers cold tolerance by repressing oswrky45–1 at the seedling stage in rice. *Rice*, 14(1):1–15, 2021.

[115] Jingyu Zhang, Yunyuan Xu, Qing Huan, and Kang Chong. Deep sequencing of brachypodium small rnas at the global genome level identifies micrornas involved in cold stress response. *BMC genomics*, 10(1):449, 2009.

[116] Xiaoming Zhang, Jing Xia, Yifan E Lii, Blanca E Barrera-Figueroa, Xuefeng Zhou, Shang Gao, Lu Lu, Dongdong Niu, Zheng Chen, Christy Leung, et al. Genome-wide analysis of plant nat-sirnas reveals insights into their distribution, biogenesis and function. *Genome biology*, 13(3):R20, 2012.

[117] Bo Zhou, Yutong Kang, Jingtong Leng, and Qijiang Xu. Genome-wide analysis of the mirna–mrnas network involved in cold tolerance in populus simonii× p. nigra. *Genes*, 10(6):430, 2019.

[118] Ying Zhu, Weiqiang Qian, and Jian Hua. Temperature modulates plant defense responses through nb-lrr proteins. *PLoS pathogens*, 6(4):e1000844, 2010.

Supplementary materials

A Supplementary figures

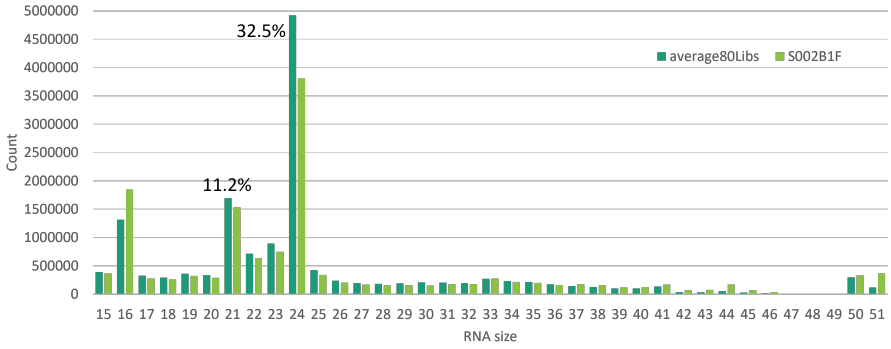


Figure SF1: Size distribution in the wheat libraries in this study. The average read counts for each size of the sRNAs in the 80 wheat libraries and the read counts of a sample library in this study are shown. On average, the most abundant sizes were 21-nt and 24-nt species, representing 11.2% and 32.5% of the RNA population, respectively.

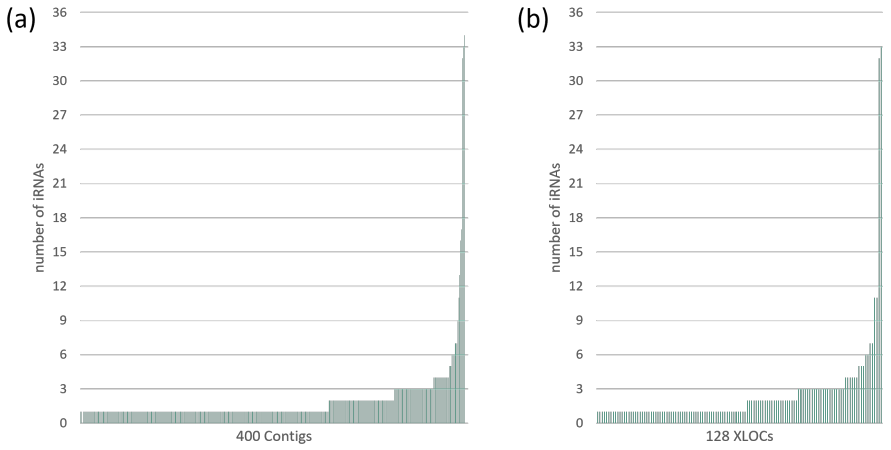


Figure SF2: The number of iRNAs coming from precursors of (a) genomic CONTIG clusters and (b) XLOC clusters are shown. The 339 mature iRNAs were derived from precursors aligned with the coordinates of 400 CONTIGs or 128 XLOCs. The number of iRNAs in a cluster ranges from 1 to 34. The majority of the clusters contained 1 to 3 iRNAs. There were 33 CONTIGs and 17 XLOCs that hosted between 4 and 34 miRNAs and/or siRNAs.

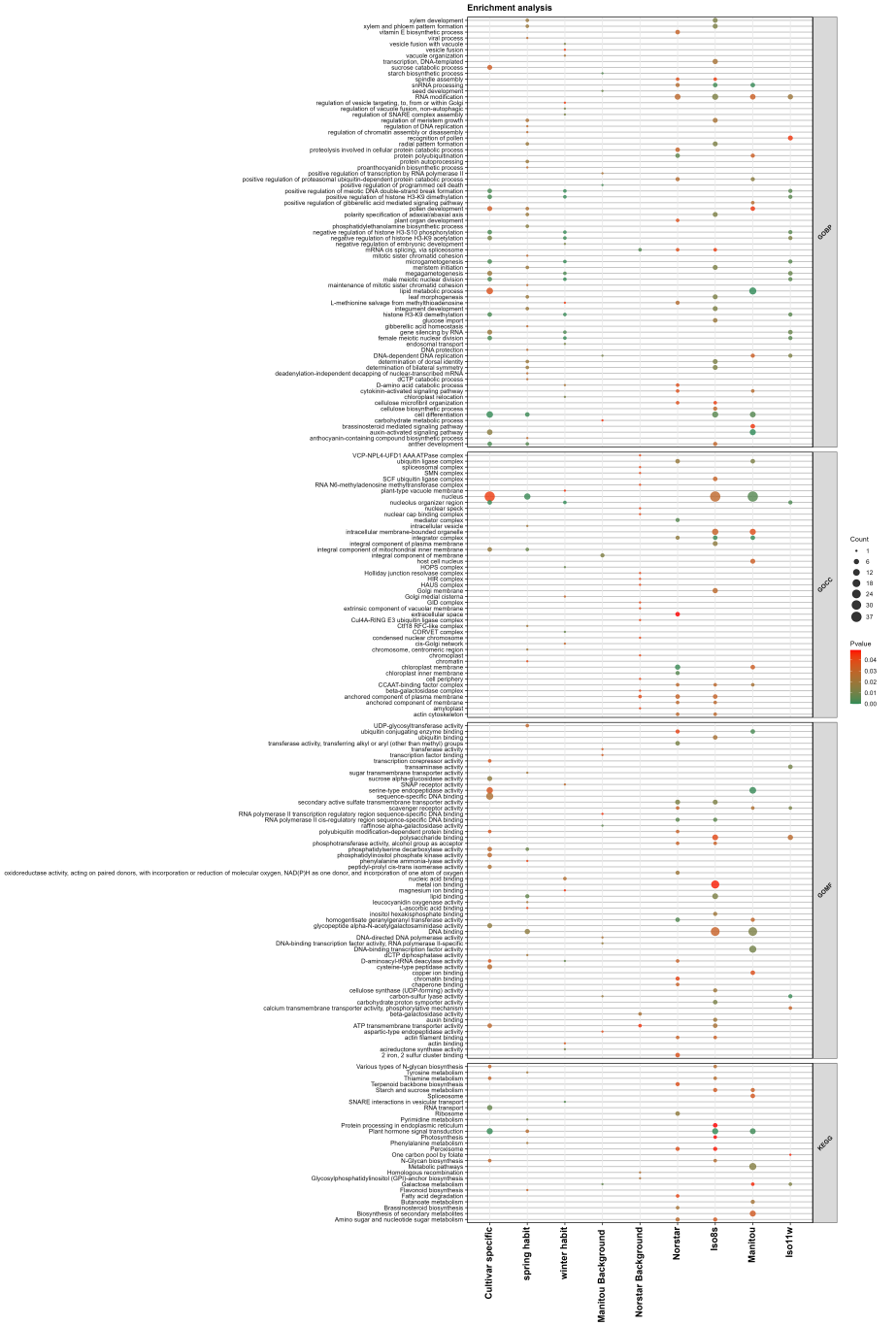


Figure SF3: Functional enrichments of target genes by cultivar-specific iRNAs. Enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the standard hypergeometric distribution test for the terms covered by the target genes of each RNA group. Terms supported by FDR ≤ 0.05 were considered enriched.

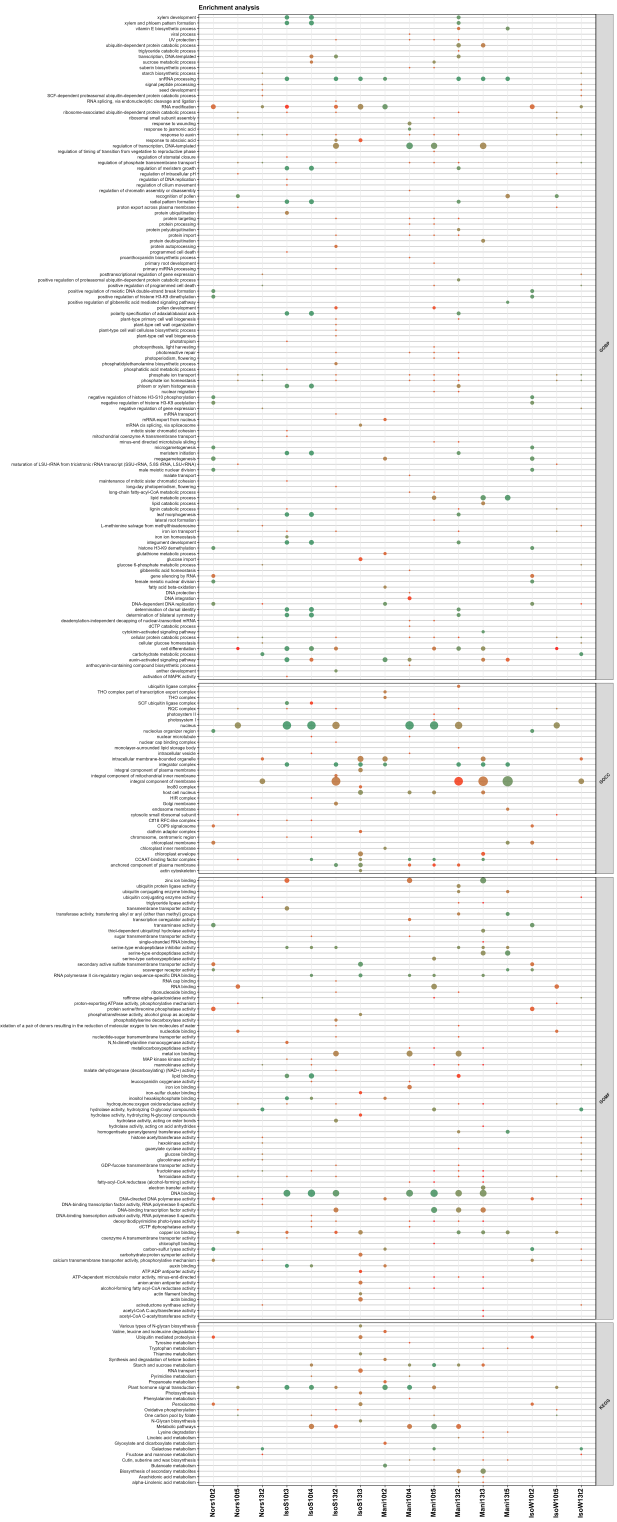

Enrichment analysis
BP
CC
MF
KEGG
Count
Padjust
Nors1to2
Nors1to5
Nors3to2
IsoS1to3
IsoS1to4
IsoS1to2
IsoS3to3
Mani1to2
Mani1to4
Mani1to5
Mani1to2
Mani1to3
Mani3to5
IsoW1to2
IsoW1to5
IsoW1to2

Figure SF4: Functional enrichments of target genes by temporal DE iRNAs. Time course data of the four cultivars (Manitou, IsoS, Norstar, IsoW) include time points t2 to t5 (each is compared to their t1) in two years 2010 and 2013. Enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the standard hypergeometric distribution test for the terms covered by the target genes of each RNA group. Terms supported by FDR ≤ 0.05 were considered enriched.

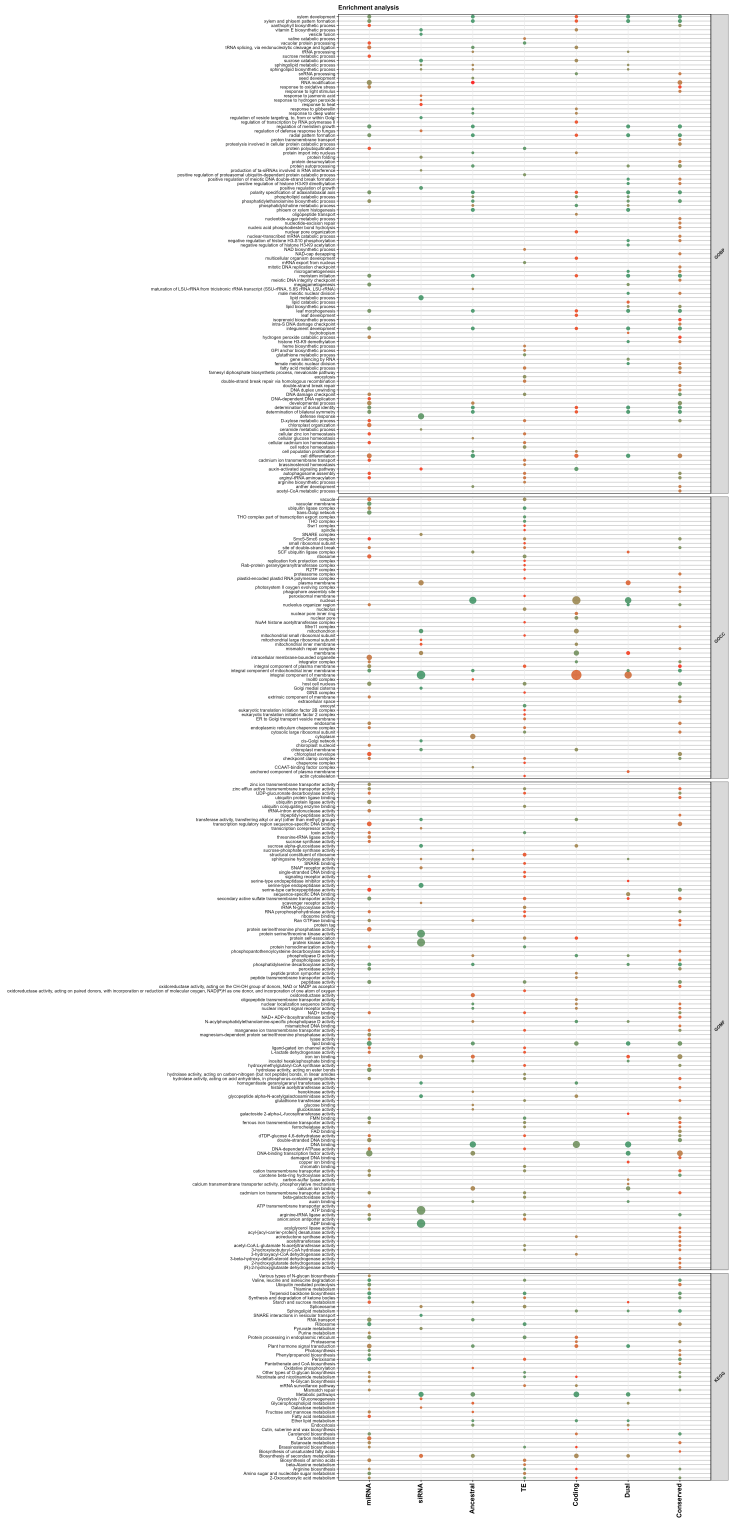

Enrichment analysis
mRNA
siRNA
Ancestral
TE
Coding
Dual
Conserved
Pvalue
Count

Figure SF5: Functional enrichments of target genes by origins of iRNAs. The origins of iRNAs include miRNA precursors, siRNA precursors, ancestral genomes, transposable elements (TE), coding regions, dual identity precursors, and iRNAs that are conserved in miRBase and literature reporting wheat iRNAs. Enrichment analyses for overrepresented GO terms and KEGG pathways were performed based on the standard hypergeometric distribution test for the terms covered by the target genes of each RNA group. Terms supported by FDR ≤ 0.05 were considered enriched.

id% >	eval <	# blast hits for all XLOCS	retention rate (%)	# select monocots	retention rate (%)	# XLOC annotated by Uniprot	% over 151691 XLOCs
0	max	2640882	100.00	-	-	-	-
50	**e-4**	**2424358**	**91.80**	**1060910**	**40.17**	**123559**	**81.45**
60	e-6	2180746	82.58	-	-	-	-
80	e-6	1677091	63.50	991611	37.55	119293	78.64
80	e-4	1693830	64.14	-	-	-	-

Figure SF6: The retention rates of blast hits in relation to filters. With the %identity and e-value thresholds tightening from 50/e-4 to 80/e-6, the resulting annotation rates were similar (81.45% down to 78.64%). In order to maximize cross-reference annotations, we chose 50/e-4 based on literature recommendations [44, 19].

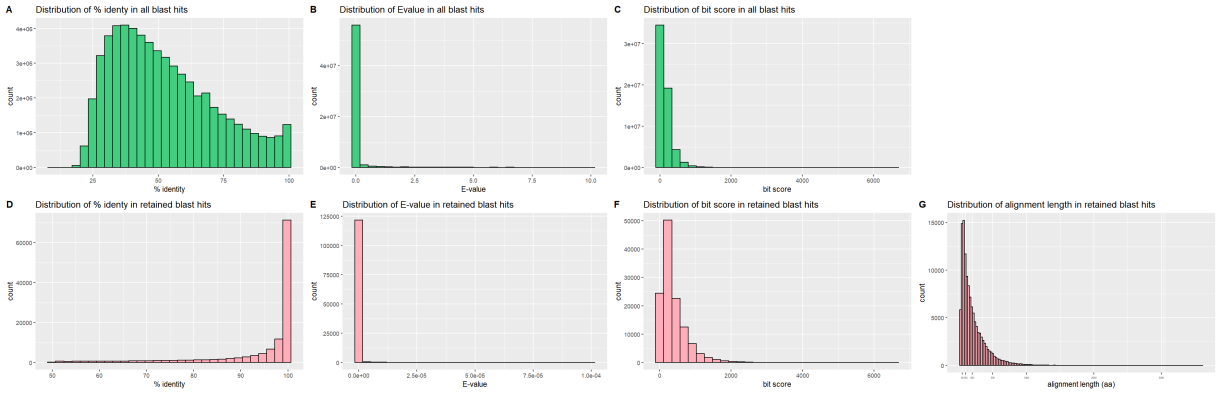


Figure SF7: (a-c) Statistics of all blast hits of XLOC sequences. (d-g) Statistics of the blast hits retained for XLOC annotation after selection: best 20 hits, $> 50\%$ identity, e-value < 0.0001, monocots. The %identity shifted towards 100 in (d), e-value had little impact in (e) and the bit score was slightly improved in (f). The majority of alignment lengths of the retained hits were longer than 50 amino acids in (g). As a result, the selection steps enhanced the information available for each XLOCs.

B Supplementary tables

B.3 Supplementary tables (legends only)

Legends of supplementary tables are as follows. For the files of these supplementary tables, please go to `https://drive.google.com/drive/folders/1AZHVy5106FnMBTuNL7uqGflaoVSpHjFe?usp=sharing`

1. {S1} Overview of 80 wheat sRNAseq field samples

File: [S1_sRNAseq_field_sample_info.xlsx]
Legend: Sequencing metadata and library sizes are shown. Sequencing metadata, including wheat cultivars, sampling time points, soil and air temperatures, were taken from Li *et al,* 2018 and columns "Calculated small RNA" and onwards were appended. Read counts of original sequencing data were in column "Number of Reads". Reads that have low quality scores were removed as recommended for illumina sequencing, and updated read counts after removal were in column "Calculated small RNA". Number and percentage of reads mapped to the wheat genome were recorded in "reads with at least one reported alignment" and "% reads with at least one reported alignment".

2. {S2} Small RNA expression read counts in 80 libraries

File: [S2_readcounts.xlsx]
Legend: Sheet 1: Raw read count data of sRNAs were retained if they passed the selection criteria for miRNA or siRNA embedded in the bioinformatics pipelines. Sheet 2: Normalized read count of sRNAs as parts per 100 millions (pp10m).

3. {S3} Publicly available small RNA databases/predictions used in this study for local mirFinder analysis

File: [S3_smallRNAdatabases.xlsx]
Legend: The small RNA databases in this file include (one database a sheet, total 7 sheets): The miRNA and siRNA databases of literature collection in year 2015 (3598 wheat unique mature miRNAs) and in year 2020 (8831 wheat unique mature miRNAs), IWGSC (14008 wheat unique mature miRNAs), PSU Plant small RNA genes catalog (439 wheat unique mature miRNAs, 57241 wheat unique mature siRNAs) and the full list of ShortStack predictions on our 80 wheat libraries (753 unique mature miRNAs, 409 unique mature siRNAs). The unique mature sequences were indexed to be searched by a local version of *microRNA Finder* [Remita et al., 2016]. The details of the wheat miRNA collection from Rahmdani *et al,* 2020 are included (db2, 10580 entries containing 8831 unique miRNA mature sequences).

Summary of databases

	Nb miRNA (wheat Tae)	Nb siRNA
miRBase v22	122 (All species 48860)	0
Literature 2015 (Agharbaoui *et al*)	3598	0
Literature 2020(Rahmdani *et al*)	8831	0
IWGSC v1.0	14008	0
PSU catalog	439	57241
ShortStack on our libraries	753	409

4. {S4} Raw outputs of microRNA Finder

Files:
[S4a_WMP_miRBaseHit_result_alignment.txt]
[S4b_local_mirFinder_outputs.xlsx]
Legend:
File a is the raw output from web portal WMP *microRNA Finder* [Remita et al., 2016] analyzing predicted iRNAs against miRBase v22. Default options were used: Min sequence length 18, Min identity percentage 80, Min query coverage 0.9, Min subject coverage 0.9, Max of gaps 0, Max of total mismatches 3. A parser script reports from this file the hits with mismatches ≤ 2.
File b keeps the sheets of the raw outputs from local miRFinder. We used the local version of miRFinder to analyze databases other than miRBase, including literature collections in 2015 and 2020 [Agharbaoui et al, 2015; Rahmdani et al, 2020], and collections based on ShortStack (psuss, unique mature miRNAs and siRNAs) [Axtell 2013; Lunardon et al, 2020]. The local output format is different from the portal.

5. {S5} Precursors

File: [S5_precursors.xlsx]
Legend: For each precursor of the miRNAs and siRNAs, the genomic coordinate, relative position of the small RNAs, the dot-bracket notation of the the RNA secondary structure, siRNA phasing score or miRCheck result are recorded in **Sheet 1**. Our predicted miRNAs and siRNAs are clustered in **Sheet 2** based on their genomic coordinates, represented by the CONTIG they are located in. Our predicted miRNAs and siRNAs are clustered in **Sheet 3** based on XLOC precursors they originated from. **Sheet 4** extracted the precursors of TASiRNAs from Sheet 1.

6. {S6} precursor blastn results against nr database for similarity search.

Files:
[S6a_blastN_precursor_alignment.txt]
[S6b_precursor_blastn_report.tsv]
Legend:
In table (a), Raw text output of blastN.
In table (b), Parsed raw text to an organized table.

7. {S7} Expression analysis by DEseq2

File: [S7_Expression_analysis_by_DEseq2.xlsx]
Legend: DEseq2 is a statistical package for analyzing differential expression. Raw read counts and their metadata (cultivar, year, time point and replicate) were used as input for DEseq2. DEseq2 normalized read frequencies based on the input read frequency population. We considered time point 1 of each experiment (for each year and each cultivar) as control, and compared time points T2,T3, T4 and T5 with time point T1 of the same experiment. We computed the log2 fold change (log2fc) and the significant difference. If the absolute log2fc was greater than 1 (ie. abs(log2fc) ≥ 1) and the significance value was smaller than 0.05 (adjusted pvalue considering FDR ≤ 0.05), the value was retained, otherwise the value was annotated as NA. The log2fc was recorded in **Sheet 1 DESeq2_L2fc_filtre_1**, and the

pvalue was recorded in **Sheet 2 DESeq2_pvalue_filtre_.05**. The sequences that had reported values were the sRNAs differentially expressed between later time points (T2, T3, T4 and T5) and time point T1 in each year (2010 and 2013) and each cultivar (Norstar, Iso8S, Manitou and Iso11W), respectively. **Sheet 3** keeping_consistent_replicates. Inconsistent replicates were removed from Sheet 1. There are two biological replicates for each time point by the design of this study. The inconsistent cases of a RNA sequence between replicates were identified by the gauge formula (see methods). We replaced the log2 fold change values of these cases by zero in DESeq2 output (from Sheet 1 to Sheet 3). The results of differentially expressed miRNAs and siRNAs from this table were used for further analysis in this study. **Sheet 4** Retention rate after removing inconsistent cells. Comparing replicates r1 and r2 using the gauge formula, 20 pairs of the replicates (40 libraires) were not consistent and were removed entirely. As a result, Iso11W T3 and T4 and Iso8S T5 did not have usable expression data in both 2010 and 2013. The retention rates for the remaining 20 pairs (40 libraries) ranged from 96% to 31%.

8. {S8} Identification for cold regulated miRNAs and siRNAs.

File: [S8_cold_regulation_of_RNAs.xlsx]
Legend: **Sheet 1** (profiles of each RNA) is the expression profiles of 8 winter-spring comparisons for each candidate RNA. See below for more details how the comparisons were made. **Sheet 2** (unique winter-spring profiles) Initially 735 iRNAs were predicted. After removal of random phasing, 339 iRNAs were retained for further analysis. The 735 iRNAs generated 127 unique winter-spring profiles and the 339 iRNAs generated 102 profiles. **Sheet 3** (CR profile2quote) The cold response profiles were converted to a simple quote denoting the role in cold response (ie, Dw for down-regulated, Up for up-regulated). This analysis allowed us to identify the differentially expressed RNAs as cold responsive if they expressed at least partially coordinately between winter and spring cultivars. To explain with a few examples: Based on the expression profiles, we considered an iRNA that was expressed coordinately (eg. up profile [0u,0u,uu,uu] or down profile [00,0d,dd,d0]) as a cold response candidate; not-regulated (eg. profiles [00,0u,0u,u0] or [00,00,00,00]) as a cold-unresponsive and constitutively expressed iRNA. **Sheet 4** (RNAs in each CR profile) is the list of miRNAs and siRNAs in each profile. Note that certain RNAs were removed later by expression filters. **Sheet 5** (CR profile of iRNA) is the reverse key-value list of sheet3, showing the profile of each cold response iRNA. **Sheet 6** (cold regulation of iRNAs) is based on the profile quote analysis for each miRNA and siRNA in the final list reported in this study. All the 339 miRNAs and siRNAs were listed. Based on these profiles and quota we assigned the regulation category to each iRNA: **cold response (CR) category**, which can be **upregulated (CRup)** or **down regulated (CRdw)**. The sequences that are not in these categories are considered **constitutive (Cstv)** and unresponsive to cold. **Sheet 7** (clustering of cold response iRNAs) The profiles of the 66 cold response iRNAs were clustered into 8 groups based on their winter-spring expression patterns.

Additional description for Sheet 1: Comparison of expression profile in four pairs of winter and spring wheat (Norstar_Manitou, Norstar_Iso8S, Iso11W_Manitou, Iso11W_Iso8S) in four time points. Input matrix is S3b which is restricted by abs (log2 Fold Change)>= 1 and padj (adjusted P value) =< 0.05. Our objective is finding the log2 FC profiles by comparing the winter and spring cultivars, in 4 time points (T2/T1, T3/T1, T4/T1, T5/T1) in each year. In each profile when we compare the winter and spring cultivars, always the first sign (Up (p), Down (n) and unchanged (0) log2) is for winter cultivar (Norstar or Iso11W) and the second one is for spring cultivar (Manitou or Iso8S)

•uu: Both log2 FC are positive

•ud: log2 FC for cultivar N is positive, log2 FC for cultivar M is negative

•du: log2 FC for cultivar N is negative, log2 FC for cultivar M is positive

•dd: Both log2 FC are negative

•u0: log2 FC for cultivar N is positive, log2 FC for cultivar M is unchanged

•d0: log2 FC for cultivar N is negative, log2 FC for cultivar M is unchanged

•0u: log2 FC for cultivar N is unchanged, log2 FC for cultivar M is positive

•0d: log2 FC cultivar for N is unchanged, log2 FC for cultivar M is negative

•00: Both log2 FC are unchanged

The profiles are from these comparisons:
Norstar_Manitou, Norstar_Iso8S, Iso11W_Manitou, Iso11W_Iso8S in 4 time points and for each year separately. A few examples of the observed 34 profiles in this study are [00,00,00,dd], [00,00,00,uu], [00,00,0d,dd], [00,00,0u,uu] and [00,dd,00,00].

We defined a short_profile for each profile base on the below definition:
uu:: **U**
u0 or ***0u***: ***u***
dd: **D**
d0 or **0d**: **d**
00: **-**

For quote_profile we counted the ***u*** and ***U*** together also ***D*** and ***d*** together.
In brief, all profiles are **winter-spring coordinate (or cold response profiles)**. The 34 profiles and their annotations are as follows.

profile	short_profile	quote_profile	Coor_NoCoor
00,00,00,dd	-,-,-,D	D1	Coor
00,00,00,uu	-,-,-,U	U1	Coor
00,00,0d,dd	-,-,d,D	D2	Coor
00,00,0u,uu	-,-,u,U	U2	Coor
00,dd,00,00	-,D,-,-	D1	Coor
00,dd,d0,00	-,D,d,-	D2	Coor
00,dd,d0,0d	-,D,d,d	D3	Coor
00,uu,u0,00	-,U,u,-	U2	Coor
00,uu,u0,0u	,-,U,u,u	U3	Coor
0d,00,0d,dd	d,-,d,D	D3	Coor
0d,dd,00,00	d,D,-,-	D2	Coor
0d,dd,00,0d	d,D,-,d	D3	Coor
0d,dd,d0,00	d,D,d,-	D3	Coor
0d,dd,d0,0d	d,D,d,d	D4	Coor
0u,00,0u,uu	u,-,u,U	U3	Coor
0u,uu,00,00	u,U,-,-	U2	Coor

0u,uu,u0,00	u,U,u,-	U3	Coor
0u,uu,u0,0u	u,U,u,u	U4	Coor
d0,00,0d,dd	d,-,d,D	D3	Coor
dd,00,00,00	D,-,-,-	D1	Coor
dd,00,00,d0	D,-,-,d	D2	Coor
dd,00,0d,dd	D,-,d,D	D3	Coor
dd,0d,00,00	D,d,-,-	D2	Coor
dd,0d,00,0d	D,d,-,d	D3	Coor
u0,00,0u,uu	u,-,u,U	U3	Coor
uu,00,00,00	U,-,-,-	U1	Coor
uu,00,00,0d	U,-,-,d	U1D1	Coor
uu,00,00,0u	U,-,-,u	U2	Coor
uu,00,00,u0	U,-,-,u	U2	Coor
uu,00,0u,00	U,-,u,-	U2	Coor
uu,00,0u,uu	U,-,u,U	U3	Coor
uu,0u,00,00	U,u,-,-	U2	Coor

9. {S9} Expression analysis of cold response iRNA

File: [S9_Expression_analysis_ColdResponse_iRNA.xlsx]
Legend: The cold response iRNAs were extracted from table S7 for reading convenience. In this table, **sheet 1** showed the log2fc of 66 cold response iRNAs (extracted from S7, sheet3); their corresponding pvalues in **sheet 2** and their corresponding fold changes in **sheet 3**. As described in S7 and S8, the cold response iRNAs that had reported values were the sequences differentially expressed between later time points (T2, T3, T4 and T5) and time point T1 in each year (2010 and 2013) and each cultivar (Norstar, Iso8S, Manitou and Iso11W), respectively.

10. {S10} psRNAtarget webtool output results

File: [S10_psRNAtarget_combined_outputs.txt]
Legend: The target genes of the 339 iRNAs were predicted using psRNATarget from the total RNA transcripts of the same 80 wheat samples (denoted as XLOCs).

11. {S11} Target annotation table

File: [S11_iRNA_target_annotations.tsv]
Legend: The functional annotations of the iRNA targets were extracted from the XLOC database.

12. {S12} XLOC annotation database

[S12_field_homolog_annotations.tsv]
Legend: The wheat mRNA transcripts -- or XLOCs -- collected from the same field samples for sRNA analysis were used to annotate precursors in this study.

13. {S13} spearman correlation analysis of all iRNA-targetXLOC pairs

File: [S13_spearman_correlation_of_all_iRNA_targetXLOC_pairs.xlsx]

Legend: Spearman correlation analysis of all the pairs of iRNA-targets was conducted, and significant correlation results with p-values greater than 0.05 were shown.

14. {S14} Kinases targeted by cold-response iRNAs and the literature supporting the role of kinases in cold response

File: [S14_CR_iRNA_targeted_kinases_and_their_literature.xlsx]
Legend: **Sheet 1** (targeted kinase): All the kinases targeted by the 66 cold-response iRNAs are shown. Among 66 cold response iRNAs, 13 iRNAs are targeting 63 kinases (unique Uniprot IDs). The expression correlation between the iRNA and its target in 80 libraries was evaluated by the spearman correlation coefficient. We set the correlation strength to four levels by their spearman correlation coefficient's (co) absolute values: strong (1 > co >= 0.7), moderate (0.7 > co >= 0.5), fair (0.5 > co >= 0.2) and weak (0.2 > co > 0). Note that "pvalue = -1" indicates the spearman correlation could not be computed. XLOC with a positive sign (+) is the precursor. XLOX with a negative sign (-) is the complementary strand of precursor. XLOC with (number) is the target transcript and the number indicates the expectation value of psRNATarget. **Sheet 2**: Keep the lines with 0 <= p <= 0.05. The pvalue evaluates the significance of the correlation strength between the iRNAs and their target kinases (XLOC long RNA expressions). The pvalues smaller than 0.05 are generally considered significant. **Sheet 3** (reduce columns and descriptions) Only the lines of the unique pair of the iRNA and the Target Uniprot ID were kept. The column "Description" was manually edited to remove redundant phrases. Some columns were removed for concise presentation. A new column was appended to the sheet to indicate if the **highest** expressions of the target among the 80 libraries were greater than **50 reads** in one library. XLOCs with low expression levels were removed (cite see below). **Sheet 4** (final presentation) The kinases targeted by cold-response iRNAs with high confidence, and a list of literature supporting the role of kinases in cold response were shown in the table. **Sheet 5** (XLOC sequences of kinases) The XLOC sequences of the kinases targeted by the cold response iRNAs.

15. {S15} Transcription factors/DNA-binding activities targeted by cold-response iRNAs and the literature supporting the role of transcription factors in cold response

File: [S15_CR_iRNA_targeted_transcriptionfactors_and_their_literature.xlsx]
Legend: Similar to S14 with simplified presentation. **Sheet 1** (targeted DNA binding). **Sheet 2** (2) 0 <= p <= 0.05). **Sheet 3** (citation). **Sheet 4** (reduce and format). **Sheet 5** (XLOC sequences of DNA binding).

16. {S16} iRNA annotations, all information

File: [S16_iRNA_annotations_all_info.xlsx]
Legend: This a wide table containing all the annotations we added to each iRNA. **Sheet 1** is the main table. **Sheet 2-the last** are subtables extracted from Sheet 1 for the cold response, the spring-habit specific, the winter-habit specific, the Manitou-background specific, the Norstar-background specific, the Iso11W-specific, the Iso8S-specific, the Manitou-specific, the Norstar-specific, the differentially-expressed but not cultivar-specific, the dual identity iRNA subsets and tasiRNAs, respectively.

Column O, 6 species are: Hits were retained if they matched wheat Triticum aestivum or monocots Oryza sativa, Hordeum vulgare, Arabidopsis thaliana, Brachypodium distachyon and Zea mays miRNAs.

Because miRBase does not include all known and published miRNAs and siRNAs, we searched in blastn nr database for precursor hits, literature publications and Penn State small RNA's wheat database for wheat miRNAs on our 80 libraries to expand the known miRNA and siRNA annotations. For the same objective, we also ran ShortStack (a software used by Penn State Database) to predict miRNA and siRNA on our 80 libraries. MiRNA and siRNA precursors were searched using blastn in the nr database (nucleotides database from all kingdoms). Hits containing keywords miRNA and siRNA were extracted and annotated the query precursors as known. We also annotated the hits coming from selected species. The keywords and species are as follows. **Keywords** = miRNA, microRNA, MIR, siRNA, TAS, TASiRNA, SiRNA, VRN, cold-responsive, stress, piRNA, mirtron, ncRNA, mRNA, tRNA, rRNA, snoRNA, virus. **Species** = Triticum, aestivum, Brachypodium distachyon, Hordeum vulgare, Oryza sativa, Arabidopsis thaliana, Zea mays. PSU https://plantsmallrnagenes.science.psu.edu/genomes.php?genomes_id=3

Column names with "(**mirfinder**)" meaning the query sequences are annotated against the database using mirfinder. Mirfinder parameters are t_len=18, t_iden=80, t_qcov=0.9, t_scov=0.9, t_gap=0, t_totMis=2
For miRFamily hits, it is searched against mirBase.v22 using mirfinder (wheat.bioinfo.uqam.ca/) and only 2 mismatches are allowed. Some sequences find more hits using miRBase (www.mirbase.org) but they might not satisfy our stringent hit criteria.

A list of keywords have been searched in the target keywords. For example, "**isTargeting**HeatShockProtein'' shows that the targets contain the keyword "Heat shock protein" and the string variant "HSP". The keywords are: kinase, sucrose (glycolytic), dehydrin (DHN), Rab protein, DREB, heat shock protein, ice crystallization protein, IBP, TaVRT2, VRN, SOC, CO, cor413, cor410, WCS120, C-repeat/DRE/LTRE-binding protein, C-repeat/dehydration-responsive element, CRT/DRE, CRT-Binding Factor, LTRE-binding protein, chitinase, CBF, Vernalization genes, Mads-Box, phosphatase, aquaporin, HSP, Auxine response factor, ABA signaling, DEAD-BOX, LIPOXYGENASE, GIBBERELLIN OXIDASE, GLUTATHIONE PEROXIDASE, SUPEROXIDE DISMUTASE, WCOR, ARF, GAMYB, THAUMATIN-LIKE PROTEIN (TLP). The full list of the variants for the keywords is not shown. Note that the enrichment showed that this phrase of GOMF term "protein kinase activity" was enriched in siRNA population and CR/CRdw population. On the other hand, **isTargetingKinase** used a tolerant keyword "kinase"in table S14. As a result, an expanded pool of functional descriptions containing kinase were flagged. Example phrases are protein serine/threonine kinase activity, protein kinase binding, pyruvate kinase activity, phosphatidylinositol phosphate kinase activity, aspartate kinase activity, etc.

17. {S17} Alignment and phasing potential of TAS precursor CONTIGs

File: [S17_TAS_contigs.pdf]
Legend: Alignment and phasing potential of representative TAS precursor CONTIGs in this study are shown. The TAS precursor CONTIGs were identified using blastn. For each CONTIG, the panel (a) showed the nucleotide counts on the 3-prime and 5-prime extremities of the 21- and 22-bp sRNA species. (b) showed the size distribution of the sRNA species on the CONTIG. (c) showed the information of the CONTIG, including the begin (deb), end (fin) of the phasing segment, length of the segment, maximum phasing cycles observed, accumulated phasing score on the segment. Strand ratio and dominant strand

indicated that the phasing RNAs were located on one strand or both strands of the precursor. The precursor was considered double-stranded if the ratio was smaller than 0.8; otherwise, single-stranded. (d) showed the read counts of sRNAs of all sizes or only the 21-nt species aligned on a given position on the positive strand of the reference CONTIG, when counting was based on the 5-prime position of the sRNAs on the positive strand of the reference, or plus 2 nts on the negative strand, honoring the Dicer stagger cuts. The NewIDs of the iRNAs originating from the CONTIG were labeled at the corresponding positions above the plot. The potential cleavage sites of miR390 were pointed by a triangle. (e) showed the H-score values on each position of the CONTIG, as well as the sum of k values that were used in the H-score formula.

Summary of TAS CONTIGs

	CONTIG	Nb of miR390 clevage sites	blastn	Max phasing cycles	Length (nt)	Originating strand
1	6D__285740000	2	Tae TAS3b	8	232	+-
2	6B__446720000	1	Tae TAS3b	5	77	-
3	6B__446715000	2	Tae TAS3b	7	232	+-
4	6A__406360000	2	Tae TAS3b	8	231	+-
5	4D__316535000	0	Plant TAS3a	8	282	+-
6	4A__158175000	2	Plant TAS3a	8	273	+-
7	2D__450395000	2	Tae TAS3b	7	169	-
8	2A__598975000	2	Tae TAS3b	7	130	+-

18. {S18} miRNA precursor structure visualization of a representative library

File: [S18_precursorindex.pdf]
Legend: “precursorindex.html” is one of the output files mirLibSpark generates. It shows the predicted miRNA mature sequence, coordinates in the genome, precursor sequence and its secondary structure in dot-bracket notation and visualization. Here such a file of a representative library in this study was shown.

19. {S19} iRNA annotations of the 339 iRNAs

File: [S19 iRNA annotations of the 339 iRNAs.pdf]
Each iRNA sequence was given a NewID. For each iRNA, the following annotations were shown: the miRBase family, the chromosomes of the precursor loci, the functional description of precursors, the target XLOC IDs identified by psRNATarget, the functional description of the target XLOCs, the cold response regulation, our prediction as a miRNA or a siRNA, and the iRNA subcategories. iRNA subcategories were A: miRBase Tae miRNA, B: miRBase monocot miRNA, C: miRBase all miRNA, D: Literature Tae MiRNA 2015, E: Literature Tae MiRNA 2020, F: ShortStack (and PSU) wheat miRNA and siRNA, G: blastn miRNA and siRNA precursor and H: novel. Known mature sequences were step-wisely put aside from category A to category G until not found in any referenced databases and assigned as A to H.

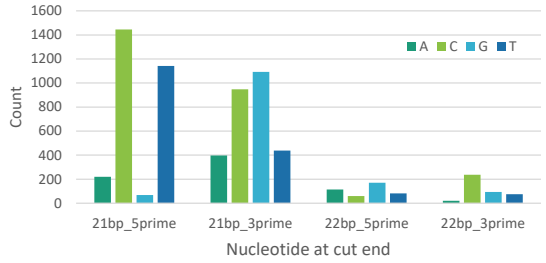
TAS contigs identified by blastn

1. 6D__285740000_285745250
2. 6B__446720000_446725250
3. 6B__446715000_446720250
4. 6A__406360000_406365250
5. 4D__316535000_316540250
6. 4A__158175000_158180250
7. 2D__450395000_450400250
8. 2A__598975000_598980250

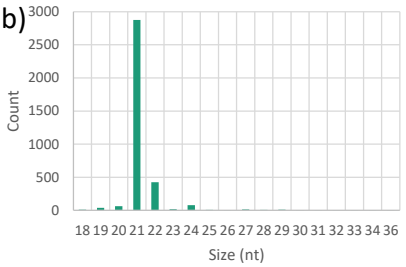
1

1. 6D__285740000_285745250

(a)



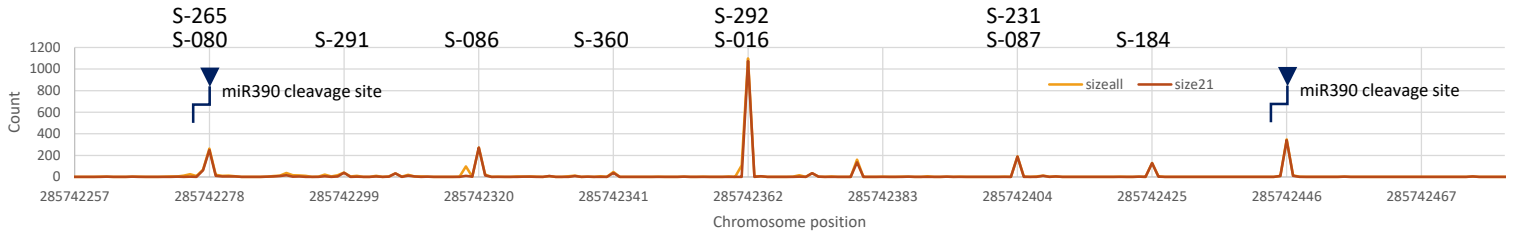
(b)



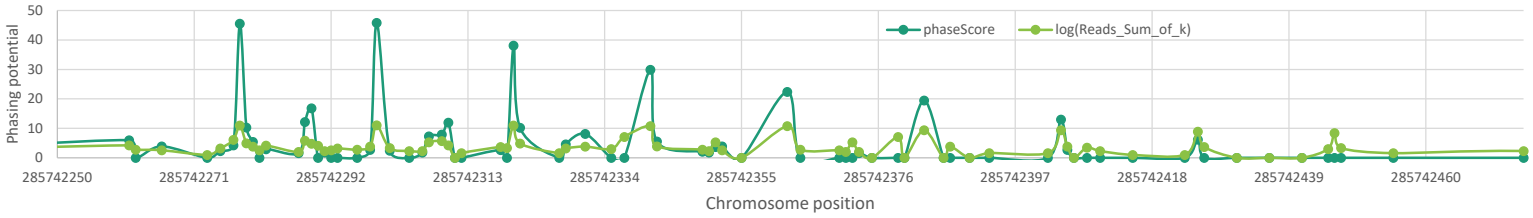
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
285742193	285742425	232	8	45.8	667.7	0.68	+-

(d)



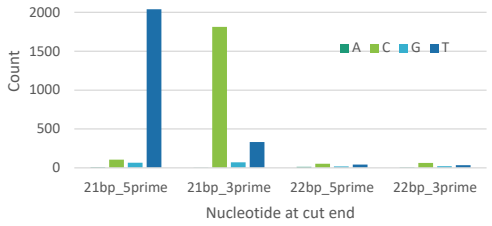
(e)



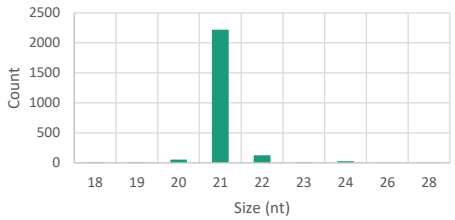
2

2. 6B__446720000_446725250

(a)



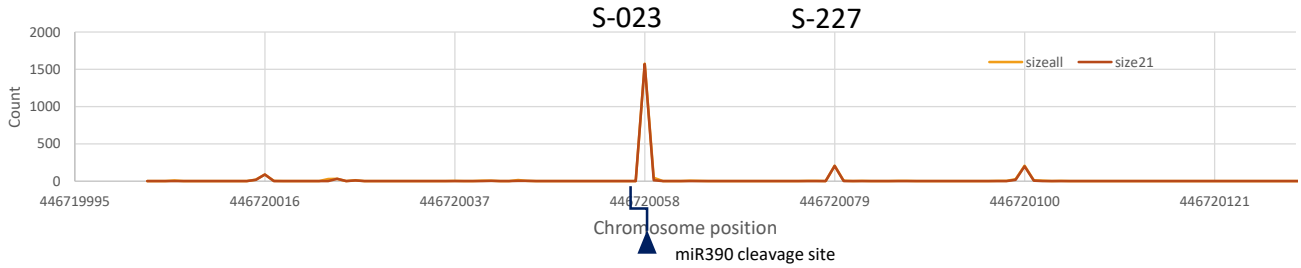
(b)



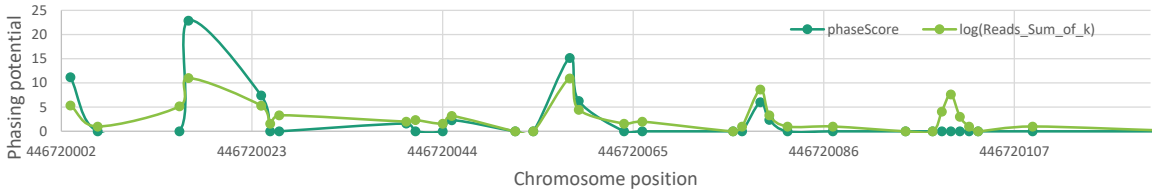
(c)

Deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
446720003	446720080	77	5	22.9	121.7	0.19	-

(d)



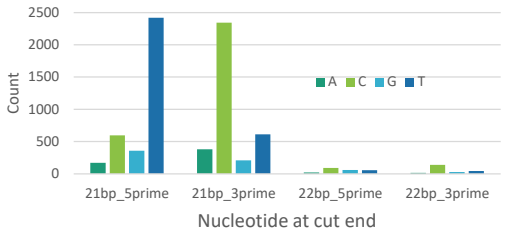
(e)



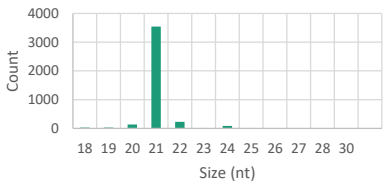
3

3. 6B__446715000_446720250

(a)



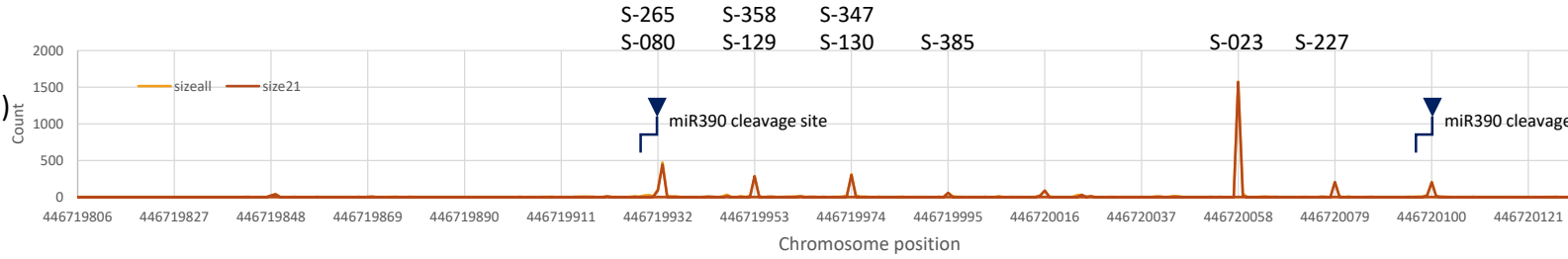
(b)



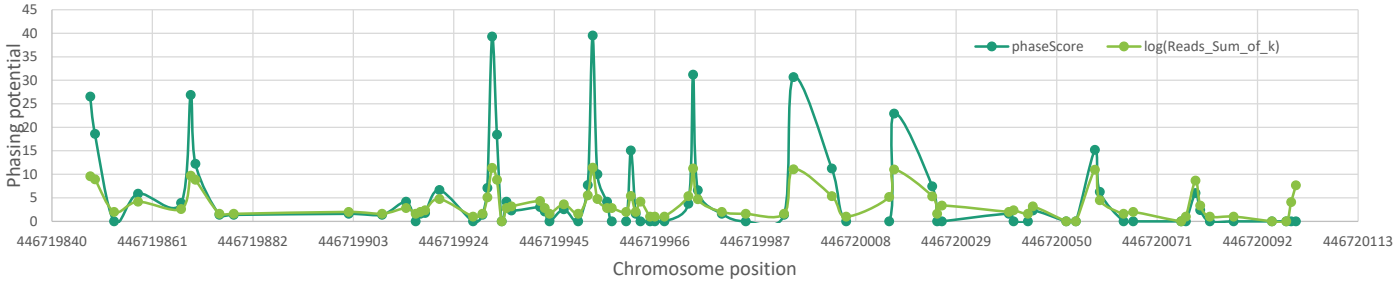
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
446719848	446720080	232	7	39.5	670.8	0.29	+-

(d)



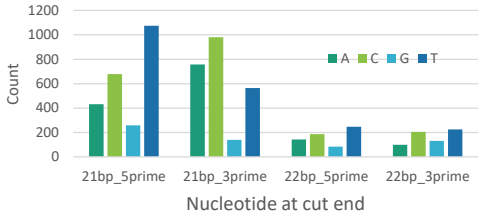
(e)



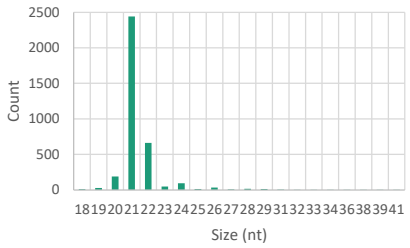
4

4. 6A__406360000_406365250

(a)



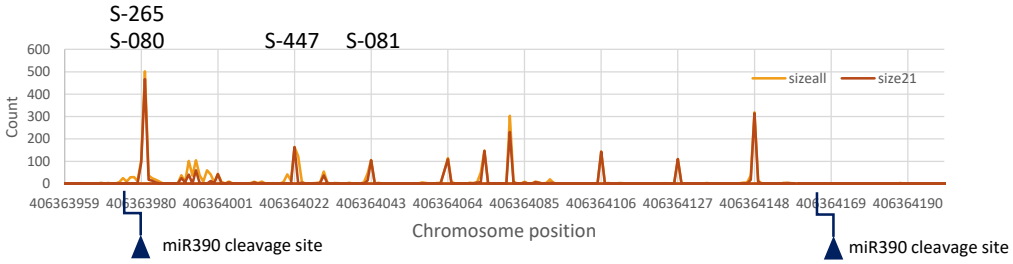
(b)



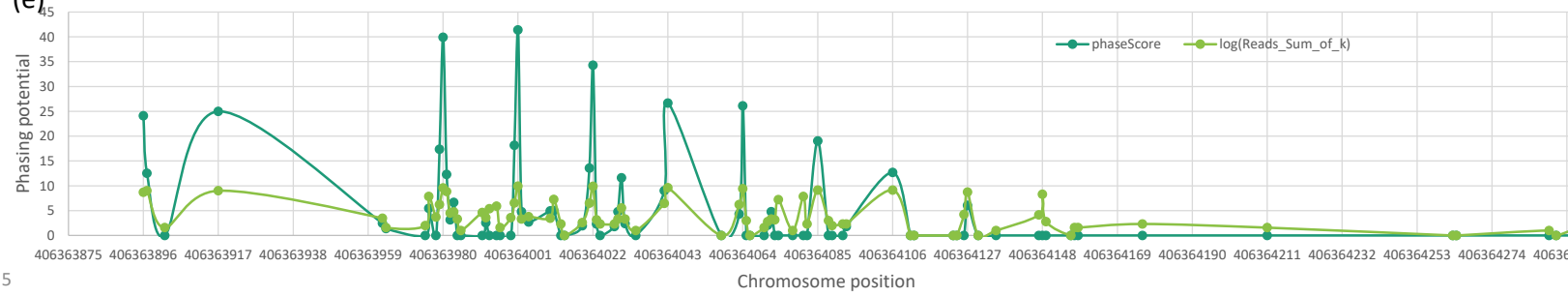
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
406363896	406364127	231	8	41.4	687.1	0.37	+-

(d)



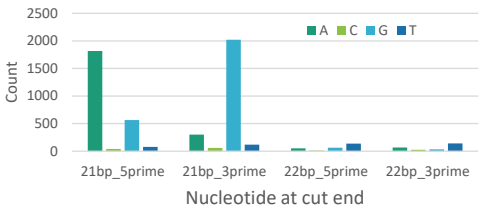
(e)



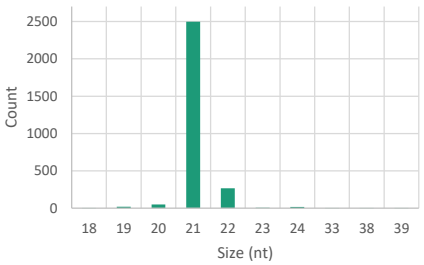
5

5. 4D__316535000_316540250

(a)



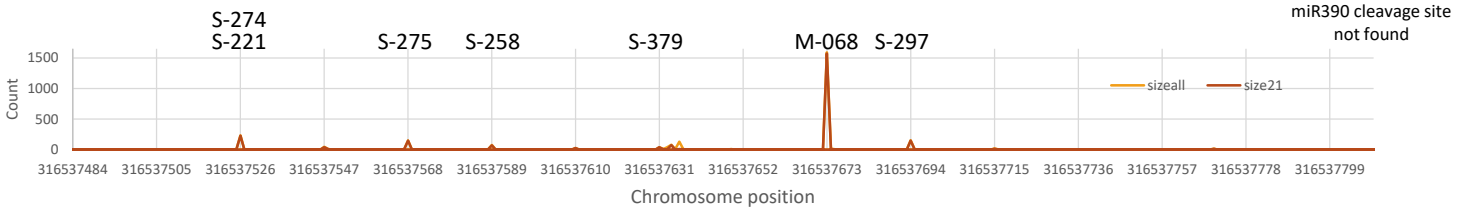
(b)



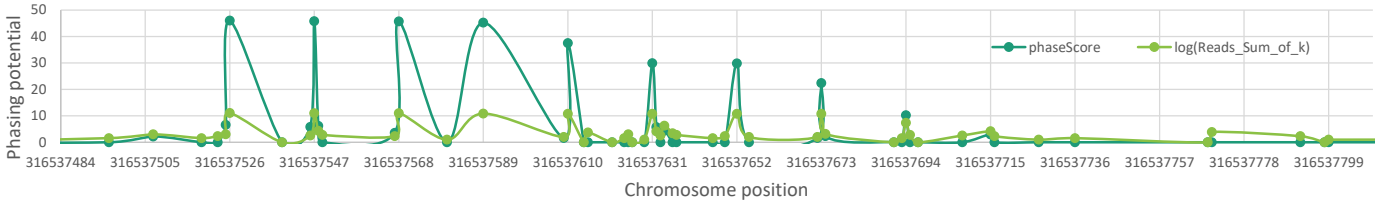
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
316537433	316537715	282	8	46	612.8	0.22	+-

(d)



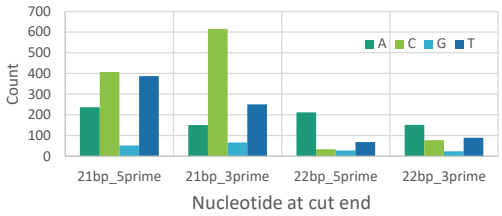
(e)



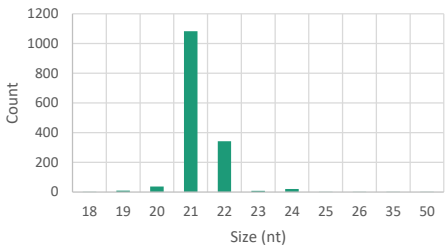
6

6. 4A__158175000_158180250

(a)



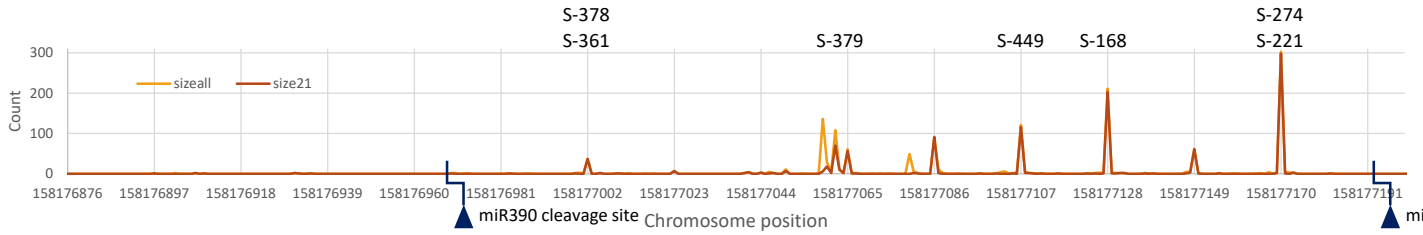
(b)



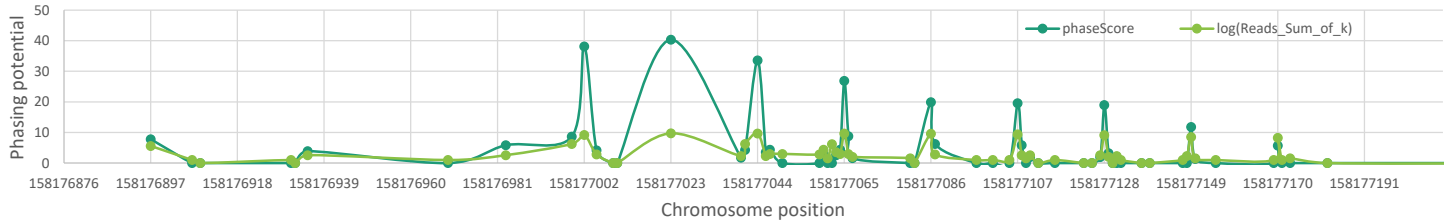
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
158176897	158177170	273	8	40.4	420.3	0.46	+-

(d)



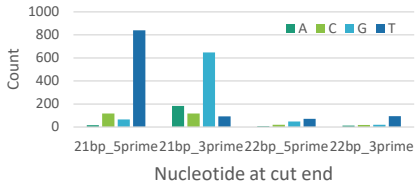
(e)



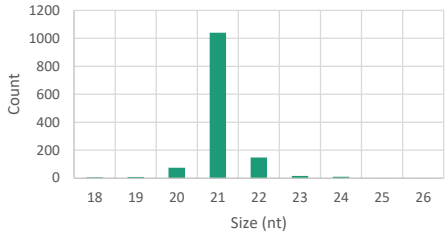
7

7. 2D__450395000_450400250

(a)



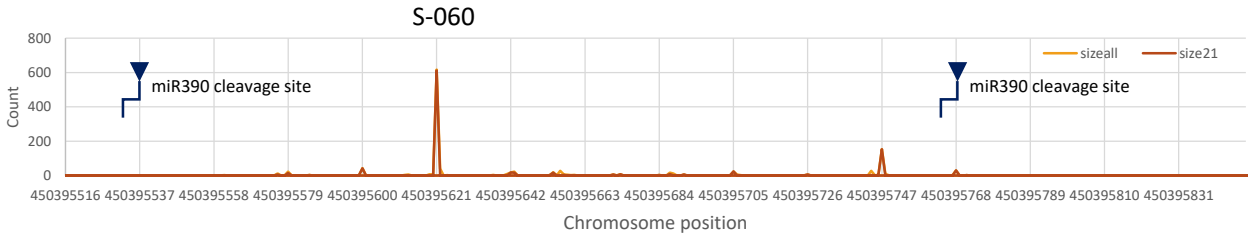
(b)



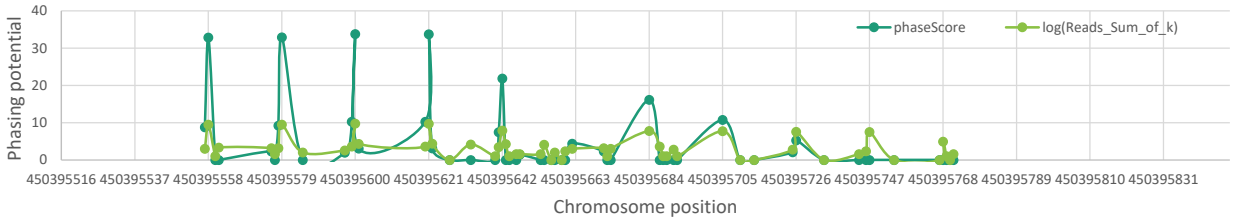
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
450395557	450395726	169	7	33.8	410.8	0.19	-

(d)



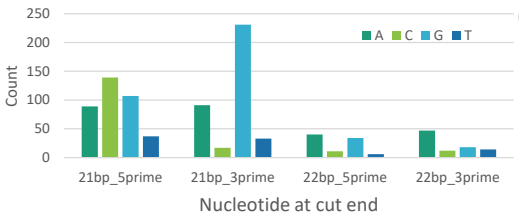
(e)



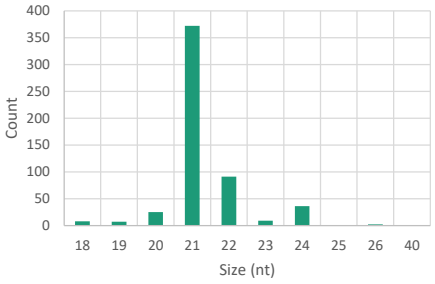
8

8. 2A__598975000_598980250

(a)



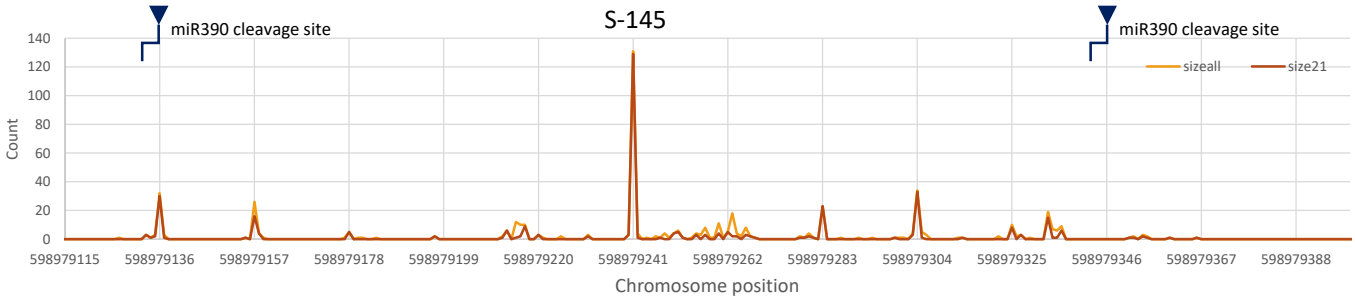
(b)



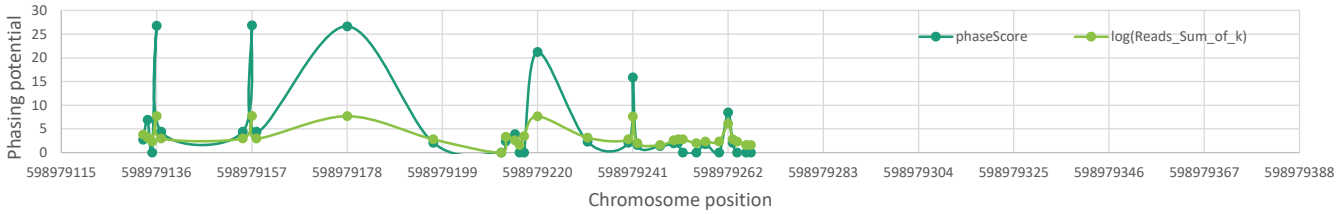
(c)

deb	fin	length	maxnbcycle	maxphasescore	accphasescore	strandRatio	domstd#
598979133	598979263	130	7	26.9	298.5	0.35	+-

(d)

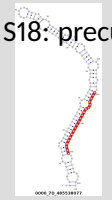
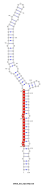
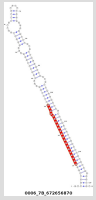
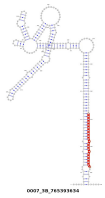


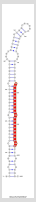
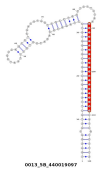
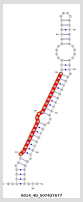
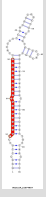
(e)



9

miRNAs and their genomic precursors

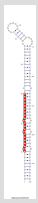
Serial	pre-miRNA.Structure.Visualization	miRNA.Sequence & Coordination / pre-miRNA.Sequence / pre-miRNA.Structure	
B.5 S18: precursor index			
0000		AAAAATTGCGGATTTTTGTCT	Chr 7D:485538077 [+]
		CGCTATTGAATGATCAAGGAATTTGTGAGGTTTACATTATCTTTGCAAGAGTTGTATTGTTCAGTTAGACCTAGACATGCAAGTATATTTCTAGGCGAAAATATATT GTTAAAAATTGCGGATTTTTGTCTCTCAAGTTGC	
		.((..((((..(((..(((((((((((..((((((.(((.(((((.((((.((((((... (((((.......))).))..)))))))..)))))..))))...))).)).)))))..)))))))))))))))..)))))...))	
0001		AAGCTCAGGAGGGATAGCGCC	Chr 5A:649565478 [-]
		ACAATCCTTGAAGCTCAGGAGGGATAGCGCCTAGGTTTATTGCAACCAGCAAGAGACCGGCCGGCCTCTTGATGGTTCATCGATCTGTTCGTCCGAAGCGCTATCTAT CCTGAGCTCCACGGTTTGT	
		((((.((.((.((((((((((.((((((((((.((((((...((.(((((.((((((.((....)))))))).)))))))..)))))) (((....)))))))))))))).))))))))).)).)).))))	
0002		AAGCTCAGGAGGGATAGCGCC	Chr 5B:657146708 [-]
		ACAATCCTTGAAGCTCAGGAGGGATAGCGCCTAGGTTTATTGCAACCAACAAGAGAGACCGGCCGGCCCGCCGGCTGGCTCTTGGTGGTTCATCGATCTGTTCGTCC GAAGCGCTATCTATCCTGAGCTCCACGGTTTGT	
		((((.((.((.((((((((((.((((((((((.((((((...((.(((((.((((((...(((((((((....)))))))))))))))).)))))))..)))))) (((....)))))))))))))).))))))))).)).)).))))	
0003		AAGCTCAGGAGGGATAGCGCC	Chr 5D:521092316 [-]
		ACAATCCTTGAAGCTCAGGAGGGATAGCGCCTAGGTTTATTGCAACCAGCAAGAGACCGACCGGCCGGCCGGCCGGCTCTTGATGGTTCATCGATCTGTTCGTCCGA AGCGCTATCTATCCTGAGCTCCACGGTTTGT	
		((((.((.((.((((((((((.((((((((((.((((((...((.(((((.((((((.(((.((((....)))).)))))))))).)))))))..)))))) (((....)))))))))))))).))))))))).)).)).))))	
0004		ACCTGCAGTTGGGCCAATGAC	Chr 3A:608695359 [+]
		CCGTAACTGCACCTGCAGTTGGGCCAATGACATGTGGGCCTAACAGACGGCTGGCCCACATGCCATTGACCCAACTGCAGGTGCAGTTACGG	
		((((((((((((((((((((((((.(((((.((((((((((.............)))))))))).)))))).)))))))))))))))))))))))	
0005		AGAAAGCAATGTCAGAGGAGT	Chr 7A:688193527 [-]
		ATTACTCAAGAGAAAGCAATGTCAGAGGAGTTTTAGCAACAGTCTGAAGAGAAGCTGATTATGTATATGTGAGGAGAGTGTATATTCGGCAGCAGTATATTCGAGGA GAGTGTATATTCAGACTGCTGTTGAAGCTCATTTGGCATTGCTTTCCCTTGAAGTGT	
		..(((((((((.((((((((((((((((.((((((((((.(((((((((.....((((...(((((((((.........)))))))))....)))))). (((((((((.....)))))))))..)))))))))).)))))))))))))))).)))))).)))).	
0006		AGAAAGCAATGTCAGAGGAGT	Chr 7B:672656870 [-]
		ATTACTCAAGAGAAAGCAATGTCAGAGGAGTTTTAGGAACAGTCTGAAGAGAAGCTGATTATTTATATTTGAGGAGAGTTTATATTCAGACTGCTGTTGAAGCTCAT CTGGCATTGCTTTCTCTTGAAGTGT	
		..((((((((((((((((((((((((((.(((((((((...(((((((((...(((((..... (((((....))))...))))))...)))))))))..)))))))))..)))))))))))))))))))))))).)))))).	
0007		AGAAGATTAGAAGATTAAGCA	Chr 3B:765393634 [+]
		CAGTTGGTCTCTCTTAAATTTGTAATCTTCTGGAAATATGCTTAAGTATAATAAATGATGCGAAGGTAGATCATAGAACATGAACACATGGTGTAAAAGAGGCACTC ATGCACCAAAACTATAAAAAGCAGCAAAGAAAATGAGTAAAAGAGGCACTTGACTTGTAGATTATACAAGAGTGATGATCTGGAAGTCAACCCTCTTTTATCAGTAG TTTTGATCTACCTTCTCATCTATAGGATTAGTTATTATGTCTAAGCATGTTTTTAGAAGATTAGAAGATTAAGCAAGCTCATTCG	
		....(((.((..((((((.(((.(((((((((((((((((((((((((.((((((((..((((.((((((((((((.......(((....)))))))))(((... (((...))).))))))))))))))).....................(((.(((((((((...(((((((. (((((((((....)).)))))))))..)))))))..)))))))))))))))))))))))

		TTTATATTTTAGGACTTTTGTCCCTTCGCCGCCGCCGGCACCTCCCATCCATCCCTCAACTCCGCCGCCTCTCTCTCTTCAATTTCAGCCCCCAGCCAACAACCTCCTAGTTCCCAACCGTGTG .((((.((((((((((((((((((.(((((((((.......(((((((((((((((((((((((.(((..(((((((.....((((((((((..(((((((. (((((((((........)))))))).)))))))).)))))))))))))).....)))))).)).(((((((((((((((((((((((.)))))))))).)))))))))))))))))))).)))))))))).)))))))))))))))))))).))))
0010		AGGGTCGAACTGAGAACACAT Chr 3D:523446499 [-] AAGTTCCTGCAGGGTCGAACTGAGAACACATGAGATATATTGACGAGCGACTAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCT ..(((((.((.(((.(((.((((((((((((((((.((..(((.....))).)).)))))))))))))))).))).))).)).))))))..
0011		AGGTGGGAGGTTGTTGGCTGG Chr 5B:130320968 [+] TACGGCTCGGAGGTGGGAGGTTGTTGGCTGGTGGCTGAAATGAAAATAGAGGAGGGAGGGGCGGTGGAGCTGGTTGATGGACATGCGGCGGTGGCGGCAAAGGGGACCAAAGTCTTTATATTTTAGGACTTTTGTCCCTTCGCCGCCGCCGCCACCTCCCATCCATCCCTCAACTCCGCCGCCTCTCTCTCTTCAATTTCAGCCCCCAGCCAACAACCTCCTAGTTCCCAACCGTG (((((...((((.((((((((((((((((((((.(((((((((.......(((((((((((((((((((((((.(((..(((((((.....(((((((((((.. (((((((. (((((((((........)))))))).)))))))).)))))))))))))).....)))))).)).(((((((((((((((((((((((.)))))))))).)))))))))))))))))))).)))))))))).)))))))))))))))))))).))))
0012		AGTTCCAGACGATGCAGGCCT Chr 7B:732433647 [-] ACCTCGAGGAAGCCCTGCATCATTTGGAACTCGCCGGTGCGATAGCACGCAGCATGACCTGGTGTTCATCACCGGCAAGTTCCAGACGATGCAGGCCTTCATCAAGGC .(((.((.((((.(((((((((.(((((((((.(((((((....(((((. (((......)))))))).....)))))))).)))))))).)))))))).))))).)).))).
0013		ATACCACGACTGTCATTAAGC Chr 5B:440019097 [-] TGAAGGCGATGCTTAATGACAGTCGTGGTGTCCTAGGCGAAGACGATGCTTAATGGCCAGCATAGACATAGTCGCCTGGCTCTTCTAAGATACCACGACTGTCATTAAGCATCATCTTCA ((((((.((((((((((((((((((((((((((((((((((..... (((((........)))))........))))))))........)))))))))))))))))))))))))))))))).))))))
0014		ATAGCATCATCCATCCTACCA Chr 4D:507427677 [+] GAATTTGCCCATAGCATCATCCATCCTACCATCCAGCGGCGGCGGCTGCTGCCCCACGATGGGTAGGATGGCCGGTGCTGTTGGAAATATG ..((((.((.((((((((..(((((((((((((....((((((....))))))....))).)))))))))..)))))))).)))))...
0015		ATAGCATCATCCATCCTACCA Chr 4D:507476194 [+] GAATTTGTCCATAGCATCATCCATCCTACCATTCAACGGCGGCTGCTGCTGCCCCACCGGTGGGTAGGATGGCTGGTGCTATGGATAGCTTC (((.((((((((((((((((.((((((((((..((...((((((....))))))).....))..)))))))))).)))))))))))))))).)))
0016		ATAGCATCATCCATCCTACCA Chr 5A:691797378 [-] GAATTTGTCCATAGCATCATCCATCCTACCATTCAATGGTGGCTGATGCTGCCCAACGGTGGGTAGGTGGCTGGTGCTATGGATAGCTTC (((.((((((((((((((((.(((.(((((((..((...((((((....))))))....)).)))))))))).)))))))))))))))).)))
0017		ATAGCATCATCCATCCTACCC Chr 4B:654900524 [+] GAATTTGTCCATAGCATCATCCATCCTACCCTTCAACGGTGGTTGCTGATTCCCCACGGTGGGTAAGATGGTTGATGCTATGGATAACTTC (((.((((((((((((((((.((((((.((((((....((.((((...........)))).)))))))).)))))).)))))))))))))))).)))
0018		ATAGCATCATCCATCCTACCC Chr 6A:21074997 [-] GAACTTGTCCATAGCATCATCCATCCTACCCTACCCTTCAACGGCGGCCGCTGATGCCCCACGGCGGGTAGGATGGCTGGTGCTATTGTTAGCTTC (((.(((...((((((((((.((((((((((.....(((.(((...))).))).(((....)))))))))))))))).)))))))...))).)))
0019		ATCAGGAAGGATGACACCGAC Chr 2B:210254927 [+]

GGGTCCTGTGATCAGGAGAGATGACACCGACGCCGATCAGATGGGTCGGCTTAACCAGGCCGTGCATCCGTCTGTTCGGAGTTGGTGTCACCTCGCCTGAACATAGG
GCTC

((((((((((.(((((.(((.((((((((((.((((.
((((((((((((((....))))))...)))))))))).)))))..)))))))))).)))..))))))..)))))))))))

0020

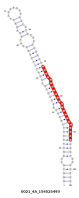


ATCAGGAGAGATGACACCGAC Chr 2D:151985102 [+]

GGGTCCTGTGATCAGGAGAGATGACACCGACGCCGATCAGATGGGTCGGCTTAACCAACCAGGCCATGCATCCGTCTGGTCGGAGTTGGTGTCACCTCGCCTGAACA
TAGGGCTC

((((((((((.(((((.(((.((((((((((.(((((((((((((.
(((((.........)))))....))))))))))))))).)))))))))).)))..))))))..)))))))))))

0021



ATGCCGTGTTGTTCTGAAAGAA Chr 4A:154525493 [+]

GGGTGTCAGTTTCTTCAGAAGAGCATGTCATTTTCTGGAATGAAGGCTCAAAACTGCCATCCTCCCAGAAAATGCCGTGTTGTTCTGAAAGAAACTACCACCC

(((((..(((((((((((((((.(((((((.((((((((((...((.(((........))).))...)))))))))).)))))))..)))))..)))))))))..)))))

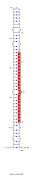
0022

ATGGACCGCGGACAACGAGGAT Chr 1B:459805474 [-]

TATCGTCTTCATCCTCGTTGTCCGCAGTCCATAAAAAAGTATTTTCTGATGAAAACTTTAGGTATGATTCACCCTTACTCTGTTGCTTCATCCATCTTCCATGCAAC
CTGAGAAGTTCTTTCCGTATATCCTCAAAGGTCCTGGAAAGAACTTCTCAGGTTGCATGGAAGATGGATGAAGCAATAGAGTAAGGGTGAATCATACCTAAAGTTTG
CAGCAGAAAATACTTTTTATGGACCGCGGACAACGAGGATGAAGACGATA

((((((((((((((((((((((((((.((((((((((.((((((((((.((.
((.....
(((...)))...)).)).)))))))))))))))))))))))).)))))))))))

0023

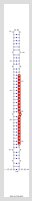


ATTTGTAGTGTTCGGATTGAGTTT Chr 5A:419366109 [-]

CCACTTTTGCAAACTTAGTCTTGGCACTATAAATTTAAAATGTGCATGTGAAATATGCTCATTATAAATTTGTAGTGTTCGGATTGAGTTTATAATAGTCG

..(((.(((.((((((((((((.((((((((((((((((.((((.((((((...)))))).)))))..)))))))))))))))))))))))))).)))..

0024

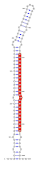


ATTTGTAGTGTTCGGATTGAGTTT Chr 5B:374449053 [-]

CCACTTTTGCAAACTTAGTCTTGGCACTATAAATTTAAAATGTGCATGTGAAAATATGCTCATTATAAATTTGTAGTGTTCGGATTGAGTTTATAATAGTCG

..(((.(((.((((((((((((.((((((((((((((((.((((.((((((....)))))).)))))..)))))))))))))))))))))))))).)))..

0025

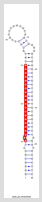


ATTTGTAGTGTTCGGATTGAGTTT Chr 5D:322788342 [-]

CCACTTTTGCAAACTTAGTCCTGGCACTATAAATTTAAAAATGTGCATGTGAAAATATGCTCATTATAAATTTGTAGTGTTCGGATTGAGTTTATAATAGTCG

..(((.(((.((((((((((((.((((((((((((((((..((((.((((((....)))))).)))))..)))))))))))))))))))))))))).)))..

0026



CAGAACCAGAATGAGTAGCTC Chr 2D:537495892 [-]

GAGGCGCCTCCAGAACCAGAATGAGTAGCTCATGCACCTACTCAACATAGTGAATGTGAGCTGCTCATTCTGGTTTTAGATGCGCTTC

(((((((.((.((((((((((((((((((((((.(((............)))...)))))))))))))))))))))).)).)))))))

0027

CATGTGCCTTTCTTCTCCACC Chr 1A:493586793 [+]

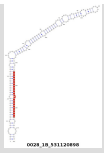
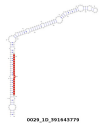
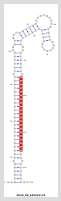
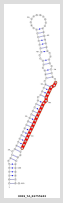
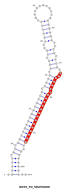
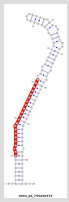
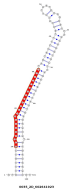
AGGCGCGCGAGGTGGAGAAGCAGGGCACGTGCATCCATTTCCAGCTCGGCATTCCCGGCGTCCGGCCGGCCGGCTGCCGCGGCCTTGCCTGGCTGGGTAGTGCGTCG
CTCGATCCGGCCGTGCGCCGGCGGCCGGCCCTTGCATGCATGTGCCTTTCTTCTCCACCGTGCACACCT

(((...(((.(((((((((((.((((((((((((((.((......((.((((....(((((((.(((((((.(((((((.
(((((......)))))))))))).)).........)))))))..)))))))..)))))))...)).)))))))))))))).)))...)))

0028

CATGTGCCTTTCTTCTCCACC Chr 1B:531120898 [+]

AGGCGCGCGAGGTGGAGAAGCAGGGCACGTGCATTCATTTCCAGGTCGCTTCCCGGCGTCCGGCCGGCACGCACCGCCGCGGCGGCCTTGCTTGGTCAGTGTGTCGC
TCGATCCGGCCGTGCGCCGGCAGCCGGCCCATGCATGCATGTGCCTTTCTTCTCCACCGTGCACACCT

		(((...(((.((((((((((.(((((((((((((.(((.....(((((((..(((((((.(((((((..((...((.((((.. ((((......)))))..)))))..))..))..)))))))).)))))))).))).)))))).))).))))))))))))).))))))))))...)))
0029		CATGTGCCTTTCTTCTCCACC Chr 1D:391643779 [-] AGGCGCGCGAGGTGGAGAAGCAGGGCACGTGCATTCATTTCCAGCTCGGTTCCCGGCGTCCGGCCGGCTGGCTGGCACGCACCGCCCCGGCCTTTCCTGGCTAGTGT GTCGCTCGATCCGGCCGTGCGCCGGCGGCCGGCCCATGCATGCATGTGCCTTTCTTCTCCACCGTGCACACCT (((...(((.((((((((((.(((((((((((((.(((.....((.(((((.(((((((.(((((((..(((.(((((((...(((.. ((.....)).)))..)))))))))))....)))))))..)))))))..)))))))..)))).))))))))))))))).))))))))...)))..)))
0030		CCACGACTGTCATTAAGCATC Chr 5B:440019179 [+] AAGTGAAGATGATGCTTAATGACAGTCGTGGTATCTTAGAAGAGCCAGGCGACTATGTCTATGCTGGCCATTAAGCATCGTCTTCGCCTAGGACACCACGACTGTCA TTAAGCATCGCCTTCACCT ..((((((.((((((((((((((((((((((((((.((((((..((...((((((.......... ((((........))))))))))))))..)))))))..)))))))))))))))))))))))))))))).)))))))..
0031		CCACGACTGTCATTAAGCATC Chr 5D:370732244 [+] AAGTGAAGATGATGCTTAATGACAGTCGTGGTATCCTAGAAGAGCCAGGCGACTGTCTACGCTGGCCATTAAGCATCGTCTTCACCTAGGACACCACGACTGTCATT AAGCATCGCCTTCACCT ..((((((.((((((((((((((((((((((((((.((((((((((((((((. ((........))))))).............)))))...))))))..)))))))))))))))))))))))))))))).)))))))..
0032		CCATGATGAGGTCGTTCAACC Chr 7A:62755243 [-] CGTTTTTGCCGGTTGAACGACCTCACCATGTCGACCGCCTCTTCCTCGACAAGTGCCAGGGAGATCTGCTCGTGCCATGATGAGGTCGTTCAACCAGCAAAACGG ((((((.((.(((((((((((((((((.((((.(((.((..(((((((...........)))))))..))).)))...)))))..)))))))))))))))))))))))))))))))).
0033		CCATGATGAGGTCGTTCAACC Chr 7D:58259009 [-] CGTTTTTGCCGGTTGAACGACCTCACCATGTCGACCGCCTCTCCTCGACAAGTGCCAGGGAGATCTGCTCGTGCCATGATGAGGTCGTTCAACCAGCAAAACAA .(((((.((.(((((((((((((((((.((((.(((.((..((((((.............)))))))..))).)))...)))))..))))))))))))))))))))))))))))))))..
0034		CCTGTTTGTCATTAAGTTTCTT Chr 2A:735236373 [-] GTGTTTGTCACCTGTTTGTCATTAAGTTTCTTAGCGCCTTTGCCATCCTAGTGACTAGTGAAAGGTCAAAAGGTACTTAGAAGCTTAATGAGAACATGTGACAAACG T ..(((((((((.(((((.(((((((((((((.((.(((((((((. ((((((...)))))..))..))))..)))))))..))..))))))))))))))))))))))))))..))))))))..
0035		CCTGTTTGTCATTAAGTTTCTT Chr 2D:602641923 [-] GTGTTTGTCACCTGTTTGTCATTAAGTTTCTTAGTGCCTTTGCCATCTTAGTGAAGGGTCAAAAGGTACTTAGAAGCTTAATGAGAACATGTGACAAACGT ..((((((((((.(((((.(((((((((((((((.(((((((((((((.((.....)).)..)))..)))))))))).))))))))))))))))))))))))))..)))))))))..
0036		CGCGGACAACGAGGATGAAGACGA Chr 1B:459805466 [-] CGTGAAGATATCGTCTTCATCCTCGTTGTCCGCAGTCCATAAAAAAGTATTTTCTGATGAAAACTTTAGGTATGATTCACCCTTACTCTGTTGCTTCATCCATCTTC CATGCAACCTGAGAAGTTCTTTCCGTATATCCTCAAAGGTCCTGGAAAGAACTTCTCAGGTTGCATGGAAGATGGATGAAGCAATAGAGTAAGGGTGAATCATACCT AAAGTTTGCAGCAGAAAATACTTTTTATGGACCGCGGACAACGAGGATGAAGACGATATCTTGACA .((.((((((((((((((((((((((((((((((((((.((((((((((.((((((((((.((. ((((((((((((((((((((((((((((((((((

0038

CGTTGAGATGAGATTACCCCA Chr 5D:526099140 [-]

CTGCAGCGGCCGTTGAGATGAGATTACCCCATACCCTGTAGCGGCAGCTCGACTCCGTGCATGCGTAGGGGGCAATCTCACCTCAACAGCTGCTCCCA

....((((((.((((((.(((((((.((((.(((..((((.(((.((.....)))))))))...))).)))))).)))))))..)))))).))))))....

0039

CTCGGACCAGGCTTCATTCCCC Chr 1B:357916236 [-]

TTGCTTCTGGGTGGAATGTTGTCTGGTTCAAGGTCTCGCTCTGGTTTGAGGATGGAAAACTTTCATTTCTCTCAGCTTCCTGCGATCTCGGACCAGGCTTCATTCCC
CTCAGAGATAG

....(((((((.(((((((..((((((((((.((((((..((...((((.(((((..
(((....)))..)))))))..)))))....)))))))).)))))))))..)))))))..)))))))....

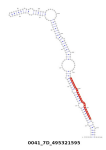
0040

CTCGGACCAGGCTTCATTCCCC Chr 7B:523239739 [-]

TTCGTTTTGAGGGGAATGTTGTCTGGCTCGGGGCTCTCTTTCTTTCTCTCTCTCTCTGTGTTCCTCCCCTCGATCCCTCCTTGATCAGATCGAGAGAAAGATCGAAG
ATCCGATGAGGGGGGAGCAGAGAGTAGACGTGCGTAGATCTCGGACCAGGCTTCATTCCCCCCAAAGACAA

...((((((.(((((((((..(((((((.((((((.......(((.......(((((((.(((((((.(((((((.(((..((.((((((...
((....))..)))))))..))...)))))))))))))))))).))))).)))).......)))))))))..)))))))..))))))..))))..

0041

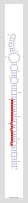


CTCGGACCAGGCTTCATTCCCC Chr 7D:495321595 [-]

TTCGTTTTGAGGGGAATGTTGTCTGGCTCGGGGCTCTCTTTCTTTCTCTCTCTCTATCCTCCTCCCCTCGACCCCTCCTTGATCAGATCGAGAGAAAGGTCGAAGAT
CCGATGAGGGGGGAGAAGAGAGTAGACGTGCGTAGATCTCGGACCAGGCTTCATTCCCCCCAAAGACAA

...((((((.(((((((((..(((((((.(((((((.(((.......((((((((((((.(((((((((...((((((..
(((((((((...)))))))..))..)))))))).........))))))))))).)))))))..))).)))))))..)))))))..))))))..))))..

0042

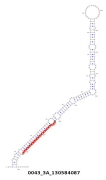


CTTGAACTTCTCCATAGCATC Chr 4D:507462046 [+]

CAATCCTTTGCTTGAACTTCTCCATAGCATCATCCATCCTGCCATTCTACTGAGGTAGCCGCTGCTCTTTTGCATGATGGGTACAATGGCTGGTGCTATGGGTAAAT
TCAACCAGATGGTTG

(((((.((((.(((((.((.((((((((((((.((((..((((..((....((((.
(((....)))))))).....))..)))))..)))).))))))))))))).))..))))))..)))))

0043

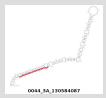


CTTGGACTGAAGGGTGCTCCC Chr 3A:130584087 [+]

ATGGACGGAAGAGAGCGTCCTTCAGTCCACTCACAGGCGGTGCTAGGATCGAATTAGCTGCCGACTCATTCGTCCACTTGCCGAGCAAAGTTATTTGCCGAGGAGCT
TAGCAGGTGAGTGAATGAAGCGGGAGGTAAAGTCTTTGATCTCGCACCGTCGTTGCTTGGACTGAAGGGTGCTCCCTCCTCCTCTC

..(((.((.((.(((((((((((((((((..((..(((((((..(((((((((((.((.(((..(((((((..(((((((.
((((.....................)))).)))))))..)))))))..))).))..)))....)))))))).)))))))..))..)))))))))))))))))....

0044



CTTGGACTGAAGGGTGCTCCCT Chr 3A:130584087 [+]

GATGGACGGAAGAGAGCGTCCTTCAGTCCACTCACAGGCGGTGCTAGGATCGAATTAGCTGCCGACTCATTCGTCCACTTGCCGAGCAAAGTTATTTGCCGAGGAGC
TTAGCAGGTGAGTGAATGAAGCGGGAGGTAAAGTCTTTGATCTCGCACCGTCGTTGCTTGGACTGAAGGGTGCTCCCTCCTCCTCTCT

((.(((.((.((.(((((((((((((((((..((..(((((((..(((((((((((.((.(((..(((((((..(((((((.
((((.....................)))).)))))))..)))))))..))).))..)))....)))))))).)))))))..))..)))))))))))))))))...

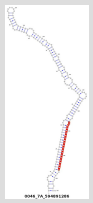
0045

CTTGGATTGAAGGGAGCTCTC Chr 3B:174450524 [-]

TGGCGAAAAAGATGAGCTCCCTTCGACCCAATCCAGGAGGGGGAAGTGGTCGGATGCAGCTGCCGGTTCATGGATACCTGTCCGGTGCATGCAGTTCGGTTGAGCCTG
CGCACTTTGTATAGGTTTTGCATGACCCGGGAGATGAACCTACCAGCATCTTTCCCTCGTGCTTGGATTGAAGGGAGCTCTCTCTCTCTATC

(((.((...(((.(((((((((((((((.(((((..((.(((((((((((((.((...((.((.((((.(((((((.(((((((..(((((..
(((((((....))).)))))))))....)))))))..))..))))))..))))).))..))..))..))))....))))))))).))..))))).))))))))))))))..))..))))..

0046

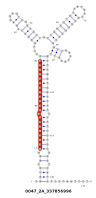


CTTGGATTGAAGGGAGCTCTC Chr 7A:594891286 [-]

GGCGAAAAAGAAGAGCTCCCTTCGATCCAATCCAGGAGGGGAGGTGGTCGGATGCAGCTGCCGGTTCATGGATACCTGTCCGGTGCATGCAGTGCGGTTGAGCCTGC
GCACCTGTATAGGTTTTGCATGACCCGGGAGATGAACCTACCAGCATCTTTCCCTCGTGCTTGGATTGAAGGGAGCTCTCTCTCTCTCTC

((.((...(((((((((((((((((((((((..((.((((((((((((.((...((.((.((((.(((((((.(((((((..(((((..((((.
((......)))))))))))..)))))))..)))..))))).))))).))..))..))..))))....))))))))).))..))))))))))))))))))..)))..)).

0047

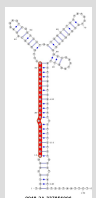


GAGAGGACAGAGCGAGCAGGC Chr 2A:337856996 [+]

TGGTGCGGGAGAGAGGACAGAGCGAGCAGGCACTGGGGCAGAGGAGTCGGTCCTGCTGGCCTGTTGTGCTGCGGGCCGGGACGACTGGGTTGTTTGCTCGTTTTTTT
TTCTCTGCCTCATCTCCTAATTCT

.((((.((.(((((((((.(((((((((((((...(((((.((....)).)))))).((((((((((......)))))))))
(((......))))))))))))))))).))))).))..))..))..))).)))))..........

0048



GAGAGGACAGAGCGAGCAGGCA Chr 2A:337856996 [+]

TGGTGCGGGAGAGAGGACAGAGCGAGCAGGCACTGGGGCAGAGGAGTCGGTCCTGCTGGCCTGTTGTGCTGCGGGCCGGGACGACTGGGTTGTTTGCTCGTTTTTTT
TTCTCTGCCTCATCTCCTAATTCT

.((((.((.(((((((((.(((((((((((((...(((((.((....)).)))))).((((((((((......)))))))))
(((......))))))))))))))))).))))).))..))..))..))).)))))..........

0049

GATAGGGAAACTGACATGCCT Chr 6A:4769710 [-]

TGAAGGGTGAGATAGGGAAACTGACATGCCTTGAAGAGCTGCAGCTATATTCGGTGGACAAGTGCCCAAACTTCTTCACAATGGTGGGGAAGCTGACAGAGTTAAGG
GTGCTTAACATCTTTTTCGGTGAAATGGAGGAGTCTGCAGGCAAGGCTCTGATGGTATCTCTGTGCAACCTTCAC

((((((.((..(((((((.((((.((.((((((.....(((((((.....((((...(((((((((...(((((((((....)))))))))....
((((...)))))..)))))))))..)))))).((((((.....)))))))).)))))))..)))).)))))..)).)))))).

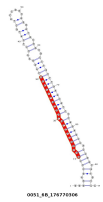
0050

GCGGACAACGAGGATGAAGACGAT Chr 1B:459805465 [-]

ACGTGAAGATATCGTCTTCATCCTCGTTGTCCGCAGTCCATAAAAAAGTATTTTCTGATGAAAACTTTAGGTATGATTCACCCTTACTCTGTTGCTTCATCCATCTT
CCATGCAACCTGAGAAGTTCTTTCCGTATATCCTCAAAGGTCCTGGAAAGAACTTCTCAGGTTGCATGGAAGATGGATGAAGCAATAGAGTAAGGGTGAATCATACC
TAAAGTTTGCAGCAGAAAATACTTTTTATGGACCGCGGACAACGAGGATGAAGACGATATCTTGACAT

..((.((((((((((((((((((((((((((.((((((((((.((((((((((.((.
((.....
(((...)))...)).)).)))))))))))))))))))))))).)))))))))))

0051

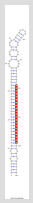


GCGGCTCTGTGGTGTTCAAGC Chr 6B:176770306 [-]

GGGGAGCCCGGCGGCTCTGTGGTGTTCAAGCAGGAACCTCATGCTACCGGCAGGATGCGGCGCTTGCTTGAACATCCCAGAGCCACCGGCGTGCCA

..((.(((((((.(((((((.(((((((((((((((..((.(((.((......)).))).)).)..)))))))))))))).)))))))).)))).)).)).

0052

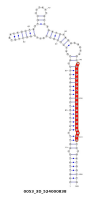


GGTGCTATGGATAAATTCAAC Chr 3A:658832986 [+]

CAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCATCCTGCCATTCAAATGCATTTTTGCACGATGGGGAAGATGGCTGGTGCTATGGATAAATTCAACCCGATGG
TTG

(((((.((.((((((((((((((((((((((((((((((((...((..((...
((((....)))).)).)..)).)...)))))))))))))))))))))))))))))))).)).)))))

0053



GGTGCTATGGATAAATTCAAC Chr 3D:524000838 [+]

CAGTTATCCGGTTTAATTTGTCCATAGCATCCGCGATCCCGCCATTCAGCTGTGATAGCTGCTGCCCTATTGCACGGTGTGGAAGATAGCTGGTGCTATGGATAAAT
TCAACCCGATGGCTG

(((((((((.((((.((((((((((((((((((.((..(((((((...((((((...)))))))..
(((......))).)))))..)))).....)).)))))))))))))))))).)))).)))))))))

0054

GGTGCTATGGATAAATTCAAC Chr 4D:507434720 [+]

CAATCCTCTGGCTGAATTTGTCCATAGCATCAGCACCGGCCGTTCAACTCGGACGGCGGTAGTAGTTGCTGCCTTTCGGCGCTGGTGCTATGGATAAATTCAACCGG
ATGGTTG

(((((.(((((.((((((((((((((((((((((((.((((((((((.....)))))))
((((((....)))))))...)))).)))))))))))))))))))))))).)))))).)))))

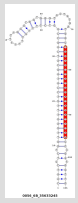
0055

GGTGCTATGGATAAATTCAAC Chr 4D:507477627 [+]

CAATCCTCTGGTTGAATTTTTCCATAGCATCAGCACCGGCCATTCAACTCTGGTGGTAGTAGTTGCTGCCTTTTGGCGCTGGTGCTATGGATAAATTCAACCGGATG
GTTG

(((((.((((((((((((((.((((((((((((((.(((((((.........))))
((((((....)))))))...)))).)))))))))))))))))))))))))))))).)))))

0056

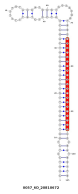


GGTGCTATGGATAAATTCAAC Chr 6B:35633245 [+]

CAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCAGTCCTCTGCTTTTTTTGCACGATGGGTACTACGATGGCTGGTGCTATGGATAAATTCAACCCGATGGTTG

(((((.((.((((((((((((((((((((((((((((((((((((((.......))).)).))))).......)))))))))))))))))))))))))))))))).)).)))))

0057



GGTGCTATGGATAAATTCAAC Chr 6D:20810672 [-]

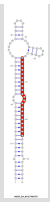
CAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCAATCCTCTGTTTTTTTGTACGATGGGTACTACGATGGCTGGTGCTATGGATAAATTCAACCCGATGGTTG

(((((.((.((((((((((((((((((((((((((((((((..(((........))))..))))).......)))))))))))))))))))))))))))))))).)).)))))

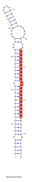
0058

GTAGGATGGCCGGTGCTATGG Chr 5A:691798576 [-]

TTGAAATTGTCCATAGCATCATCCATCCTACCATTCAACGGCGACAGCTGCTGCCCCACGGTGGGTAGGATGGCCGGTGCTATGGGTAGCTTCAA

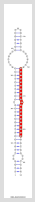


(((((.(((((((((((((((..((((((((((.....((((....)))))..(((....))).)))))))))))..)))))))))))))))).)))))

0059		GTAGGATGGCTGGTGCTATGG	Chr 4D:507476247 [+]

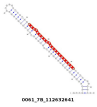
TTGAATTTGTCCATAGCATCATCCATCCTACCATTCAACGGCGGCTGCTGCTGCCCCACCGGTGGGTAGGATGGCTGGTGCTATGGATAGCTTCAA

(((((.(((((((((((((((.((((((((((..((...((((((....)))))).....))..)))))))))).)))))))))))))))).)))))

0060		GTAGGATGGCTGGTGCTATGG	Chr 4B:654094213 [-]

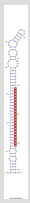
TTGAATTTGTCCATAGCATCATCCATTCTACCATTCAACGGTGGCTGCGCTGCCCGCAGTGGGTAGGATGGCTGGTGCTATGGATAGCCTCAA

((((..(((((((((((((((.((((((((((.......((((((...))))))).......)))))))))).))))))))))))))..))))

0061		GTCTCTGCCAATTCTTCGTGT	Chr 7B:112632641 [+]

ATTGCGAAGAGCGCGAAGGATTTGCAGATACTCCGTAAAATCACTTTATGGAGTCTCTGCCAATTCTTCGTGTTCTTAACGTC

...((((((((((((((((((.(((((.((((((((((.....)))))))))).))))))..))))))))))))))))..)))).

0062		GTGCTATGGATAAATTCAACC	Chr 3A:658832987 [+]

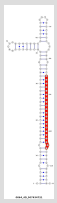
CCAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCATCCTGCCATTCAAATGCATTTTTGCACGATGGGGAAGATGGCTGGTGCTATGGATAAATTCAACCCGATG
GTTGA

.(((((.((.((((((((((((((((((((((((((...((..((...
(((....)))).))..))...)))))))))))))))))))))))).)).)))))).

0063		GTGCTATGGATAAATTCAACC	Chr 3B:691444913 [+]

CCAGTTATCCGGTTTAATTTGTCCATAGCATCCGCGATCCCGCCATTCAGCTGTGATAGCTGCTGCCCTATTGCACGTTGCGGAAGATAGCTGGTGCTATGGATAAA
TTCAACCCGATGGTTGT

.(((((((((.((((.((((((((((((((((((.((.(((((((.....
((((((((((.........)))))))).)))))))..)))).)).)))))))))))))))))).)))).)))))))).

0064		GTGCTATGGATAAATTCAACC	Chr 4D:507434721 [+]

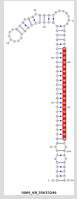
CCAATCCTCTGGCTGAATTTGTCCATAGCATCAGCACCGGCCGTTCAACTCGGACGGCGGTAGTAGTTGCTGCCTTTCGGCGCTGGTGCTATGGATAAATTCAACCG
GATGGTTGC

.(((((.(((((.((((((((((((((((((((((((((.((((((((((.....)))))))
((((((....)))))))...)))).)))))))))))))))))))))))))).)))))).

0065		GTGCTATGGATAAATTCAACC	Chr 4D:507477628 [+]

CCAATCCTCTGGTTGAATTTTTCCATAGCATCAGCACCGGCCATTCAACTCTGGTGGTAGTAGTTGCTGCCTTTTGGCGCTGGTGCTATGGATAAATTCAACCGGAT
GGTTGT

.(((((.((((((((((((((.((((((((((((((.(((((((........))))
((((((....)))))))...)))).)))))))))))))))))))))))))).)))))).

0066		GTGCTATGGATAAATTCAACC	Chr 6B:35633246 [+]

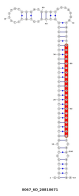
CCAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCAGTCCTCTGCTTTTTTTGCACGATGGGTACTACGATGGCTGGTGCTATGGATAAATTCAACCCGATGGTTG
T

.(((((.((.
((((((((((((((((((((((((((((((((((((((.......))).)).)))).......)))))))))))))))))))))))))))))).)).)))))).

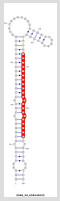
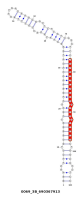
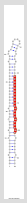
0067

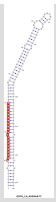
		GTGCTATGGATAAATTCAACC	Chr 6D:20810671 [-]

CCAATCCTCTGGTTGAATTTGTCCATAGCATCAGCCAATCCTCTGTTTTTTTGTACGATGGGTACTACGATGGCTGGTGCTATGGATAAATTCAACCCGATGGTTGA



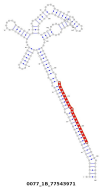
.(((((.((.(((((((((((((((((((((((((((((((((..(((........)))..))))).......)))))))))))))))))))))))))))))).)).)))))).

0068		GTGTTCTCAGGTCGCCCCCGC	Chr 3A:658244035 [-]
		GGAAATTCCTGCGGGGTCGAGCTGGGAACACATGAGATATGAACCGCTTGATTGAATATCAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTCC	
		(((..((((.((((((.(((.(((((((((((((((((...........(((((.....)))))))))))))))))))))).)))).)))))))..)))	
0069		GTGTTCTCAGGTCGCCCCCGC	Chr 3B:690367913 [-]
		GGAAATTCCTGCGGGGTCGAACTGGGAACACATGGGATACGAACCGCTTGATTGACGAGCGACTCTTATCAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTCC	
		(((..((((.((((((.(((.((((((((((((((((((..((.. ((((((.....)))))))..)).)))))...)))))))))))))))).)))).)))))))..)))	
0070		GTGTTCTCAGGTCGCCCCCGC	Chr 3D:523446450 [-]
		GGAAGTTCCTGCAGGGTCGAACTGAGAACACATGAGATATATTGACGAGCGACTAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTCC	
		(((.(((((.((.(((.(((.((((((((((((((((.((..(((.....))).)).)))))))))))))))).)))).)))).)))))).)))	
0071		TAAAAAATCTGTAGGCACTAC	Chr 5B:430318744 [+]
		ATAGTGCCTATAAATTTTTTATAAAGTCAAACTTTATAAACTTTGACCAAGTTTATACAGAAAATTATTTATATATACAATGTTAAATATATAAAATATGAAAATAC ATCTTACGATGAATCTAATAATATATGTTTGGCACTCTAAATGTAAATGCTTTTCTCCGTAAACTTGGTCAAAGTTTGTAAGGTTTGACTTTTTAAAAAATCTGTAG GCACTACATTATGAAAC	
		.((((((((((((.(((((((((.((((((((((((((((((((((((((((((((((..((((((.((((((((((.((...(((((((((((...(((....((. ((((....)))).))...)))).)))))))))))))...)).)))))))))).))))))..)))))))))))))))))))))))))))))))).)))))))).))))))))))).....	
0072		TAAAAAATCTGTAGGCACTAC	Chr 7A:516590345 [-]
		GTTCCATAATGTAGTGCCTATTGATTTTTTGAGAAGTCAAACTTTACAATCTTTGACCAAATTTGTAGAGAAAAATATTTATATTTACAATACCAAAAATATATAAT GTGAAACTACATCTTATGATGAATCTAGCGATATATGTTTCCCACTCTAGATGGAAATATTTTTCTCCACAAACTTGGTCAAAGTTTGTGAGGTTTGACTTTTTAAA AAATCTGTAGGCACTACATTATGGAAC	
		((((((((((((((((((((((.((((((((((((((((((((((((((((((.((((((((((.(((((.((((((((((((((.(((((((.........(((((((((. (((..... ((((....)))).....))).)))))))))).....)))))))).)))))))))))))).)))))).)))))))))))))))))))))))))))))))



```
((((((.((.(((((((.((((((((((((((((((.(((((((((((((.(((((((...
(((.....)))..)))))))).)))))))))))))))))))))))))))))))).))).))))))
```

0077



TAATAAACAAGTCTTTAGATG Chr 1B:77543971 [+]

```
ATCTGCTTCTCATTTGAAGACTAGTTTATTACATATAAGCATATTCTTGAATGTTTGAGAGAGAGATCACAAGCAGATCTCGTGCAGAAGAGAGCATCATTGATACA
ACTATAATGATTCTTCCACTTCAAGAGCCATTGGGCCGCTGAGAAATATGTTTAGATATAATAAACAAGTCTTTAGATGAGAAGGCAGA
```

```
.((((((((((((((((((((((.(((((((((.((.((((((((((((((....)))))))..((((((.......)))))))))..((((((....((((((.
(((....)))))))))))))))))))))))(((...(((....)))...)))))..)))))))))).)).)))))))).)))))))))))))))))).)))))
```

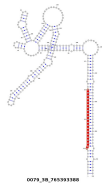
0078

TAATAAACATGTCTTCAGATG Chr 1A:49996930 [+]

```
ATCTGCTTCTCATCTGAAGACTAGTTTATTACATATAAGCATATTCTTGAAAGTTTGAGAGAGAGAGAGAGATCACAAGCAGATCCCGTGCAGAAGAGAACATCA
CTGATACAACTATAATAATTCTTCCACTTCAAGAGCCATTGGGTCGCTGAGAAATATCTTTAGATATAATAAACATGTCTTCAGATGAGAAGGCAGA
```

```
.((((((((((((((((((((..(((((((.............(((((((.((..(((((.............((((.......)))))..((((((..
((.....)).)))..)))...........))))))..)))))))))))(((...)))((((...(((((....)))))..)))).)))))))..)))))))))))))))))).)))))
```

0079

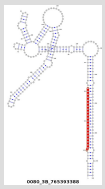


TAATCTTCTGGAAATATGCTT Chr 3B:765393388 [+]

```
CTTAAATTTGTAATCTTCTGGAAATATGCTTAAGTATAATAAATGATGCGAAGGTAGATCATAGAACATGAACACATGGTGTAAAAGAGGCACTCATGCACCAAAAC
TATAAAAAGCAGCAAAGAAAATGAGTAAAAGAGGCACTTGACTTGTAGATTATACAAGAGTGATGATCTGGAAGTCAACCCTCTTTTATCAGTAGTTTTGATCTACC
TTCTCATCTATAGGATTAGTTATTATGTCTAAGCATGTTTTTAGAAGATTAGAAGATTAAG
```

```
(((((.(((.((((((((((((((((((((((((.(((((((((..((((.((((((((((.......(((....)))))))))...(((...))).))))))
(((((((.....................(((.((((((((...(((((((.
((((((((((....)).)))))))).)))))))).)))))))))))))))))))))))))))))))))))))))))))).)))).)))))))))).)))))))))))))))))))))))).)))
```

0080



TAATCTTCTGGAAATATGCTTA Chr 3B:765393388 [+]

```
CTTAAATTTGTAATCTTCTGGAAATATGCTTAAGTATAATAAATGATGCGAAGGTAGATCATAGAACATGAACACATGGTGTAAAAGAGGCACTCATGCACCAAAAC
TATAAAAAGCAGCAAAGAAAATGAGTAAAAGAGGCACTTGACTTGTAGATTATACAAGAGTGATGATCTGGAAGTCAACCCTCTTTTATCAGTAGTTTTGATCTACC
TTCTCATCTATAGGATTAGTTATTATGTCTAAGCATGTTTTTAGAAGATTAGAAGATTAAG
```

```
(((((.(((.((((((((((((((((((((((((.(((((((((..((((.((((((((((.......(((....)))))))))...(((...))).))))))
(((((((.....................(((.((((((((...(((((((.
((((((((((....)).)))))))).)))))))).)))))))))))))))))))))))))))))))))))))))))))).)))).)))))))))).)))))))))))))))))))))))).)))
```

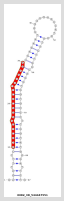
0081

TACGGCCTGATGACATCCACA Chr 3D:143964741 [+]

```
AGCTTCACACCGTGGATGTCATCGTGGCCGTACATGAATCGTCCCCCACATCCAGCTCTGTTCAATTTCAGACAAAGGTGGACGTTTCATTGTACGGCCTGATGACA
TCCACAGCGTGGCTCT
```

```
(((..(((...(((((((((((((((.((((((((((((((.(((((.((.........
((((........))))....)).)))))).)))))))))))))))))))))))))...)))))))..
```

0082

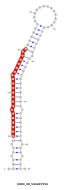


TACTGTGGGCACTTATTTGAC Chr 3D:544447551 [-]

```
GAGATTTCCATACTGTGGGCACTTATTTGACAAGAATGAGGAGAAACATGAATTCTTGGCAAATAGTGCCCTTAGTATGAAATTCTG
```

```
.(((.((.(((((((.(((((((.(((((.(((((((.............)))))))..)))))))))))))).)))))))..))).
```

0083

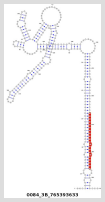


TACTGTGGGCACTTATTTGACA Chr 3D:544447550 [-]

```
GAGATTTCCATACTGTGGGCACTTATTTGACAAGAATGAGGAGAAACATGAATTCTTGGCAAATAGTGCCCTTAGTATGAAATTCTG
```

```
.(((.((.(((((((.(((((((.(((((.(((((((.............)))))))..)))))))))))))).)))))))..))).
```

0084

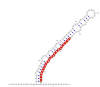


TAGAAGATTAGAAGATTAAGC Chr 3B:765393633 [+]

```
AGTTGGTCTCTCTTAAATTTGTAATCTTCTGGAAATATGCTTAAGTATAATAAATGATGCGAAGGTAGATCATAGAACATGAACACATGGTGTAAAAGAGGCACTCA
TGCACCAAAACTATAAAAAGCAGCAAAGAAAATGAGTAAAAGAGGCACTTGACTTGTAGATTATACAAGAGTGATGATCTGGAAGTCAACCCTCTTTTATCAGTAGT
TTTGATCTACCTTCTCATCTATAGGATTAGTTATTATGTCTAAGCATGTTTTTAGAAGATTAGAAGATTAAGCAAGCTCATTC
```

```
...(((.((..(((((.(((.(((((((((((((((((((((((.(((((((((..((((.((((((((((.......(((....)))))))))...
(((...))).))))))(((((((((.....................(((.((((((((...(((((((.
((((((((((....)).)))))))).)))))))).)))))))))))))))))))))))))))))))))))))))))))).)))).)))))))))).)))))))))))))))))))))))).)))
```

0085



TAGAAGATTAGAAGATTAAGC Chr 3D:575552486 [+]

```
TTTACTACCAGTTTTGATCTACCTTATCATCTAGAAGATTAGTTATTATATCTAAGTATGTTTTTAGAAGATTAGAAGATTAAGCAAAGTCACTT
```

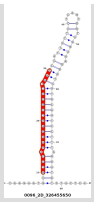
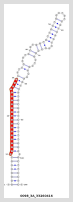
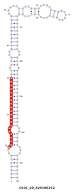
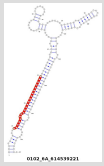
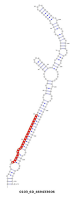
```
..........((((.((((((..((.(((.(((((((((..((.(((.......)))..))..)))))))).)))).)).)))))))))..........
```

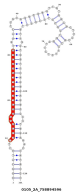
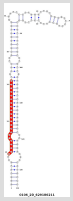
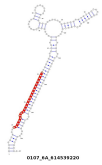
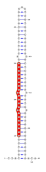
0086

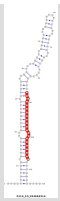
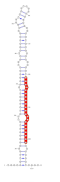
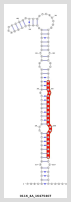
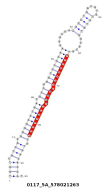
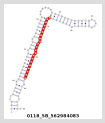
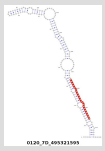
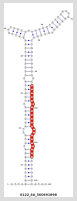
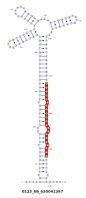
TAGGGTGTAGAATAAGCTATT Chr 4B:120579573 [+]

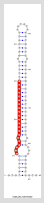
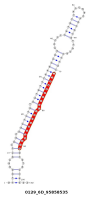
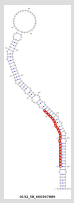
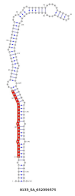
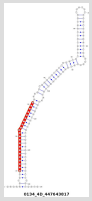
```
GTACCGTTACTTAGGGTGTAGAATAAGCTATTCTACACTCACGCTTAGAATATCATTACCCTCATGTTATTCTAAGCGTGAGTATAAAATAGGGTAACACTAAACTAA
GCGCGGGTGTAGAATAACTTATTCTACACCTCCAAGTAATGCTAT
```

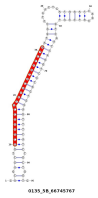
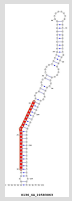
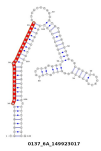
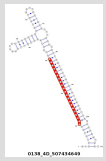
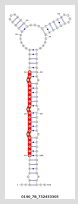
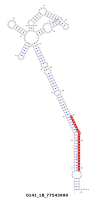
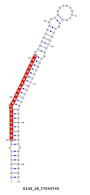
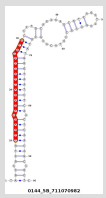
		(((.(((((((.(((((((((((((((.(((((((((((((.(((((((..... (((((((((((((.....)))))))..........))))))).....)))))))).))))))))))))))...)))))))..)))	
0087		TAGGGTGTAGAATAAGCTATT	Chr 4D:142764943 [+]
		GTGAGTGTAGAATAGCTTATTCTACATCCCAGTAGCGGTACATGCAAAATATAGTAATTTCCGTAAAAATATACTAAATTACAAACTTGATAAGTAAAAGAATATTT TCTGGATTACTACATTTTATACACCGTTTTACTAGATTTTGTATATAGTACTACTAGGGTGTAGAATAAGCTATTCTATAGTCAC	
		((((.((((((((((((((((((((((((.(((((...((.(((((((((.((((((..((((..((((((((.((... ((((...........)))).)).)))))))).))))..................))))))).)))))))).)))...)))))).))))))))))))))))))))))))).))))	
0088		TAGGGTGTAGAATAAGCTATT	Chr 7B:123286001 [+]
		GTGAGTGTAGAATAGCTTATTTCACACTGCCTAATAACGCTATATACAAAATCTAGTAAAATGGTATCTAAATATTAGTAATTCTGAAAATAATTTTTTTACTACCC AAGTTTATAAAATTTGCTACATTTTTATAAAAAACACTATATTTTATATATATCGTGTTACTATATTTTGTATGTGCCGTTACTAGGGTGTAGAATAAGCTATTCTA CCCTCAC	
		(((((.((((((((((((((((.(((((..((((.(((((.(((((((((((((.((((((.(((((((.(((((...((((..... (((((.....))))))......... ((((((((((.........)))))))))....))))...))))...)))))))).)))))).)))))))))))))).)))))).)))))))).)))))))))))))))))))))))))

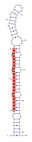
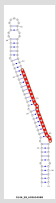
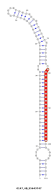
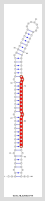
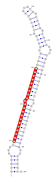
		.....((((((.((((((((((((.(((((.(((((((.....)).)))))..)))))))))))))))))))).))))))..... TCCAAAGGGATCGCATTGATC	Chr 2D:650221338 [+]
0097		AGTGGAGGATTCCAAAGGGATCGCATTGATCGATCTCTCCGTCATCGGCGTGCAAGGTCGATGGATCAGTGCAATCCCTCTGGAATTCTCCGCT ((((((((((((((.(((((((.((((((((((.....((((((((.....))))))))))))))))))).)))))))..))))))))))))))	
0098		TCCAAAGGGATCGCATTGATC TGGGGAAGCATCCAAAGGGATCGCATTGATCCTTCCATCCATGGTGTTGATGGCTGTCATCAATCAACCGGCTCGACAGATCATGCGATCCTTTTGGAAGCTTCCTCT	Chr 3A:33260416 [-]
0099		.(((((((((.(((((((((((((((((.(((((..((...((.. ((((((((((.....))))))))..)))))...))..)))))))))))))))))))).)))))))). TCCACAGGCTTTCTTGAACTG	Chr 2B:772062273 [+]
0100		GGGCATGCTCTCCACAGGCTTTCTTGAACTGTGAACTCGCGGGCGGATTGTGGCGCTCCCGCCTTCTCTTCGGCCGCCTCCGCCTGTTGATGGTTCAAGAAAGCCCA TGGAAACCATGCCC	
		(((((((...((((..(((((((((((((((((.(((....(((((((....((((..... (((........))))))))))))))))).))))))))))))))..))))...)))))))	
0101		TCCACAGGCTTTCTTGAACTG GGGCATGCTCTCCACAGGCTTTCTTGAACTGTGAACTCGCGGGGATGGGGATTCCCTGGATTGTGGCGCTCCTGCCTTCTCTTCGGCCGTCTCCGCCTGTTGATGGT TCAAGAAAGCCCATGGAAACCATGCCC	Chr 2D:629186212 [-]
		(((((((...((((..(((((((((((((((((.(((..((((((((((.....((..(((.... (((......)))...)))..))))))))))))))..))).))))))))))))))..))))...)))))))	
0102		TCCACAGGCTTTCTTGAACTG GGCCATGCTCTCCACAGGCTTTCTTGAACTGTCAACACGCGCGCGCCGGCCAGCCATGGCTTTCCTTTGCCGCCCAACTCGATCGCATATGCTCTTGCATGTCTCGA TCTCTGCCTGCCTTGTTGATGGTTCAAGAAAGTCCTTGGAAACCATGCCC	Chr 6A:614539221 [-]
		(((.(((...((((..(((((((((((((((((((((.(((.((...(((.......(((........)))))).....(((((. ((((((....))))))..))))))...)).)))..)))))))))))))))))))..))))...))))))).	
0103		TCCACAGGCTTTCTTGAACTG GGCCATGCTCTCCACAGGCTTTCTTGAACTGTCAACTCGCGCGCTCCGGCCATCCATGGCCTTCCTTTGCCGCCGAATTCGCATATGCTCTTGCATGTCTCGATCTC GTGCGATCCGATGGATCTCTGCCTGCCTTGTTGATGGTTCAAGAAAGTCCTTGGAAAACATGCCC	Chr 6D:469433606 [-]
		(((.(((...((((..(((((((((((((((((((..(((.((...(((((....)))))).(((.(((.(((((((..(((. ((((((....))))))..))))..))))...)))..)))..)))))..)))).)))))))))))))))))..))))...))))))).	
0104		TCCACAGGCTTTCTTGAACTGT GGGCATGCTCTCCACAGGCTTTCTTGAACTGTGAACTCGCGGGGATTGTGGCGCTGCTGCCTTCTCTTCGGCCGCCTCCGCCTGTTGATGGTTCAAGAAAGCCCATG GAAACCATGCCC	Chr 2B:772062273 [+]
		(((((((...((((..(((((((((((((((((.(((..((((((... (((((((....)).........)))))))))))..))))..))))))))))))))))..))))...)))))))	
0105		TCCACAGGCTTTCTTGAACTGT GGGCATGCTCTCCACAGGCTTTCTTGAACTGTGAACTCGCGGGCGGATTGTGGCGCTCCCGCCTTCTCTTCGGCCGCCTCCGCCTGTTGATGGTTCAAGAAAGCCCA TGGAAACCATGCCC	Chr 2A:758894596 [-]

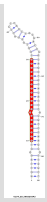
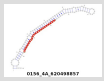
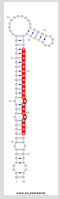
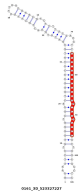
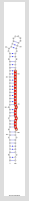
			
		(((((((...((((..(((((((((((((((((.(((....(((((((....((((..... (((........)))))))))))))))))).))))))))))))))))..))))...)))))))	
0106		TCCACAGGCTTTCTTGAACTGT	Chr 2D:629186211 [-]
		GGGCATGCTCTCCACAGGCTTTCTTGAACTGTGAACTCGCGGGGATGGGGATTCCCTGGATTGTGGCGCTCCTGCCTTCTCTTCGGCCGTCTCCGCCTGTTGATGGT TCAAGAAAGCCCATGGAAACCATGCCC	
		(((((((...((((..(((((((((((((((((.(((..((((((((((.....((..(((.... (((......)))...)))..)))))))))))))..))))).))))))))))))))))..))))...)))))))	
0107		TCCACAGGCTTTCTTGAACTGT	Chr 6A:614539220 [-]
		GGCCATGCTCTCCACAGGCTTTCTTGAACTGTCAACACGCGCGCGCCGGCCAGCCATGGCTTTCCTTTGCCGCCCAACTCGATCGCATATGCTCTTGCATGTCTCGA TCTCTGCCTGCCTTGTTGATGGTTCAAGAAAGTCCTTGGAAACCATGCCC	
		(((.(((...((((..(((((((((((((((((((((((.(((.((...(((........(((........))))))).....(((((. ((((((....)))))))..))))))...)).)))..)))))))))))))))))))))))..))))...))))))).	
0108		TCCACAGGCTTTCTTGAACTGT	Chr 6D:469433605 [-]
		GGCCATGCTCTCCACAGGCTTTCTTGAACTGTCAACTCGCGCGCTCCGGCCATCCATGGCCTTCCTTTGCCGCCGAATTCGCATATGCTCTTGCATGTCTCGATCTC GTGCGATCCGATGGATCTCTGCCTGCCTTGTTGATGGTTCAAGAAAGTCCTTGGAAAACATGCCC	
		(((.(((...((((..(((((((((((((((((((((..(((.((...(((((....)))))).(((.(((.(((((((..(((. ((((((....))))))..))))..)))).)))...)))).)))).....)).)))...)))))))))))))))))))..))))...))))))).	
0109		TCCGTCCCATACTATAAGAGC	Chr 7A:467421678 [+]
		TTGTACTCCCTCCGTCCCATACTATAAGAGCATTTTTGACACTACACTAGTGTCAAAAACGCTCTTATATTATGAGATGGAGGGAGTAATA	
		...(((((((((((((.(((((.((((((((((.(((((((((((((...)))))))))))))).)))))))))).)))))..)))))))))))))...	
0110		TCCGTCCCATACTATAAGAGC	Chr 1B:9707435 [-]
		AGATACTCCCTCCGTCCCATACTATAAGAGCGTTAATCGTAGTGTCAAAAATGCTCATATATTATGGAATGGAGGGAGAAGTA	
		.....((((((((((.((((((.((((.(((((((..............)))))))).)))))..))))))..))))))))))..... .

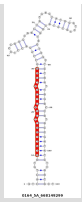
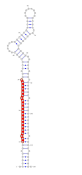
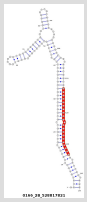
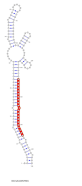
		....(((((((.(((((((..(((((((((.((((((..((...((((.(((((.. (((....)))..))))))..))))....)))))))..)))))))))..)))))))..))))))))....
0115		TCGGACCAGGCTTCATTCCCC Chr 4D:450196377 [+] TGGCAGTTGAGGGGAATGTTGTCTGGTTGGAGACCTGGCTCGCACGACGGATCGATCGGCGGCGGGCGTGCGCGCGCGGTTTCGGACCAGGCTTCATTCCCCCCAACTCAAC ....(((((.(((((((((..(((((((((.(((((((..((.((((((.((.. (((.....)))))..))))))).))..)))))))).)))))))))..)))))))).)))))....
0116		TCGGACCAGGCTTCATTCCCC Chr 4A:16475407 [-] TGGCAGTTGAGGGGAATGTTGTCTGGTTGGAGACCTGGCTCGCACGACGGCCGGACCCCGGATCGATCGACGGCGTGCGTGCGCGGTTTCGGACCAGGCTTCATTCCCCCCAACTCAAC ....(((((.(((((((((..(((((((((.((((((..((.((((((. (((((((...)))))..)))).........)))))))..))..)))))))).)))))))))..)))))))).)))))....
0117		TCGGACCAGGCTTCATTCCCC Chr 5A:578021263 [-] TGGTTGTCGAGGGGAATGACGCCGGGTCCGAAAGAGAGACGCATGGCGTGCGTGTGGCGCGTTTCGGACCAGGCTTCATTCCCCATGACACCAT ((((.((((.(((((((((.(((.(((((((((.....((((((.....)))))).....)))))))))).)))).))))))))).. 0117_5A_578021263
0118		TCGGACCAGGCTTCATTCCCC Chr 5B:562984083 [-] TGGTTGTCGAGGGGAATGACGCCGGGTCCGAAAGAGAGACGCGCGCATGGTGTGCGTGTGGCGCGTTTCGGACCAGGCTTCATTCCCCATGACTCCAT (((..((((.(((((((((.(((.(((((((.....((((((((((((.....)))))).)))))))))))))).)))).)))))))))..))).. 0118_5B_562984083
0119		TCGGACCAGGCTTCATTCCCC Chr 7B:523239739 [-] TTCGTTTTGAGGGGAATGTTGTCTGGCTCGGGGCTCTCTTTCTTTCTCTCTCTCTCTGTGTTCCTCCCCTCGATCCCTCCTTGATCAGATCGAGAGAAAGATCGAAGATCCGATGAGGGGGGAGCAGAGAGTAGACGTGCGTAGATCTCGGACCAGGCTTCATTCCCCCCAAAGACAA ...((((((.(((((((..((((((.((((((.......(((.......(((((((.(((((((.(((((((.(((..((.(((((((... ((....))..))))))).))...)))))))))))))))).)))).)))).......)))))))).))))))..)))))))).)).))))..
0120		TCGGACCAGGCTTCATTCCCC Chr 7D:495321595 [-] TTCGTTTTGAGGGGAATGTTGTCTGGCTCGGGGCTCTCTTTCTTTCTCTCTCTCTATCCTCCTCCCCTCGACCCCTCCTTGATCAGATCGAGAGAAAGGTCGAAGATCCGATGAGGGGGGAGAAGAGAGTAGACGTGCGTAGATCTCGGACCAGGCTTCATTCCCCCCAAAGACAA ...((((((.(((((((..((((((.((((((.(((........((((((((((.(((((((((...(((((((.. (((((((((...)))))).))..)))))))........)))))))))).)))))).)))).......)))))))).))))))..)))))))).)).))))..
0121		TCGGACCAGGCTTCATTCCCCT Chr 1A:326329278 [+] TTTGCTTCTGGGTGGAATGTTGTCTGGTTCAAGGTCTCGCTCTGGTTTGAGGATGGAAAACTTTCATTTCTCTCAGCTTCCTGCGATCTCGGACCAGGCTTCATTCCCCTCAGAGATAGC ...((((((((((.(((((((..(((((((((.((((((..((...((((.(((((.. (((....)))..))))))..))))....)))))))..)))))))))..)))))))..)))))))...)))
0122		TCGGACCAGGCTTCATTCCCT Chr 6A:560691898 [+] ATTAGGTTAAGGGGTATGTTGTCTGGCTCGGTGCCCCCGGTCGAGATTGATAGTCGTCGACTCTCCGACGGGAGACGTCGATTGATCGGGGACGTCGGACCAGGCTTCATTCCCTTTAACCGTGC ....((((((((((.(((..(((((((.(((((((.(((((((((((((...)))). ((((((((((....)))))).)))))))))))))))).)))))))..))).)))))))))....
0123		TCGGACCAGGCTTCATTCCCT Chr 6B:630041297 [+] ATTAGGATAAGGGGTATGTTGTCTGGCTCGGTGCCCCCGGTCGAGGTCGATAGTCGTCGACTCTCCCTTCGCCGGGAGAAAGACGTACGTGTGCGTCGATTGATCGGGGGCGTCGGACCAGGCTTCATTCCCTTTAACCGTGC ....((.(((((((.(((..(((((((.((((((((((((((((((((((((....)))))))))))))))(((((.....)))))). ((((((....))))))).)))))))))))))))))).)))))))..))).))))))))..)))).)).)))).)).)).)

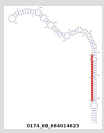
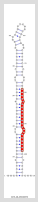
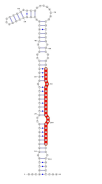
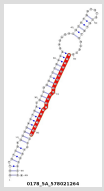
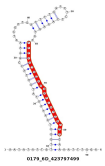
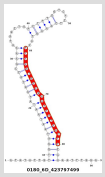
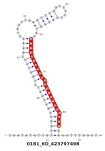
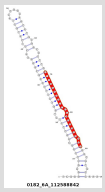
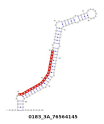
0125		TCTGAAGACTAGTTTATTACA	Chr 1A:49996769 [+]
		TATGTTTAGATATAATAAACAAGTCTTTAGATGGGAAGGCAGATCATAATAGCTGAGCGAAATGTAAGTGATACCGTTGCTCAATTGTTATGATCTGCTTCTCATCT GAAGACTAGTTTATTACATATAAGCATA	
		(((((((((..((.(((((((((.(((((((((((((((((((.(((((((((((((.(((((((... (((.....)))..)))))))).)))))))))))))))))))))))))))))))))))))).))))))))))	
0126		TCTGAAGACTAGTTTATTACA	Chr 1D:51038190 [+]
		CTGCTTCTCATCTGAAGACTAGTTTATTACATATAAGCTATTCTTGGAAGTTTGAGAGAGAGATCACAAGCAGATCCCGTGCAGAAGAGAACATCATTGATACAACT ATAGTAATTCTTCCACTTCAAGAGCCATTGGGCCGCTGAGAAATATGTTTAGATATGATAAACAAGTCTTCAGATGAGAAGGCA	
		...((((((((((((((((((.(((((((((.((.(((((((((((((((..(((((.((.......)).)))))).......(((..((((((.((((.. (((...)))..)).))..)))))))))).......(((....))))..)))))).)))))).)).)))))))).)))))))))))))))))...)	
0127		TCTGCAAACTAATATAAGACT	Chr 2A:628583651 [+]
		TACTCCATCTTCTGCAAACTAATATAAGACTTTTTAGATCACTATATGCTTTAGTGATCTAAAAAGTTTTATATTAGTTTACAGAGGGAGTATTA	
		((((((...(((((.(((((((((((((((((((((((((((((((.......)))))))))))))))))))))))))))))))))))).)))))))))...	
0128		TCTGCAAACTAATATAAGACT	Chr 2B:715343102 [-]
		ATGTACTCCCTCTGCAAACTAATATAAGACTTTTTAGATCACTTATATTAGTGATCCAAAAAGTCTTATATTAGTTTACACAAGGAGTAGTG	
		...((((((...((.((((((((((((((((((((.(((((((.....)))))))).)))))))))))))))))))).))...)))))))...	
0129		TGAACATCCCAGAGCCACCGG	Chr 6D:95858535 [+]
		GTTGGGGAGCCCGGCGGCTCTGTGGTGTTCAAGCAGGAACCTCATGCTACCGGCAGCATGTTGCGCTTGCTTGAACATCCCAGAGCCACCGGCGTGCCAAAA	
		.((((...(((.((.(((((((.(((((((((((((((.....((((((.....)))))).....)))))))))))))).)))))))).)))))...))))..	
0130		TGAACCTTAGGGAAAAGCCGCAT	Chr 5D:525785116 [+]
		CGTAGCTTCTTGAACCTTAGGGAAAAGCCGCATAACACCATTAACTTGAAAACCACTTATCTTATCCCAGAACTGCTCCTACCACAATATAAGGGGTAAAAAATGGT TTTTAGTTTGTGAGGTCATGCGGATTTTTCACCAAGATTCAAGTGGTTATC	
		.((((((.((((((.(((.(((((((((((((.((..(((.((((.(((((((((.((... (((((((.........................)))))))).)).)))))))))))))).)))..)).)))))))).))))).))))))).	
0131		TGAACCTTAGGGAACAGCCGCAT	Chr 5B:660367434 [+]
		CGTAGCTTCTTGAACCTTAGGGAACAGCCGCATAACACCATTAACTTGAAAACCACTTACCTTATCCCAAAACTGCTCCTACCACAATATAAGGGGTAAGGAATGGT TTTTGGTTTGTGAGGTCATGCGGATTTTTCACCAAGGTTCAAGTGGTTATC	
		.((((((.((((((((((.(((((.(((((((((.((..(((.((((.(((((((((.((. (((((((((.........................)))))))))))))))))))))))))))))))).)))..)).)))))))).)))).))))))))))))))).	
0132		TGAACCTTAGGGAACAGCCGCAT	Chr 5B:660367889 [+]
		CGTAGCTTCTTGAACCTTAGGGAACAGCCGCATAACACCATTAACTTGAAAACCACTTACCTTATCCCAAAACTGCTCCTACCACAATATAAGGTTTAAGGAATGGT TTTTAGTTTGTGAGGTCATGCGGATTTTTCACCAAAGGTTCAAGTGGTTATC	
		.((((((.((((((((((.(((((.(((((((((.((..(((.((((.(((((((((.((.(((((.. ((.........................))..)))))))))))))))))))))))))))))))).)))..)).)))))))).)))).)).)))))))))))))).	
0133		TGAACCTTGGGGAAAAGCCGCAT	Chr 5A:652359575 [+]
		CATAGCTTCTTGAACCTTGGGGAAAAGCCGCATAACACCATTAACTTGAAAACCATGTATTTCCTATCCCAAAACTGCTCCTAGCACAAGGGGTAAAAGATGGTTTT TAGTTGTGTGGTCATGCGGATTTTTCACCAAGATTCAAGTGGTTATC	
		.((((((.((((((.(((((.((((((((((((...((((((((((.(((((((((...(((..((((((..... ((((...))))...)))))))))).)))))))))))))))).))))))).)))))))).)))))))))))))).	
0134		TGAAGCTGCCAGCATGATCTGA	Chr 4D:447643817 [+]
		CACAAGTAGATGAAGCTGCCAGCATGATCTGAAAGCTAGCTATGTTGCATGTTGATCTTGATGGTCATCTCCATCAAGATTGAAGACCATGGGCATGGCTGCAATCA GGTCATGTGGCAGCTTCATTTTCTCATG	
		....((.((((((((((((((.((((((((((..((.(((((((((. ((((((((((((((((((((((((((((((

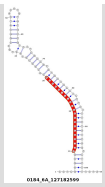
0135		TGAAGCTGCCAGCATGATCTGA	Chr 5B:66745767 [+]
		CACGAGCTGGTGAAGCTGCCAGCATGATCTGATGACCTAACTCATGGATCAGAGTCCATGTCAATCAGGTCATGCTGGAGTTTCATCTGCTGGTC	
		.((.(((.((((((((((.(((((((((((((((((((((.......((((((....)))))))))).)))))))))))))))))))))))).))).)).	
0136		TGAAGCTGCCAGCATGATCTGA	Chr 4A:19583063 [-]
		CACAAGTAGATGAAGCTGCCAGCATGATCTGAAAGCTATGCTGCATCTTGATCTTGATGGTCATCTGCATCAAGGTTGAAGAACATGCCCGGCAATCAGGTCATGTG GCAGCTTCATTTTCTCATG	
		....((.(((((((((((((((.((((((((((..(((..((.... ((((((((((((((.........)))))))))).))))....))..))))..))))))))))))))))))))).))....	
0137		TGAAGCTGCCAGCATGATCTGC	Chr 6A:149923017 [+]
		TGTGAGAGACTGAAGCTGCCAGCATGATCTGCTCGTCCAGCATGCGAGCCGAAAACGCGTTCCATCGCGCGGGGGTTGTATATTTGGAATTTTGGCTCCATCAGATCA TGCTGCAGCTTCATTTTCTCACG	
		.(((((((((.((((((((((.((((((((((((.........(((.(((((((((..(((((.....)))))..(((. ((....)).)))))))))))))))))))))))))))))))))))))).)))))))))).	
0138		TGAATTTGTCCATAGCATCAG	Chr 4D:507434649 [+]
		ATCCTCTGGCTGAATTTGTCCATAGCATCAGCACCGGCCGTTCAACTCGGACGGCGGTAGTAGTTGCTGCCTTTCGGCGCTGGTGCTATGGATAAATTCAACCGGAT GGT	
		..(((((((.((((((((((((((((((((((((((.((((((((((.....))))))) ((((((....))))))...))).)))))))))))))))))))))))))).)))))).)).	
0139		TGAATTTGTCCATAGCATCAG	Chr 4D:507476338 [+]
		ATCCTCTGGTTGAATTTGTCCATAGCATCAGCGTCGGCCGTTCAACTGCGGGCGGTGGTAGTTGCTGCCTTTTGGCGCTGGTGCTATGGATAAATCCAACCGGATGG T	
		..((((((((((.((((((((((((((((((((((((((((((((((.....))))))) ((((((...)))))...))))))))))))))))))))))))))).)))))).)).	
0140		TGAGAAAGGACTGCATCATCT	Chr 7B:732433305 [-]
		TTGGCCACCTTGAGAAAGGACTGCATCATCTGGAAGGATAGCGCACAGCGTAACCTGGTGTTCTCACCAGCAAGTTCCAGAGGATGCACTCCTTCCTCAAGGTGACC TA	
		..((.(((((((((.((((((.((((((.(((((((.....((((...)))))... ((((((....))))))....)))))))).)))))).)))))).)))))))))..)).	
0141		TGAGAAGGCAGATCATAATAGC	Chr 1B:77543990 [+]
		TTGCTCAATTGTTATGATCTGCTTCTCATTTGAAGACTAGTTTATTACATATAAGCATATTCTTGAATGTTTGAGAGAGAGATCACAAGCAGATCTCGTGCAGAAGA GAGCATCATTGATACAACTATAATGATTCTTCCACTTCAAGAGCCATTGGGCCGCTGAGAAATATGTTTAGATATAATAAACAAGTCTTTAGATGAGAAGGCAGATC ATAATAGCTAAGCAATAA	
		(((((...(((((((((((((((((((((((((((((((.(((((((((.((.((((((((((((((....)))))))..(((((((.......)))))))))((.. (((((....((((((.(((....)))))))))))))))))))))))))))))))... (((....)))...)))))..)))))))))).)).)))))))).)))))))))))))))))).)))))))))))))...)))))...	
0142		TGAGAAGGCAGATCATAATAGC	Chr 1D:51038361 [+]
		GTTGCTCAATTGTTATGATCTGCTTCTCATCTGAAGACTAGTTTATTACATATAAGCTATTCTTGGAAGTTTGAGAGAGAGATCACAAGCAGATCCCGTGCAGAAGA GAACATCATTGATACAACTATAGTAATTCTTCCACTTCAAGAGCCATTGGGCCGCTGAGAAATATGTTTAGATATGATAAACAAGTCTTCAGATGAGAAGGCAGATC ATAATAGCTGAGCAGTAA	
		.(((((((.(((((((((((((((((((((((((((((((.(((((((((.((.((((((((((((((..(((((.((......)).)))))).......(((.. ((((((.((((..(((...)))..)).)..))))))))))..... (((....))).)))))).))))).))))).)).)))))))).)))))))))))))))))).))))))))))))))...)))))))...	
0143		TGAGAAGGTAGATCATAATAGC	Chr 1B:77543745 [+]
		AGTCTTTTGATGAGAAGGTAGATCATAATAGCTGAGCGAAATGAATGTGATACAGTTGCTCAATTGTTATGATCTGCTTCTCATTTGAAGACT	
		(((((((.(((((((((.((((((((((((((.(((((((..((.........)).)))))))).)))))))))))))))))))))))))))))).))))))))	
0144		TGAGATGAGATTACCCCATAC	Chr 5B:711070982 [+]
		TAGCAGCTGTTGAGATGAGATTACCCCATACCCTGCAGAGAGCGTCAGCTCCGACGCCGTACACGCGTATGGGTAATCTCAACTCAACAGCCGCTC	
		.(((.((((((((((.((((((((((((.((((...((.....(((((.....)))))).....)))))))))))))))))))))))))).))))))))).))).	
0145		TGAGATGAGATTACCCCATAC	Chr 5D:526099137 [-]

		CAGCGGCCGTTGAGATGAGATTACCCCATACCCTGTAGCGGCAGCTCGACTCCGTGCATGCGTAGGGGGCAATCTCACCTCAACAGCTGCTC	.(((((((.(((((((.(((((((.((((.(((..((((.(((.((.....)))))))))...))).)))).)))))))).))))))).....
0146		TGAGCCGAACCAATATCACTC	Chr 2B:603654988 [-]
		AAGAAGACATGCTGGTATTGTTTCGGCTCATATCCTCGTCGAGCACATCGTCGGATGAGATGTGAGCCGAACCAATATCACTCATGTATTCTT	(((((.(((((.(((((((((.(((((((((((((((.(((((..((.....))..)))))))))))))))))).)))))))..)))))).)))))
0147		TGATTGAGCCGTGCCAATATC	Chr 4B:55643747 [+]
		GGGAGAGTCCGATGTTGGCATGGCTCAATCAATTCGAGTGGGCAGCGATCGAGGCGTGATATGATTGTTTAACTCGATCTGATTGAGCCGTGCCAATATCGCAAGCT CCT	(((((....(((((((((((((((((((((((.. (((((((((((((((((.....)))).)).)))))))..)))))))))))))))))))))....)))))
0148		TGATTGAGCCGTGCCAATATC	Chr 4D:38287834 [+]
		GGGAGAGTCCGATGTTGGCATGGCTCAATCAAATCGAGTGGGCGGCGATCGAGGCGTGAAGATATGATTGTTTAACTCGATCTGATTGAGCCGTGCCAATATCGCAA GCTCCT	(((((....(((((((((((((((((((((((. (((((((((((((((((..........)))).)).)))))))..)))))))))))))))))))))....)))))
0149		TGATTGAGCCGTGCCAATATC	Chr 7B:526800482 [-]
		TGGATGGAGCGATATTGGTGGGGTTCAATCCAATGCTAGGTTTTACAACAGTTTGTAAGACCAGTATCTGATTGAGCCGTGCCAATATCTCTCCCTCTC	.(((.((((.((((((((((.(((((((((..((((((.((((((((((....))))))))))))))))..)))))))))).)))))))))).)))).
0150		TGATTGAGCCGTGCCAATATC	Chr 7B:526810774 [-]
		TGGATGGAGCGATATTGGTGAGGTTCAATCCGATGCTAGGTCTTACAACAGTGGTAAGACCAGTATCTGATTGAGCCGTGCCAATATCTCTCCCTCGC	..((.((((.((((((((((.(((((((((.((((((((((.......)))))))))))))))).)))))))))).)))))..))..
0151		TGCATCATCTCGAACTCGTCG	Chr 7A:725637594 [+]
		GTGGAAAACCTGCATCATCTCGAACTCGTCGGTGCGATAGCGTGCAGCGTGACATGGTGTTCATCACCGGCGAGTTCCAGATGATGCAGGCCTTCATC	(((((...((((((((((((.((((((((((((((.((((.((((((...)).)))).)))))...)))))))))))))).)))))))))))..)))))).
0152		TGCCAAAGGAGAGTTGCCCTG	Chr 2D:507824577 [-]
		AGTCCAGTTTCAGGGCTCCTCTTTATTGGCAGGGAGCGTGTGAGGCCATGTAGCTCCATTCAGCGCTCTGCCAAAGGAGAGTTGCCCTGTAACTGGAAA	..(((((((.((((((..((((((.((((((((((.((..(((.(((......))).)))...)).)))))))))).)))))).)))))))..)))))))..
0153		TGGACGAGGATGTGCAACTGC	Chr 5B:610499054 [+]
		AGCTAGAGCTTGGACGAGGATGTGCAACTGCGGTACGGTAGGACCTTCAGCTACGACCGGCAGCTGCACATCCATGTCCAAGCTCTAGCT	((((((((((((((((.((((((((((.(((((((.((.(((.........)))))))).)))))).)))))))))).))))))))))))))))
0154		TGGACGAGGATGTGCAGCTGC	Chr 2D:568229242 [+]

		AGTTAGAGTTTGGACGAGGATGTGCAGCTGCGGTAGGACTCTCAGCTACGACCGGCAGCTGCACATCCACGTCCAAGCGCTAGCT ((((((.(((((((((.((((((((((((((((((.((.((...)))).)))).)))))))))))))).)))))))))).))))))
0155		TGGGCAAGTCACCCTGGCTACC Chr 4A:620467156 [-] GGCCATCTTCGATAGCCAAGGATGATTTGTCTGTGAAACTCACCTGGCAGCCAAGCTCTCTGCCACAGAGAACAGGCGTGCAGCAGTTCCATGGGCAAGTCACCCTG GCTACCCAAGTACCCC ((....(((...(((((((.((.((((((((((((((.((((...(((.(((((.... ((((((...)))))))..))).)))))).)))))))))))))))))))))))...)))...)).
0156		TGGGCAAGTCACCCTGGCTACC Chr 4A:620498857 [-] GGCCATCTTTGATAGCCAAGGATGATTTGCCTGTGAAACTCGCCTGGCAGCCGAGAGAGCGGCGTCCGGCGGTTTCATGGGCAAGTCACCCTGGCTACCAGAGTACC TC ((....(((((.(((((((.((.(((((((((((((((((((.((((.(((. ((((.....)))))))).)))))))))))))))))))))))))))))).))))).)).
0157		TGGGCAAGTCACCCTGGCTACC Chr 5B:700650125 [-] GGACATATTCGATAGCCAAGGATGATTTGCCTGTTAAACTCACCTGGCAGCCAAGCTCTCTGCCACAGAGAATGGCGTACGGCAGTTCCATGGGCAAGTCACCCTGG CTACCCAAGTACCCC ((...((((...(((((((.((.((((((((((((..((((...(((...(((((... ((((((...)))))))..))))...))).)))))..)))))))))))))))).))))...)))))..)).
0158		TGTCTGAAAAAACTTGTCATC Chr 6D:113464134 [+] ATACTCCCTCTGTCTGAAAAAACTTGTCATCAAAATGAATAAAAAGGCATGTATCTAGGACTAATATACATCTAAATACATCCCCTTTTATTCATTTTGATGACAAG TATTTCCGGACGGAGGGAGTAT ((((((((((((((((.(((.(((((((((((((((((((((((((((.((.((((((. ((((...........)))))..)))))))..)))))))))))))))))))))))))))).)))))))))))))))))
0159		TGTCTGAAAAAACTTGTCATC Chr 6A:585007709 [-] GTACTCCCTCTGTCTGAAAAAACTTGTCATCAAAATGGACAAAAATGGGTGTATCTAGAAATGATGACAAAAAGGGGTTTTGTCCATTTTGATGACAAGTATTTCCG GACAGATGGAGTAT (((((((.(((((((((.(((.(((((((((((((((((((((((((((.....((((((.......)))))..)).)).........)))))))))))))))))))))))))))).)))).))))))))
0160		TGTGTTCTCAGGTCGCCCCCG Chr 3A:658244036 [-] GAAATTCCTGCGGGGTCGAGCTGGGAACACATGAGATATGAACCGCTTGATTGAATATCAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTC ((..((((.(((((((.(((.((((((((((((((((((...........(((((.....)))))))))))))))))))))))).)))).)))))))..))
0161		TGTGTTCTCAGGTCGCCCCCG Chr 3D:523327227 [-] GAAATTCCTGTGGGGTCGAGCTGGGAACACATGAGATATGAACCGCTTGATTGACGAGCGTCTCTTATCAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTC ((..((((.(((((((.(((.((((((((((((((((((((.((..((((((.....)))))))..)).)))))...)))))))))))))))).)))).)))))))..))
0162		TGTGTTCTCAGGTCGCCCCCG Chr 3D:523446451 [-] GAAGTTCCTGCAGGGTCGAACTGAGAACACATGAGATATATTGACGAGCGACTAACTCATGTGTTCTCAGGTCGCCCCCGCTGGAGCCTC ((.(((((.((.(((.(((.(((((((((((((((((.((..(((.....))).)).)))))))))))))))).))).))).)).)))))).))
0163		TGTTGGCTCGACTCACTCAGA Chr 4D:487088261 [+] GTCACTATGATGTTGGCTCGACTCACTCAGACCACGCCGGCCGGCCGGCCGGAGCCATCCGCGGCGGTGGTTCTGATTGAGCCGTGCCAATATCTTAGTGAT (((((((.((((((((((.((.((((.((((((((((.((((((....)))))))..(((......))).)))))..)))))).))))).)).)))))))))).)))))))
0164		TGTTGGCTCGACTCACTCAGA Chr 5A:668149299 [+] GTCACTATGATGTTGGCTCGACTCACTCAGACCACGCCTGCCGGCCGGCCGTAGCCATGCATCTGCATGCGGTGGTGGCTCTGATTGAGCCGTGCCAATATCTCAGT GCT

			
		..((((..((((((((((.((.((((.(((((((((((....))))..((((.... (((((....)))))))))).)))))..))))))..))))..)).))))))))))..)))))..	
0165		TTAATTTGTCCATAGCATCCG	Chr 3D:524220889 [-]
		GTTATCCGGTTTAATTTGTCCATAGCATCCGCGATCCCGCCATTCAGCTGTGATAGCTGCTGCCCTATTGCACGGTGCGGAAGATAGCTGGTGCGATGGATAAATTC AACCCGATGGT	
		.((((((.((((((.(((((((((((((.((((((.((.(((((((..... ((((((((((........))))..)))))))))))..)))).)).))))))..)))))))))))).)))))..)))))).	
0166		TTAGATGACCATCAGCAAACA	Chr 2B:528817821 [+]
		GCTCTGAACTTGTTTTGTTGGTTGTCATCTAACCATCGTTCGATCTGCCACCTGGTGCATGGTGGATGCCTGGGTGCATGGCGATTGCGCCAACACGCAGACGGTGA TGATGGTTAGATGACCATCAGCAAACATGTTCATGAG	
		.(((((((((.((((((.(((((((.(((((((((((((((((((((.((((((((((((.((((.....)))))))))))).. ((((((....))))))....)))))))..)))))))))))))))))))).))))))))))).)))))))))))))	
0167		TTAGATGACCATCAGCAAACA	Chr 2D:449570581 [-]
		GCTCTGAACTTGTTTTGTTGGTTGTCATCTAACCATCGTTCGATCTGCCACCTGGTGCATGGATGGATGCCTGGGTGCATGGCGATTACGCCAACACACACACGGTG ATGATGGTTAGATGACCATCAGCAAACATGTTCATGAG	
		.(((((((((.((((((.(((((((.(((((((((((((((((((.....(((.(((((((.((((.....)))))))))))))))))))((((....))))))... (((.....)))))))))))))))))))))).))))))))))))))).))))))))))))))))))))

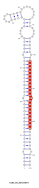
		(((((..(((((((((((((((((.((((((((((((((.(((.....(((((((...((..(((((((((.((. ((........)).)).)))))))))).)).)))))).))).....))).)).)))))).)))))).)))))))))))))))))..))))))
0174		TTCCACAGCTTTCTTGAACTT Chr 6B:684014623 [+] TGCCATGCCTTTCCACAGCTTTCTTGAACTTCTCTTGCGCCCCTCCTTAATTACTCCCCACTCTCTCCTCTCCTCTCTGTTGGTAGAGAGAGACATGAACGG GGAGGGAGAACAGGAGCAGGAGCAGCAAGAAGAGGGTCAAGAAAGCTGTGGGAAGAAATGGCA ((((((..(((((((((((((((((.((((((((((((((((((...((.(((((..(((..((. ((((((((((........)).)))))))))...))..)))..)))))))).))))))..)).)).)))))).))))))))))))))))))..))))))
0175		TTCCACAGCTTTCTTGAACTT Chr 6D:450527454 [+] TGCCATGCCTTTCCACAGCTTTCTTGAACTTCTCTTGCGCCTCTCCTTAATTACTCCCCACTCTCTCTGTCTCTCCCTTTCTCTCCGTTGGTAGAGAGAGACATGAA TGGGGAGAGAGAACAGGAGCAGGAGCAGCAAGAAGAGGGTCAAGAAAGCTGTGGGAAGAAATGGCA ((((((..(((((((((((((((((.((((((((((.((((((((...((.(((((((...((..(((((((((.((. ((........)).)).)))))))))).)).)))))))).))....))))))..))).))))..))))))).))))))))))))))))))..))))))
0176		TTCGGACCAGGCTTCATTCCC Chr 4D:450196376 [+] GGCAGTTGAGGGGAATGTTGTCTGGTTGGAGACCTGGCTCGCACGACGGATCGATCGGCGGCGGGCGTGCGCGCGCGGTTTCGGACCAGGCTTCATTCCCCCCAACT CAA ...(((((.(((((((((..(((((((((.((((((..((.((((((.((.. (((.....)))))..)))))))).))..)))))))).)))))))..)))))))).))))))...
0177		TTCGGACCAGGCTTCATTCCC Chr 4A:16475408 [-] GGCAGTTGAGGGGAATGTTGTCTGGTTGGAGACCTGGCTCGCACGACGGCCGGACCCCGGATCGATCGACGGCGTGCGTGCGCGGTTTCGGACCAGGCTTCATTCCC CCCAACTCAA ...(((((.(((((((((..(((((((((.((((((..((.((((((. (((((((...)))))..)))..)))))))).))..))))))).)))))))..)))))))).))))))...
0178		TTCGGACCAGGCTTCATTCCC Chr 5A:578021264 [-] GGTTGTCGAGGGGAATGACGCCGGGTCCGAAAGAGAGACGCATGGCGTGCGTGTGGCGCGTTTCGGACCAGGCTTCATTCCCCATGACACCA (((.((((.((((((((((.(((.((((((((((......((((((.....))))))).....)))))))))).)))).)))))))))..))))))).
0179		TTCGTGCCGGATTATCATTTC Chr 6D:423797499 [-] AATATATTGTTGGATGATAATATGGTACCGAAAATAGCTGACTTTGGTCTCTCAAAGTACTTTCGTGCCGGATTATCATTTCAAAATTTGGA .........((((((((((((.(((((((.((((.......((((((......)))))))..)))))))))).)))))))).))))).........
0180		TTCGTGCCGGATTATCATTTC Chr 6D:423797499 [-] ATATATTGTTGGATGATAATATGGTACCGAAAATAGCTGACTTTGGTCTCTCAAAGTACTTTCGTGCCGGATTATCATTTCAAAATTTGGA ........((((((((((((.(((((((.((((.......((((((......)))))))..)))))))))).)))))))).))))).........
0181		TTCGTGCCGGATTATCATTTCA Chr 6D:423797498 [-] TAATATATTGTTGGATGATAATATGGTACCGAAAATAGCTGACTTTGGTCTCTCAAAGTACTTTCGTGCCGGATTATCATTTCAAAATTTGGAT .........((((((((((((.(((((((.((((.......((((((......)))))))..)))))))))).)))))))).))))).........
0182		TTGAACATCCCAGAGCCACCG Chr 6A:112588842 [-] TCGGGGAGCCCGGCGGCTCTCTGGTGTTCAAGCAGGAACCTCATGCTACCGGCAGCATGCGGCGCTTGCTTGAACATCCCAGAGCCACCGGCGTGCCAAA ....((.(((((((.(((((((..((((((((((((((..((.((((((.....)))))))..))..)))))))))))))..)))))))..)))))..)).))...
0183		TTGACAGAAGAGAGTGAGCAC Chr 3A:76564145 [+] TGTGTGGAAGTTGACAGAAGAGAGTGAGCACCCATGGTGTTTCCTTAGCATCAAGGGGCATGCCAGGGAGCTATGCGTGCTCACTGCTCTATCTGTCAGCCACTGAT CA ...((((...((((((((((((((((((((.(((((((...(((((. ((((........)))))..)))))))))))).)))))))))).)))))..))))))))))).....
0184		TTGACAGAAGAGAGTGAGCAC Chr 6A:127182599 [+]



CCTCGAGAGATTGACAGAAGAGAGTGAGCACACGGCGTGATGCCGGCATAACATGTATGCCGTCTTCGCCGCGTGCTCACTCCTCTTTCTGTCAGCCTCTTTCTA

....(((((.((((((((((((((((((((.(((((.((...(((((((.....)))))))))..))))))..)))))))))).))))..)))))))))).(((((....

0185

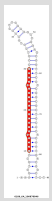


TTGAGACGAAAACAGACCAAC Chr 5D:83923044 [-]

GGAGTCTCTGGTTGGTCTGTGTTTGTTTCAAACAAGGTGGTCAGGCGTACGTCCTGTGCCTCTCTTTGTTGCTGCACCTCGTTTGAGACGAAAACAGACCAACCGGA
GTCTCC

((((.(((((((((((((((.((((((((((((.(((((....
(((((((.....))))).........)))....))))))..)))))))))))))).)))))))))))))))).))))

0186

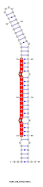


TTGAGTGCAGCGTTGATGAAC Chr 6A:106878846 [-]

AAAGGTGCCATTGAGTGCAGCGTTGATGAACCGTCCTACTCCTCCGAGGCCGGAGCGGTTCACCGGCGCTGCACACAATGACGGCTCT

..((.((.(((((.((((((((((((.((((((((((..(((....)))...))).)))))))).)))))))))))))..))).))..

0187

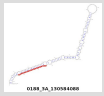


TTGAGTGCAGCGTTGATGAAC Chr 6B:169172651 [-]

AAAGGTGCCATTGAGTGCAGCGTTGATGAACCGTCCGGCCTCCTCCGAGGCCGGAGCGGTTCACCGGCGCTGCACACAATGACGCCTCT

..(((((.(((((.((((((((((((.((((((((((((((....)))))))))).)))))))).)))))))))))))..)))))..

0188



TTGGACTGAAGGGTGCTCCCT Chr 3A:130584088 [+]

GATGGACGGAAGAGAGCGTCCTTCAGTCCACTCACAGGCGGTGCTAGGATCGAATTAGCTGCCGACTCATTCGTCCACTTGCCGAGCAAAGTTATTTGCCGAGGAGC
TTAGCAGGTGAGTGAATGAAGCGGGAGGTAAAGTCTTTGATCTCGCACCGTCGTTGCTTGGACTGAAGGGTGCTCCCTCCTCCTCTCT

((.(((.((.((.((((((((((((((((((..((..(((((((((..(((((((((((.((.(((..(((((((..(((((((.
((((.....................)))).)))))))..)))))))..))))..)).)))....)))))))).)))))))..)).)))))))))))))))))...

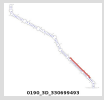
0189

TTGGACTGAAGGGTGCTCCCT Chr 3A:446540935 [-]

CTTCTGTGTGAGAGAGCTCTCTTCAGTCCACTCTCAGATGGTTGTAGGGTTTAATTAGCTGCCGATTCATTCATTCACCTACCAAGAACCAAGGCGGAGCGCATGTC
TTGGTAGCCGACTGGATGACGCGGGAGCTAATATTTAGCTCGACGTTGTTTGTGGTTGGACTGAAGGGTGCTCCCTTGCTCGCAGT

...((((((((((.((((.((((((((((((.((.((((((..(((((.((((((.(((..(((((((.(((..(((((((..((..(((((((((..((..
(((...))).)))))))))))..))..))))))))..))))..)))))))....)))))).)))..)))))))..)).)))))))))))))).)))))..)))))..

0190



TTGGACTGAAGGGTGCTCCCTT Chr 3D:330699493 [-]

CCTTCTGTGTGAGAGAGCTCTCTTCAGTCCACTCTCAGATGGTTGTAGGGTTTTATTAGCTGCCGATTCATTCATTCACCTACCAAGAACCAAGGCGGAGCGCATGT
CTTGGTAGCCGACTGGATGACGCGGGAGCTAATATTTAGCTCGACGTTGTTTGTGGTTGGACTGAAGGGTGCTCCCTTGCTCGCAGTT

....((((((((((.((((.((((((((((((.((.((((((..(((((.((((((.(((..(((((((.(((..(((((((..((..(((((((((..((..
(((...))).)))))))))))..))..))))))))..))))..)))))))...)))))).)))..)))))))..)).)))))))))))))).)))))..)))))..

0191

TTTAAAACTTTGACTTGGAAC Chr 4D:494149427 [+]

TCTAGCTTTTGTCTTAAGTCAAAATTTTAAAATTTTGACCACCTTTTTAGAAAAAAGTAAAAACATTTATGACACCAAATTAGTACCACTAGATTTGTTTTGAAATG
TACTTTAATATTATACCAATTTGATGTCATATATGTTACTACTATTTTCTTAAAGTTGGTCAAAATTTTAAAACTTTGACTTGGAACAAAAGCTAGA

((((((((((((.((((((((((.((((((((((((((((((.((((..((((((.((((..((((((.(((((((((((.....
((.....))).....)))))).............))))))..)))))))..)))))..))))..)))))))..)))))..))))))))))))))))))).)))))))))).)))))))))))))

0192

TTTAAAACTTTGACTTGGAAC Chr 5A:177623725 [-]

TCTAGATTTGGTCCCAAGTCAAAGTTTTAAAACTTTGACCAACTTTATCGGGAAAAGTAGCAGCATTTATGGCACTAAAGTAGTATCACTAGATTCGTTTTGAAATG
TACTTTCGTAATATACCAATTTTATGTCATATGTTACTATTATTTTTTATAAAGTTGGTCAAAATTTTAAAACTTTGACTTGGAACAAAACGTAGA

((((..(((.((.((((((((((((((((((((.((((((((((((((((.(((((.(((((.((((..(((((((.((((((.((((.....((..
((((....)))).....)).....))))..))))))))))))))))))).)))))).)))))).)))))))))))))))).)))))))))))))))))))..))..))))

0193

TTTAAAACTTTGACTTGGAAC Chr 7B:196712840 [-]

TCTAGCTTTTGTCTTAAGTCAAAGTTTTAATACTTTGACCACCTTTCTAGGAAAGAGTAAAAGCATTTATGGCACCAAATTAGTATCACTAGATTCGTTTTGAAATG
TACTTTCATAATATATCAATTTGATGTCATATATGTTATTACTATTTTCTATAAAGTTGGTCAAAATTTTAAAACTTTGACTTGGAACAAAATCTAGA

(((((.((((((.((((((((((((((((((...(((((((((.((((.((((((((.((((((.((((((.((((((((...)))))))(((..(((.
(((((.....))))))..)))..))).))))))).)))))))).))))).)))))).)))))))...))))))))))))))))))))))).))))))).)))))

0194

TTTATGGACCGCGGACAACGAGGA Chr 1B:459805475 [-]

ATCGTCTTCATCCTCGTTGTCCGCAGTCCATAAAAAAGTATTTTCTGATGAAAACTTTAGGTATGATTCACCCTTACTCTGTTGCTTCATCCATCTTCCATGCAACC
TGAGAAGTTCTTTCCGTATATCCTCAAAGGTCCTGGAAAGAACTTCTCAGGTTGCATGGAAGATGGATGAAGCAATAGAGTAAGGGTGAATCATACCTAAAGTTTGC
AGCAGAAAAGTTTTTATGGACCGCGGACAACGAGGATGAAGACGAT

		((((((((((((((((((((((((.((((((((((.((((((((((.((. (((((((((((((((((((((((((

B.6 S19: iRNA annotations of 339 iRNAs

#NewID	#iRNAseq	miR_family	preChrLoci	preDesc	psRNATarget(Expectation)	tarXLOCdesc	Regulation (Cstv:constitutive; CRup/CRdw:cold response up or down)	Prediction in this study	iRNA novelness (from A to H categories)
M-001	AAAAAUUGCGGAUUUUUGUCU	noHit	7A,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_081366(1.5),XLOC_088588(1.5),XLOC_139134(1.5),XLOC_130378(1.5),XLOC_071855(1.5),XLOC_063850(2.0),XLOC_133505(2.0)	Cysteine-rich receptor-like protein kinase 15;WD_REPEATS_REGION domain-containing protein;GRF-type domain-containing protein;Protein kinase domain-containing protein	Cstv	miRNA	H_novel
M-005	AAGCUCAGGAGGGAUAGCGCC	miR390	5A,5B,5D	MIR,microRNA;	XLOC_030418(2.0)	Protein kinase domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot
M-025	AGCCAACAACCUCCUAGUUCC	noHit	5B	DNA-directed RNA polymerase subunit beta	XLOC_016489(0.5),XLOC_054813(1.0),XLOC_019229(1.5),XLOC_028841(1.5),XLOC_142228(2.0),XLOC_102719(2.0),XLOC_048649(2.0),XLOC_057174(2.0),XLOC_074186(2.0)	Dof-type domain-containing protein;Cellulose synthase-like protein D3;1,4-beta-D-glucan synthase 7) (Cellulose synthase-like protein F7) (OsCslF7);Protein kinase domain-containing protein;DNA-directed RNA polymerase subunit beta;TPT domain-containing protein;Os06g0284500 protein;Omega-3 fatty acid desaturase, chloroplastic;Putative mixed-linked glucan synthase 7;Probable mixed-linked glucan synthase 7	CRdw	miRNA	F_PSUSStaeMirSir
M-027	AGGUGGGAGGUUGUUGGCUGG	noHit	5B	DNA-directed RNA polymerase subunit beta	XLOC_094640(1.5),XLOC_049350(1.5),XLOC_047884(1.5),XLOC_145163(2.0),XLOC_053987(2.0),XLOC_054353(2.0)	Glucan endo-1,3-beta-glucosidase 14;Glucan endo-1,3-beta-glucosidase GII;Beta-1,3-glucanase;Glucan endo-1,3-beta-glucosidase GI;Os06g0607000 protein (Fragment)	CRupCRdw	miRNA	F_PSUSStaeMirSir
M-028	AUACCACGACUGUCAUUAAGC	miR9672	5B	mRNA	XLOC_107815(0.5)	F-box only protein 7	CRdw	miRNA	A_MiRBaseTae
M-030	AUAGCAUCAUCCAUCCUACCC	miR9674	4B,4D,5A,6A,6B,6D		XLOC_123488(1.0),XLOC_117057(1.0),XLOC_007609(1.5),XLOC_017237(1.5),XLOC_150040(1.5),XLOC_067022(2.0),XLOC_129417(2.0),XLOC_116887(2.0),XLOC_108117(2.0),XLOC_092129(2.0),XLOC_050782(2.0),XLOC_052955(2.0),XLOC_054186(2.0),XLOC_079645(2.0)	Putative bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1;NB-ARC domain-containing protein;PPR_long domain-containing protein;ANK_REP_REGION domain-containing protein;F-box domain-containing protein;Os04g0514500 protein (cDNA clone:J033104I01, full insert sequence);Protein Rf1, mitochondrial	Cstv	miRNA	A_MiRBaseTae
M-032	AUCAGGAGAGAUGACACCGAC	miR1432	2A,2B,2D		XLOC_102168(0.0),XLOC_003288(0.5),XLOC_008047(0.5),XLOC_000994(0.5),XLOC_015506(0.5),XLOC_058856(0.5),XLOC_115353(1.0),XLOC_014564(2.0),XLOC_073716(2.0)	Os03g0812800 protein;Calcium-transporting ATPase;EF hand family protein, expressed (Os03g0812400 protein) (cDNA clone:002-184-C05, full insert sequence);Tryptophan aminotransferase 1;DNA-directed DNA polymerase;Calmodulin-like protein 5;Alliinase_C domain-containing protein	CRdw	miRNA	B_MiRBaseMonocot
M-045	CCACGACUGUCAUUAAGCAUC	miR9672	5A,5B,5D	mRNA;F-box only protein 7;	XLOC_107815(0.0)	F-box only protein 7	Cstv	miRNA	A_MiRBaseTae
M-046	CCAUGAUGAGGUCGUUCAACC	noHit	7A,7D	mRNA;Os07g0417800 protein;	XLOC_103275(2.0),XLOC_038702(2.0)	Clathrin light chain 1;Clathrin light chain;Os12g0554800 protein (Fragment);Putative polygalacturonase	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-050	CCCGCCUUGCACCAAGUGAAU	miR168	6A,6B	MIR,mRNA,miRNA,microRNA;			Cstv	miRNA	B_MiRBaseMonocot
M-067	CUUGAACUUCUCCAUAGCAUC	noHit	4D		XLOC_058416(1.5),XLOC_026243(2.0),XLOC_078142(2.0)	Epimerase domain-containing protein;Exostosin domain-containing protein;Os01g0926700 protein (Fragment);Dihydroflavonol-4-reductase	Cstv	miRNA	E_LiteratureTaeMiRNA2020
M-068	CUUGCUGCGUUGUUUGGGUCU	noHit	4D	siRNA,TAS,TASiRNA,SiRNA,mRNAAlkaline/neutral invertase	XLOC_082991(0.0),XLOC_053216(1.5),XLOC_028889(2.0),XLOC_039142(2.0),XLOC_076613(2.0),XLOC_069814(2.0)	Disease resistance protein RPP13;Kinesin-4;NB-ARC domain-containing protein;Alkaline/neutral invertase;Target of rapamycin complex subunit LST8;Target of rapamycin complex subunit LST8 (TORC subunit LST8) (Lethal with SEC13 protein 8 homolog) (WD40 repeat-containing protein 81) (OsWD40-81)	Cstv	siRNA	F_PSUSStaeMirSir
M-069	CUUGGACUGAAGGGUGCUCCCU	miR319	3A,3B,3D	MIR,mRNA,microRNA;Os01g0222001 protein (Fragment);	XLOC_060579(0.0),XLOC_061121(1.0),XLOC_049362(1.0),XLOC_001904(2.0),XLOC_006959(2.0),XLOC_012831(2.0),XLOC_145942(2.0)	Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase;Phospholipase D;Os01g0222001 protein (Fragment)	Cstv	miRNA	A_MiRBaseTae
M-073	GAGCUCCUAUCAUUCCAAUGA	miR159	3B	miRNA,microRNA,MIR,mRNA	XLOC_068598(0.0),XLOC_061039(0.0),XLOC_052503(0.0)		Cstv	siRNA	B_MiRBaseMonocot
M-086	GCGGCUCUGUGGUGUUCAAGC	noHit	6B,6D				Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-104	GUAGGAUGGCCGGUGCUAUGG	noHit	5A				Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-105	UAAAAAUUGCGGAUUUUUGUC	noHit	7A,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_081366(2.0),XLOC_088588(2.0),XLOC_139134(2.0),XLOC_133505(2.0),XLOC_139391(2.0),XLOC_145885(2.0),XLOC_130378(2.0),XLOC_130100(2.0),XLOC_048529(2.0),XLOC_071855(2.0)	Cysteine-rich receptor-like protein kinase 15;1-phosphatidylinositol-3-phosphate 5-kinase;GRF-type domain-containing protein;Protein kinase domain-containing protein	Cstv	miRNA	H_novel

M-106	UAAUAAACAAGUCUUUAGAUG	noHit	1A,1B	,miRNA;	XLOC_012297(0.0),XLOC_012379(0.5),XLOC_016786(0.5),XLOC_021552(0.5),XLOC_018567(0.5),XLOC_135172(0.5),XLOC_135860(0.5),XLOC_139841(0.5),XLOC_148918(0.5),XLOC_006001(1.0),XLOC_004769(1.0),XLOC_016889(1.0),XLOC_017858(1.0),XLOC_011614(1.0),XLOC_011090(1.0),XLOC_011706(1.0),XLOC_134106(1.0),XLOC_137699(1.0),XLOC_117784(1.0),XLOC_005733(1.5),XLOC_023595(1.5),XLOC_137538(1.5),XLOC_150749(1.5),XLOC_128697(1.5),XLOC_103942(1.5),XLOC_106325(1.5),XLOC_056510(1.5),XLOC_069698(1.5),XLOC_005990(2.0),XLOC_005950(2.0),XLOC_007823(2.0),XLOC_009582(2.0),XLOC_000183(2.0),XLOC_017621(2.0),XLOC_016787(2.0),XLOC_017929(2.0),XLOC_014253(2.0),XLOC_024162(2.0),XLOC_063832(2.0),XLOC_143020(2.0),XLOC_142786(2.0),XLOC_142076(2.0),XLOC_135404(2.0),XLOC_135953(2.0),XLOC_134135(2.0),XLOC_134493(2.0),XLOC_145269(2.0),XLOC_150402(2.0),XLOC_150538(2.0),XLOC_033560(2.0),XLOC_072474(2.0)	Putative disease resistance RPP13-like protein 3;CNL-C (Disease resistance RPP13-like protein 4);Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13;AAA domain-containing protein;Disease resistance RPP8-like protein 3;Rx_N domain-containing protein; (Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;Disease resistance RPP13-like protein 4;Calcium-dependent protein kinase 5;Glutaredoxin-C8	Cstv	miRNA	F_PSUSStaeMirSir
M-107	UAAUAAACAUGUCUUCAGAUG	noHit	1A	miRNA,mRNA	XLOC_024162(0.0),XLOC_128777(0.5),XLOC_129114(0.5),XLOC_128038(0.5),XLOC_106264(0.5),XLOC_096513(0.5),XLOC_142300(1.0),XLOC_148304(1.0),XLOC_069560(1.0),XLOC_018567(1.5),XLOC_142786(1.5),XLOC_148918(1.5),XLOC_145269(1.5),XLOC_115775(1.5),XLOC_122257(1.5),XLOC_119517(1.5),XLOC_105127(1.5),XLOC_098475(1.5),XLOC_105035(1.5),XLOC_100358(1.5),XLOC_044636(1.5),XLOC_033436(1.5),XLOC_033560(1.5),XLOC_006001(2.0),XLOC_004769(2.0),XLOC_002347(2.0),XLOC_016889(2.0),XLOC_017858(2.0),XLOC_011614(2.0),XLOC_011090(2.0),XLOC_011706(2.0),XLOC_015670(2.0),XLOC_134106(2.0),XLOC_122786(2.0),XLOC_117940(2.0),XLOC_098539(2.0),XLOC_105375(2.0),XLOC_097983(2.0),XLOC_090286(2.0),XLOC_035582(2.0),XLOC_043757(2.0)	Putative disease resistance RPP13-like protein 3;BTB/POZ domain and ankyrin repeat-containing protein NPR1 (OsNPR1) (NPR1 homolog 1) (OsNH1);Os02g0177800 protein;Os02g0177800 protein (Fragment);Calmodulin binding protein;Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13;Disease resistance RPP8-like protein 3; (Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;Regulatory protein NPR1;Disease resistance RPP13-like protein 4;Fucosyltransferase	Cstv	miRNA	F_PSUSStaeMirSir
M-109	UACUGUGGGCACUUAUUUGAC	miR9669	3D		XLOC_056752(2.0)		Cstv	miRNA	A_MiRBaseTae
M-110	UACUGUGGGCACUUAUUUGACA	miR9669	3D		XLOC_056752(2.0)		Cstv	miRNA	A_MiRBaseTae
M-113	UAUUGGCUCGGCUCACUCAGG	noHit	1A,1B,1D		XLOC_094074(2.0)		Cstv	miRNA	E_LiteratureTaeMiRNA 2020
M-114	UCCAAAGGGAUCGCAUUGAUC	miR393	2A,2B,2D,3A, 3D		XLOC_003903(1.0),XLOC_011568(1.0),XLOC_017972(1.0),XLOC_022116(1.0),XLOC_030681(1.0),XLOC_037090(1.0)	Transport inhibitor response 1-like protein Os05g0150500 (TIR1-like protein);Transp_inhibit domain-containing protein;Transport inhibitor response 1-like protein Os04g0395600 (TIR1-like protein);F-box domain-containing protein;Transport inhibitor response 1	Cstv	miRNA	B_MiRBaseMonocot
M-115	UCCACAGGCUUUCUUGAACUG	miR396	2A,2B,2D,6A, 6D		XLOC_022666(0.0),XLOC_030908(0.0),XLOC_087793(0.0),XLOC_080318(0.0),XLOC_133553(0.0),XLOC_148284(0.0),XLOC_124078(0.0),XLOC_126012(0.0),XLOC_114027(0.0),XLOC_112641(0.0),XLOC_112330(0.0),XLOC_119737(0.0),XLOC_036899(0.0),XLOC_069277(0.0),XLOC_070618(0.0),XLOC_027870(1.0),XLOC_022258(1.0),XLOC_088721(1.5),XLOC_087284(1.5),XLOC_080810(1.5),XLOC_070536(1.5),XLOC_110439(2.0)	Growth regulating factor 4-D1;Os03g0729500 protein;Growth regulating factor 5-A1;Growth regulating factor 6-B1;Growth-regulating factor 4 protein (Fragment);Growth regulating factor 1-D1;Growth regulating factor 4a-B1;Importin N-terminal domain-containing protein;Growth regulating factor 6-D1;Growth regulating factor 9-A1;Growth regulating factor 1-B1;Importin-beta1 (Os05g0353400 protein) (cDNA clone:J013029L11, full insert sequence);Growth regulating factor 1-A1;Importin subunit beta-1;QLQ domain-containing protein;Growth regulating factor 5-D1;Growth regulating factor 6-A1;GRF transcription factor (Growth-regulating factor);Growth-regulating factor 4 (OsGRF4) (Protein GRAIN LENGTH AND WIDTH 2) (Protein GRAIN SIZE ON CHROMOSOME 2) (Protein PANICLE TRAITS 2) (Transcription activator GRF4);WRC domain-containing protein;Growth regulating factor 5-4A;Os04g0600900 protein;Growth-regulating factor 1 (OsGRF1) (Transcription activator GRF1);Growth regulating factor 4-A1	Cstv	miRNA	B_MiRBaseMonocot
M-121	UCGCUUGGUGCAGAUCGGGAC	miR168	6A,6B,6D				Cstv	miRNA	B_MiRBaseMonocot

M-122	UCGGACCAGGCUUCAAUCCCU	miR166	5A,5B,5D		XLOC_109041(0.0),XLOC_003329(1.0),XLOC_010176(1.0),XLOC_140819(1.0),XLOC_147220(1.0),XLOC_132764(1.0),XLOC_110868(1.5),XLOC_103056(1.5),XLOC_095467(1.5),XLOC_001247(2.0),XLOC_008070(2.0),XLOC_015254(2.0),XLOC_058789(2.0),XLOC_059246(2.0),XLOC_084967(2.0),XLOC_105100(2.0),XLOC_098448(2.0),XLOC_092135(2.0),XLOC_090000(2.0),XLOC_051179(2.0),XLOC_076905(2.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Protein kinase domain-containing protein;Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot
M-123	UCGGACCAGGCUUCAUUCCCC	miR166	1A,1B,1D,4A,4D,5A,5B,5D,7A,7B,7D		XLOC_003329(0.0),XLOC_010176(0.0),XLOC_140819(0.0),XLOC_147220(0.0),XLOC_132764(0.0),XLOC_110868(0.5),XLOC_103056(0.5),XLOC_095467(0.5),XLOC_001247(1.0),XLOC_008070(1.0),XLOC_015254(1.0),XLOC_059246(1.0),XLOC_084967(1.0),XLOC_109041(1.0),XLOC_105100(1.0),XLOC_098448(1.0),XLOC_092135(1.0),XLOC_090000(1.0),XLOC_051179(1.0),XLOC_076905(1.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot
M-124	UCGGACCAGGCUUCAUUCCCU	miR166	6A,6B,6D		XLOC_003329(0.0),XLOC_010176(0.0),XLOC_140819(0.0),XLOC_147220(0.0),XLOC_132764(0.0),XLOC_110868(0.5),XLOC_103056(0.5),XLOC_095467(0.5),XLOC_001247(1.0),XLOC_008070(1.0),XLOC_015254(1.0),XLOC_059246(1.0),XLOC_084967(1.0),XLOC_109041(1.0),XLOC_105100(1.0),XLOC_098448(1.0),XLOC_092135(1.0),XLOC_090000(1.0),XLOC_051179(1.0),XLOC_076905(1.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot
M-127	UCUGAAGACUAGUUUAUUACA	noHit	1A,1D	miRNA;	XLOC_006001(0.0),XLOC_012297(0.0),XLOC_012379(0.0),XLOC_016786(0.0),XLOC_016787(0.0),XLOC_021552(0.0),XLOC_018567(0.0),XLOC_005950(2.0),XLOC_022824(2.0),XLOC_135956(2.0),XLOC_052771(2.0)	CNL-C (Disease resistance RPP13-like protein 4);Disease resistance protein RPP13;AAA domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	miRNA	G_BlastnPreMirSir
M-128	UCUGCAAACUAAUAUAAGACU	noHit	2A,2B	,mRNA;	XLOC_143737(0.5),XLOC_005438(1.5),XLOC_006962(1.5),XLOC_002572(1.5),XLOC_000292(1.5),XLOC_006601(1.5),XLOC_008380(1.5),XLOC_002173(1.5),XLOC_004666(1.5),XLOC_007321(1.5),XLOC_002210(1.5),XLOC_021399(1.5),XLOC_022966(1.5),XLOC_017937(1.5),XLOC_013590(1.5),XLOC_015288(1.5),XLOC_017219(1.5),XLOC_014616(1.5),XLOC_013253(1.5),XLOC_015819(1.5),XLOC_012997(1.5),XLOC_019073(1.5),XLOC_017173(1.5),XLOC_014742(1.5),XLOC_030074(1.5),XLOC_022485(1.5),XLOC_025754(1.5),XLOC_026376(1.5),XLOC_062833(1.5),XLOC_059744(1.5),XLOC_056204(1.5),XLOC_066491(1.5),XLOC_064713(1.5),XLOC_067817(1.5),XLOC_066637(1.5),XLOC_054809(1.5),XLOC_067606(1.5),XLOC_084398(1.5),XLOC_088696(1.5),XLOC_089454(1.5),XLOC_084311(1.5),XLOC_083374(1.5),XLOC_084818(1.5),XLOC_087440(1.5),XLOC_087895(1.5),XLOC_091317(1.5),XLOC_080011(1.5),XLOC_085473(1.5),XLOC_138275(1.5),XLOC_140232(1.5),XLOC_138535(1.5),XLOC_134103(1.5),XLOC_136844(1.5),XLOC_141260(1.5),XLOC_135660(1.5),XLOC_147220(1.5),XLOC_142111(1.5),XLOC_140208(1.5),	Non-specific serine/threonine protein kinase;Rhodanese domain-containing protein;Os07g0530700 protein (cDNA clone:J013031M18, full insert sequence);ABC transporter B family member 29, chloroplastic;GlcNAc kinase;Na_H_Exchanger domain-containing protein;Nuclear export mediator factor Nemf;Os03g0332533 protein (Fragment);Putative inactive shikimate kinase like 2 chloroplastic;ER membrane protein complex subunit 6;Shikimate kinase;Os04g0689900 protein (Fragment);CS domain-containing protein;Os02g0664700 protein (cDNA clone:001-034-H05, full insert sequence);J domain-containing protein;AMP-binding domain-containing protein;Phosphotransferase;HHH_2 domain-containing protein;Os07g0170800 protein (Fragment);MADS-box transcription factor 29;Serine/threonine-protein kinase sepA;Potassium transporter;Glycosyltransferase;Expressed protein (Os12g0498800 protein) (cDNA clone:J013119P15, full insert sequence);Os01g0702450 protein;Transcription initiation factor TFIID subunit 2;Expressed protein (Os10g0577700 protein) (cDNA clone:001-037-H11, full insert sequence) (cDNA clone:J023053E21, full insert sequence);UVB	Cstv	miRNA	E_LiteratureTaeMiRNA2020

M-129	UGAACAUCCCAGAGCCACCGG	miR9662	6D		XLOC_127691(1.0),XLOC_114542(1.0),XLOC_114773(1.0),XLOC_115949(1.0),XLOC_115987(1.0),XLOC_116582(1.0),XLOC_109594(1.0),XLOC_037124(1.0),XLOC_079739(1.5),XLOC_127559(1.5),XLOC_128540(1.5),XLOC_127560(1.5),XLOC_115962(1.5),XLOC_115805(1.5),XLOC_122428(1.5),XLOC_121112(1.5),XLOC_109778(1.5),XLOC_053071(1.5),XLOC_033697(1.5),XLOC_033155(1.5),XLOC_033398(1.5),XLOC_034741(1.5),XLOC_007417(2.0),XLOC_143384(2.0),XLOC_121645(2.0),XLOC_128395(2.0),XLOC_121551(2.0),XLOC_127702(2.0),XLOC_121345(2.0),XLOC_114581(2.0),XLOC_114851(2.0),XLOC_115194(2.0),XLOC_121113(2.0)		Cstv	miRNA	A_MiRBaseTae
M-130	UGAACCUUAGGGAAAAGCCGCAU	miR5062	5D	H_noveleat shock cognate 70 kDa protein 1;	XLOC_106053(0.0),XLOC_058622(1.5),XLOC_106077(1.5),XLOC_092721(1.5),XLOC_092193(1.5),XLOC_105192(2.0)	Protein argonaute MEL1	Cstv	miRNA	A_MiRBaseTae
M-131	UGAACCUUGGGGAAAAGCCGCAU	miR5062	5A	H_noveleat shock cognate 70 kDa protein 2, putative, expressed (Os03g0821100 protein) (cDNA clone:J023030D03, full insert sequence);Heat shock 70 kDa protein 3;Heat shock cognate 70 kDa protein 1	XLOC_106053(0.5),XLOC_001516(2.0),XLOC_006849(2.0),XLOC_013656(2.0),XLOC_058622(2.0),XLOC_106077(2.0),XLOC_092721(2.0),XLOC_092193(2.0),XLOC_049152(2.0)	Potassium transporter 3;Reverse transcriptase Ty1/copia-type domain-containing protein;Protein argonaute MEL1;Retrovirus-related Pol polyprotein from transposon TNT 1-94	Cstv	miRNA	A_MiRBaseTae
M-132	UGAAGCUGCCAGCAUGAUCUGA	miR167	4A,4D,5A,5B,5D				CRdw	miRNA	A_MiRBaseTae
M-133	UGAAGCUGCCAGCAUGAUCUGC	miR167	6A				Cstv	miRNA	A_MiRBaseTae
M-134	UGAAUUUGUCCAUAGCAUCAG	miR9674	4D		XLOC_117006(0.0),XLOC_123290(0.0),XLOC_016763(0.5),XLOC_018015(0.5),XLOC_004676(1.0),XLOC_005411(1.0),XLOC_010779(1.0),XLOC_018569(1.0),XLOC_018328(1.0),XLOC_016985(1.0),XLOC_061178(1.0),XLOC_129417(1.0),XLOC_004051(1.5),XLOC_011106(1.5),XLOC_016938(1.5),XLOC_010625(1.5),XLOC_012188(1.5),XLOC_017611(1.5),XLOC_004857(2.0),XLOC_001369(2.0),XLOC_004348(2.0),XLOC_010447(2.0),XLOC_000139(2.0),XLOC_017200(2.0),XLOC_018327(2.0),XLOC_011604(2.0),XLOC_012065(2.0),XLOC_017617(2.0),XLOC_026243(2.0),XLOC_030741(2.0),XLOC_118500(2.0),XLOC_079645(2.0)	PDZ_3 domain-containing protein;Pentatricopeptide repeat-containing protein mitochondrial;Secologanin synthase;Disease resistance protein RPM1;Protease Do-like 10, mitochondrial;Protein Rf1, mitochondrial	Cstv	miRNA	B_MiRBaseMonocot
M-136	UGAGAAAGGACUGCAUCAUCU	noHit	7B		XLOC_039460(0.0),XLOC_124452(0.5),XLOC_101404(1.0),XLOC_097820(1.0),XLOC_100228(1.0),XLOC_007467(1.5),XLOC_128720(1.5),XLOC_122763(1.5),XLOC_043149(1.5),XLOC_071922(1.5),XLOC_006323(2.0),XLOC_017292(2.0),XLOC_024582(2.0)	Rx_N domain-containing protein (Fragment);Putative disease resistance protein RXW24L;NHR domain-containing protein;NB-ARC domain containing protein (Os12g0198900 protein);Disease resistance protein RPM1 (Resistance protein RGA2);Disease resistance protein RPP13;Os09g0517200 protein;Rx_N domain-containing protein;Resistance protein RGA2;Disease resistance RPP13-like protein 4;Disease resistance protein RPM1	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-137	UGAGAAGGCAGAUCAUAAUAGC	miR9863	1B,1D	,mRNA,miRNA;	XLOC_012297(0.0),XLOC_012379(0.0),XLOC_016786(0.0),XLOC_016787(0.0),XLOC_021552(0.0),XLOC_018567(0.0),XLOC_063832(1.5),XLOC_017858(2.0),XLOC_062400(2.0),XLOC_143295(2.0),XLOC_142872(2.0),XLOC_142786(2.0),XLOC_142076(2.0),XLOC_135404(2.0),XLOC_139832(2.0),XLOC_148918(2.0),XLOC_150402(2.0),XLOC_146746(2.0),XLOC_092044(2.0),XLOC_047671(2.0),XLOC_057515(2.0)	Putative disease resistance RPP13-like protein 3;NAC domain-containing protein;CNL-C (Disease resistance RPP13-like protein 4);Putative disease resistance RPP13-like protein 2;Protein kinase domain-containing protein;Disease resistance protein RPP13;Os07g0166700 protein (cDNA, clone: J090033M17, full insert sequence);NAC domain-containing protein 78;NB-ARC domain-containing protein;Mitochondrial import inner membrane translocase subunit;Disease resistance RPP13-like protein 4	Cstv	miRNA	C_MiRBaseAll
M-138	UGAGAAGGUAGAUCAUAAUAGC	miR9863	1B,1D	,miRNA;	XLOC_012297(0.0),XLOC_018567(0.0),XLOC_012379(0.5),XLOC_016786(0.5),XLOC_016787(0.5),XLOC_021552(0.5),XLOC_017858(0.5),XLOC_062400(0.5),XLOC_047671(0.5),XLOC_057515(0.5),XLOC_018048(1.0),XLOC_134106(1.0),XLOC_118312(1.0),XLOC_004769(1.5),XLOC_011485(1.5),XLOC_005733(1.5),XLOC_002957(1.5),XLOC_009051(1.5),XLOC_011202(1.5),XLOC_010685(1.5),XLOC_016889(1.5),XLOC_011614(1.5),XLOC_011706(1.5),XLOC_013941(1.5),XLOC_135341(1.5),XLOC_138710(1.5),XLOC_150161(1.5),XLOC_146067(1.5),XLOC_058114(1.5),XLOC_056601(1.5),XLOC_046410(1.5),XLOC_072242(1.5),XLOC_016753(2.0),XLOC_062522(2.0),XLOC_063832(2.0),XLOC_062352(2.0),XLOC_139841(2.0),XLOC_141019(2.0),XLOC_122618(2.0),XLOC_121614(2.0),XLOC_124598(2.0),XLOC_122421(2.0),XLOC_098427(2.0),XLOC_047371(2.0)	Putative disease resistance RPP13-like protein 3;CNL-C (Disease resistance RPP13-like protein 4);Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13;AAA domain-containing protein;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);Rx_N domain-containing protein; (Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;NBR1-like protein;Mitochondrial import inner membrane translocase subunit;Disease resistance RPP13-like protein 4;Peroxidase	Cstv	miRNA	C_MiRBaseAll

M-140	UGCAUCAUCUCGAACUCGUCG	miR9778	7A,7B,7D	,mRNA;Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein;	XLOC_147261(0.0),XLOC_004785(0.5),XLOC_004095(1.0),XLOC_017292(1.0),XLOC_058711(1.0),XLOC_060090(1.0),XLOC_136162(1.0),XLOC_146010(1.0),XLOC_150570(1.0),XLOC_146796(1.0),XLOC_095247(1.0),XLOC_039460(1.0),XLOC_004987(1.5),XLOC_012250(1.5),XLOC_010815(1.5),XLOC_043149(1.5),XLOC_007467(2.0),XLOC_025363(2.0),XLOC_025542(2.0),XLOC_088900(2.0),XLOC_089124(2.0),XLOC_135317(2.0),XLOC_149072(2.0),XLOC_122763(2.0),XLOC_121477(2.0),XLOC_109872(2.0),XLOC_043884(2.0),XLOC_034177(2.0),XLOC_071922(2.0)	Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein (Fragment);Cysteine-rich receptor-like protein kinase 10;Putative disease resistance protein RXW24L;Disease resistance protein RPM1 (Resistance protein RGA2);Protein kinase domain-containing protein;Disease resistance protein RPP13;Resistance protein RGA2;AAA domain-containing protein;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;DUF676 domain-containing protein;Os08g0447200 protein	Cstv	miRNA	A_MiRBaseTae
M-142	UGGACGAGGAUGUGCAACUGC	miR9776	5B,5D		XLOC_128171(1.5),XLOC_114969(2.0)		CRdw	miRNA	A_MiRBaseTae
M-143	UGGACGAGGAUGUGCAGCUGC	noHit	2D		XLOC_096486(1.5),XLOC_101906(1.5),XLOC_059959(2.0),XLOC_087762(2.0),XLOC_138060(2.0),XLOC_148961(2.0),XLOC_128171(2.0),XLOC_104247(2.0),XLOC_097916(2.0),XLOC_038068(2.0),XLOC_072111(2.0)	Subtilisin-chymotrypsin inhibitor-2A;Auxin response factor;Transcription factor HBP-1b(C1);FLZ-type domain-containing protein;DUF4220 domain-containing protein	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-144	UGUUGGCUCGACUCACUCAGA	miR171	4B,4D,5A		XLOC_094074(0.0)		Cstv	miRNA	B_MiRBaseMonocot
M-146	UUCCAAAGGGAUCGCAUUGAU	miR393	2A,2B,2D		XLOC_003903(1.0),XLOC_011568(1.0),XLOC_017972(1.0),XLOC_022116(1.0),XLOC_030681(1.0),XLOC_037090(1.0),XLOC_124694(2.0),XLOC_118338(2.0)	Transport inhibitor response 1-like protein Os05g0150500 (TIR1-like protein);Transp_inhibit domain-containing protein;Transport inhibitor response 1-like protein Os04g0395600 (TIR1-like protein);F-box domain-containing protein;Transport inhibitor response 1;Os02g0529900 protein (cDNA clone:J013154H21, full insert sequence)	CRup	miRNA	B_MiRBaseMonocot
M-147	UUCCACAGCUUUCUUGAACUG	miR396	6A,6B,6D,7A,7B,7D	MIR,mRNA,miRNA,microRNA;	XLOC_126596(1.5),XLOC_120286(1.5),XLOC_085266(2.0)	eIF2B_5 domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot
M-148	UUCCACAGCUUUCUUGAACUU	miR396	6A,6B,6D		XLOC_126596(1.5),XLOC_120286(1.5),XLOC_085266(2.0)	eIF2B_5 domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot
M-149	UUCGGACCAGGCUUCAUUCCC	miR166	4A,4D,5A	MIR,mRNA,miRNA,microRNA;	XLOC_003329(0.5),XLOC_010176(0.5),XLOC_140819(0.5),XLOC_147220(0.5),XLOC_132764(0.5),XLOC_110868(0.5),XLOC_103056(0.5),XLOC_095467(0.5),XLOC_001247(1.0),XLOC_008070(1.0),XLOC_015254(1.0),XLOC_059246(1.0),XLOC_084967(1.0),XLOC_105100(1.0),XLOC_098448(1.0),XLOC_092135(1.0),XLOC_090000(1.0),XLOC_051179(1.0),XLOC_076905(1.0),XLOC_109041(1.5)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot
M-150	UUCGUGCCGGAUUAUCAUUUC	noHit	6D	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-29;Protein kinase domain-containing protein;			Cstv	miRNA	F_PSUSStaeMirSir
M-152	UUGAACAUCCCAGAGCCACCG	miR9662	6A,6B,6D		XLOC_127691(0.5),XLOC_128395(0.5),XLOC_121645(0.5),XLOC_114542(0.5),XLOC_114773(0.5),XLOC_115949(0.5),XLOC_115987(0.5),XLOC_116582(0.5),XLOC_109594(0.5),XLOC_037124(0.5),XLOC_079739(1.0),XLOC_127559(1.0),XLOC_127560(1.0),XLOC_128540(1.0),XLOC_115962(1.0),XLOC_115805(1.0),XLOC_122428(1.0),XLOC_121112(1.0),XLOC_109778(1.0),XLOC_053071(1.0),XLOC_033155(1.0),XLOC_033697(1.0),XLOC_033398(1.0),XLOC_034741(1.0),XLOC_007417(1.5),XLOC_010439(1.5),XLOC_143384(1.5),XLOC_128652(1.5),XLOC_129024(1.5),XLOC_121551(1.5),XLOC_127702(1.5),XLOC_128242(1.5),XLOC_114716(1.5),XLOC_121345(1.5),XLOC_114581(1.5),XLOC_114851(1.5),XLOC_115194(1.5),XLOC_121113(1.5),XLOC_120949(1.5),XLOC_109595(1.5),XLOC_076575(1.5),XLOC_143433(2.0),XLOC_149171(2.0),XLOC_150149(2.0),XLOC_128162(2.0),XLOC_127495(2.0),XLOC_121653(2.0),XLOC_121559(2.0),XLOC_127310(2.0),XLOC_128993(2.0),XLOC_127362(2.0),XLOC_121528(2.0),XLOC_127328(2.0),XLOC_121603(2.0),XLOC_127979(2.0),XLOC_128737(2.0),XLOC_128435(2.0),XLOC_128305(2.0),XL	Os12g0583900 protein (Fragment);Mitochondrial transcription termination factor family protein (mTERF family protein) (mTERF transcription factor);Sulfotransferase;Os06g0225100 protein	Cstv	miRNA	A_MiRBaseTae
M-153	UUGACAGAAGAGAGUGAGCAC	miR156	3A,3B,3D,6A,6B,6D	MIR,mRNA,miRNA,microRNA;	XLOC_006637(1.0),XLOC_127342(1.0),XLOC_122219(1.0),XLOC_116249(1.0),XLOC_101060(1.0),XLOC_135366(1.5),XLOC_142136(1.5),XLOC_148120(1.5),XLOC_128905(1.5),XLOC_121975(1.5),XLOC_115110(1.5),XLOC_063101(2.0),XLOC_060958(2.0),XLOC_079815(2.0),XLOC_084330(2.0),XLOC_110043(2.0),XLOC_104831(2.0),XLOC_100923(2.0),XLOC_032645(2.0)	Squamosa promoter-binding-like protein 16;TF-B3 domain-containing protein;Pumilio-like protein 1;MND1-interacting protein 1;STAS domain-containing protein;Nucleolin;L-ascorbate oxidase-like protein;Squamosa promoter-binding-like protein 3;Os09g0491532 protein;SBP-type domain-containing protein;Hydrolase_4 domain-containing protein;GLW7;Squamosa promoter-binding-like protein 2;NAB domain-containing protein;Squamosa promoter-binding-like protein 13;GW8-B1;SBP transcription factor (Squamosa promoter-binding-like protein 2);SBP protein	CRdw	miRNA	A_MiRBaseTae

M-154	UUGGACUGAAGGGUGCUCCCU	miR319	3A,3B,3D	MIR,mRNA,microRNA;Os01g0222001 protein (Fragment);	XLOC_061121(0.0),XLOC_049362(0.0),XLOC_060579(0.0),XLOC_019622(1.5),XLOC_022594(1.5),XLOC_031031(1.5),XLOC_109753(1.5),XLOC_110435(1.5),XLOC_103310(1.5),XLOC_110530(1.5),XLOC_094745(1.5),XLOC_100277(2.0)	Transcription factor PCF6;Os07g0152000 protein;Transcription factor PCF8;TCP domain-containing protein;Transcription factor TCP2;2'-phosphotransferase;TCP transcription factor (Transcription factor TCP2) (Fragment); Phosphatidylserine decarboxylase 1 alpha chain];TCP transcription factor (Transcription factor TCP2);DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;R domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-155	UUGGACUGAAGGGUGCUCCCUU	miR319	3A,3B,3D		XLOC_061121(0.0),XLOC_049362(0.0),XLOC_060579(0.0),XLOC_019622(1.5),XLOC_022594(1.5),XLOC_031031(1.5),XLOC_110435(1.5),XLOC_110530(1.5),XLOC_103310(1.5),XLOC_109753(1.5),XLOC_094745(1.5),XLOC_100277(2.0)	Transcription factor PCF6;Os07g0152000 protein;2'-phosphotransferase;TCP domain-containing protein;Transcription factor TCP2;Transcription factor PCF8;TCP transcription factor (Transcription factor TCP2) (Fragment); Phosphatidylserine decarboxylase 1 alpha chain];TCP transcription factor (Transcription factor TCP2);DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;R domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-157	UUUCGUGCCGGAUUAUCAUUUC	noHit	6D	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-29;Protein kinase domain-containing protein;	XLOC_064557(1.5)	Glutaredoxin domain-containing protein;Glutaredoxin-C13	Cstv	miRNA	F_PSUSStaeMirSir
M-159	UUUGGAUUGAAGGGAGCUCUG	miR159	3A,3B,3D	MIR,mRNA,miRNA,microRNA;	XLOC_068598(0.0),XLOC_061039(0.0),XLOC_052503(0.0),XLOC_060535(0.5),XLOC_018519(2.0),XLOC_022024(2.0),XLOC_028123(2.0),XLOC_022594(2.0),XLOC_063612(2.0),XLOC_054670(2.0),XLOC_115896(2.0),XLOC_110530(2.0),XLOC_094745(2.0),XLOC_046151(2.0),XLOC_060579(2.0),XLOC_057263(2.0),XLOC_037194(2.0)	Transcription factor;Os11g0656500 protein (Retrotransposon protein, putative, unclassified, expressed);Transcription factor GAMYB (OsGAMyb);R2R3MYB-domain protein;Transcription factor GAMYB;Reverse transcriptase Ty1/copia-type domain-containing protein; Phosphatidylserine decarboxylase 1 alpha chain];Transcription factor (Fragment); (Fragment);DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;Integrase catalytic domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-160	UUUGUCCGGAAAUACUUGUUC	noHit	2A		XLOC_003878(0.0),XLOC_000434(0.5),XLOC_000150(0.5),XLOC_014360(0.5),XLOC_022540(0.5),XLOC_059744(0.5),XLOC_061672(0.5),XLOC_148652(0.5),XLOC_148853(0.5),XLOC_123357(0.5),XLOC_101164(0.5),XLOC_107179(0.5),XLOC_035667(0.5),XLOC_034540(0.5),XLOC_079009(0.5),XLOC_010206(1.0),XLOC_006202(1.0),XLOC_009888(1.0),XLOC_000421(1.0),XLOC_010357(1.0),XLOC_001780(1.0),XLOC_000221(1.0),XLOC_012297(1.0),XLOC_009791(1.0),XLOC_005819(1.0),XLOC_007289(1.0),XLOC_008920(1.0),XLOC_022915(1.0),XLOC_023172(1.0),XLOC_022845(1.0),XLOC_010870(1.0),XLOC_018546(1.0),XLOC_012750(1.0),XLOC_021927(1.0),XLOC_019346(1.0),XLOC_017888(1.0),XLOC_028964(1.0),XLOC_025790(1.0),XLOC_029399(1.0),XLOC_031644(1.0),XLOC_025789(1.0),XLOC_029539(1.0),XLOC_024773(1.0),XLOC_026314(1.0),XLOC_032398(1.0),XLOC_030570(1.0),XLOC_026964(1.0),XLOC_065628(1.0),XLOC_068597(1.0),XLOC_059258(1.0),XLOC_058972(1.0),XLOC_055785(1.0),XLOC_059038(1.0),XLOC_059857(1.0),XLOC_058759(1.0),XLOC_064423(1.0),XLOC_081020(1.0),XLOC_080606(1.0),XL	H_noveldroiase, alpha/beta fold family protein, expressed (Os03g0178400 protein) (cDNA clone:002-141-A10, full insert sequence);Non-specific serine/threonine protein kinase;Viviparous9;Protein CHROMATIN REMODELING 19;Cellulose synthase;2Fe-2S ferredoxin-type domain-containing protein;Os01g0248500 protein (Putative pathogenesis related protein) (cDNA clone:006-208-D11, full insert sequence);UPF0020 domain-containing protein;Zinc finger protein 1;BZIP domain-containing protein;Protein RETICULATA chloroplastic;RRM domain-containing protein;Os05g0534900 protein;Putative arylformamidase;TPR repeat-containing thioredoxin TDX;NAC domain-containing protein 74;Os04g0255600 protein (Fragment);Ethylene receptor 1;Importin N-terminal domain-containing protein;DEAD-box ATP-dependent RNA helicase 34;Pyr_redox_2 domain-containing protein;SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1;HSF_DOMAIN domain-containing protein;Phosphatidate cytidylyltransferase;DNA repair protein RAD51 homolog;Os08g0300501 protein;B2-adapt-app_C domain-containing protein;Myb_DNA-bind_3 domain-containing protein (Fragment);CRAL-TRIO domain-containing protein;I domain	CRup	miRNA	H_novel
M-161	GAGAGGACAGAGCGAGCAGGC	noHit	2A		XLOC_025967(2.0)		Cstv	miRNA	F_PSUSStaeMirSir
M-162	AGCUGCCGGUUCAUGGAUACC	noHit	3A,3B,3D,7A	MIR,microRNA;	XLOC_068836(0.0),XLOC_060535(0.0),XLOC_045718(2.0)	Phosphotransferase	CRdw	miRNA,siRNA	F_PSUSStaeMirSir
M-163	GAUAGGGAAACUGACAUGCCU	noHit	6A,6D		XLOC_129568(0.0),XLOC_129421(0.0),XLOC_024191(1.5),XLOC_042407(1.5),XLOC_032810(1.5),XLOC_020341(2.0),XLOC_052499(2.0),XLOC_046033(2.0),XLOC_075115(2.0)	1-phosphatidylinositol-3-phosphate 5-kinase;Putative disease resistance protein RGA3;NB-ARC domain-containing protein;Disease resistance protein RPM1;DUF4371 domain-containing protein	Cstv	miRNA,siRNA	F_PSUSStaeMirSir
M-164	CUAAAAUUUAGCUGUAGCGCC	noHit	4B		XLOC_021300(2.0)	Os04g0617050 protein (Fragment);Auxin-responsive protein SAUR71	CRdw	miRNA	F_PSUSStaeMirSir
M-165	AGGUGGAAUACUUGAAGAAGA	miR9670	6A,6D	Os02g0651500 protein (Fragment);	XLOC_128592(0.5),XLOC_062979(1.5),XLOC_082298(1.5),XLOC_056859(1.5),XLOC_046276(1.5),XLOC_056279(2.0),XLOC_062267(2.0),XLOC_088584(2.0),XLOC_135649(2.0),XLOC_143241(2.0),XLOC_135730(2.0),XLOC_128803(2.0),XLOC_127544(2.0),XLOC_110367(2.0),XLOC_109769(2.0),XLOC_090559(2.0),XLOC_048443(2.0)	WD_REPEATS_REGION domain-containing protein;DDT domain-containing protein;Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase;LRRNT_2 domain-containing protein;Coatomer subunit alpha-1;Os02g0651500 protein (Fragment);Os08g0541300 protein;Coatomer subunit alpha	CRup	miRNA	A_MiRBaseTae
M-166	CCAUUUGAAGACUAGUUUAUU	noHit	1A		XLOC_012297(1.0),XLOC_006001(1.5),XLOC_012379(1.5),XLOC_016786(1.5),XLOC_016787(1.5),XLOC_021552(1.5),XLOC_018567(1.5),XLOC_129421(1.5),XLOC_005950(2.0),XLOC_121586(2.0),XLOC_052771(2.0)	Protein kinase domain-containing protein;AAA domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	miRNA	H_novel

M-167	AUGGUGCGGAUGAGAUCGAGC	noHit	6D		XLOC_128777(0.0),XLOC_128038(0.0),XLOC_007779(1.5),XLOC_001186(1.5),XLOC_014806(1.5),XLOC_137538(1.5),XLOC_129114(1.5),XLOC_001119(2.0),XLOC_008396(2.0),XLOC_012834(2.0),XLOC_128697(2.0),XLOC_124624(2.0),XLOC_113783(2.0),XLOC_043950(2.0),XLOC_035103(2.0),XLOC_032580(2.0)	Lipase_3 domain-containing protein;TOM1-like protein 2;Lipase_GDSL domain-containing protein;DUF1618 domain-containing protein;Tetratricopeptide repeat protein 37;Os07g0642200 protein (cDNA clone:001-015-G02, full insert sequence);Os02g0690200 protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;Glycerol-3-phosphate acyltransferase, chloroplastic (GPAT)	Cstv	miRNA,siRNA	H_novel
M-168	GCAUCCAAGGUCGGCAGGGCC	noHit	6D		XLOC_116247(1.5),XLOC_064231(2.0),XLOC_059001(2.0),XLOC_046831(2.0)	Os01g0579800 protein (cDNA clone:002-181-C12, full insert sequence);OSJNBa0058K23.8-like protein	Cstv	miRNA	H_novel
M-169	UAAAAUUUAGCUGUAGCGCCU	noHit	4B		XLOC_115546(2.0)		CRdw	miRNA	F_PSUSStaeMirSir
M-170	GGAUUGCACCGAUGACUCGGA	noHit	3B	Os06g0219700 protein (cDNA clone:J023036O11, full insert sequence);	XLOC_055858(0.0),XLOC_049635(1.5),XLOC_046889(1.5)	F-box protein FBW2;Os06g0219700 protein (cDNA clone:J023036O11, full insert sequence)	CRdw	miRNA	H_novel
M-171	GGUGCUAUGGAUAAAUUCAAC	miR9674	3A,3B,3D,4D,6B,6D		XLOC_117006(0.0),XLOC_123290(0.0),XLOC_061178(1.0),XLOC_129417(1.0),XLOC_092129(1.0),XLOC_079645(1.0),XLOC_065748(1.5),XLOC_129409(2.0)		Cstv	miRNA	C_MiRBaseAll
M-172	UCGUGCCGGAUUAUCAUUUCA	noHit	6D	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-29;Protein kinase domain-containing protein;			Cstv	miRNA	H_novel
M-173	UGUGUUCUCAGGUCGCCCCCG	miR398	3A,3B,3D	Beta-Casp domain-containing protein	XLOC_064449(0.0),XLOC_065563(0.0),XLOC_057863(0.0),XLOC_046804(0.0)	Beta-Casp domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-174	AACCUUGCAAACUUUGAUCAA	noHit	2D		XLOC_087871(0.5),XLOC_061878(1.5),XLOC_053545(1.5),XLOC_047682(1.5),XLOC_129550(2.0),XLOC_122085(2.0),XLOC_122092(2.0)	Tetratricopeptide repeat (TPR)-like superfamily protein;rRNA N-glycosidase;Os01g0704200 protein;Protein kinase domain-containing protein;Os02g0615500 protein (cDNA clone:J013170D14, full insert sequence);LRRNT_2 domain-containing protein;Os02g0615500 protein (Fragment);F-box domain-containing protein;Os01g0652000 protein	Cstv	miRNA	H_novel
M-175	AAGAUAGGCUGUAGAUGACAUGGC	noHit	1B	mRNA	XLOC_065679(1.5),XLOC_022953(2.0),XLOC_012790(2.0),XLOC_025464(2.0),XLOC_036996(2.0)	AP2-like ethylene-responsive transcription factor ANT;AP2 transcription factor	Cstv	miRNA	F_PSUSStaeMirSir
M-176	AAGCGUAGUCGAACGAAUCUG	miR9663	6B		XLOC_148919(0.5),XLOC_126345(1.0),XLOC_128762(1.5),XLOC_126866(1.5),XLOC_096480(1.5)	Polyglutamine tract-binding protein 1 (Polyglutamine-binding protein 1)	CRdw	miRNA	A_MiRBaseTae
M-177	AAUUUAACCAACGAGACUGACUGU	noHit	1A,3A,3B,4A,6B,6D	,mRNA,stress;	XLOC_041207(0.5),XLOC_004262(1.0),XLOC_001153(1.0),XLOC_042456(1.0),XLOC_024508(1.5),XLOC_142952(1.5),XLOC_131938(1.5),XLOC_095086(1.5),XLOC_000210(2.0),XLOC_020728(2.0),XLOC_068422(2.0),XLOC_056293(2.0),XLOC_093380(2.0),XLOC_123341(2.0)	Acetyl-CoA carboxylase;Ankyrin repeat protein SKIP35;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;Methyltranfer_dom domain-containing protein;BZIP domain-containing protein;Glutathione reductase (GRase);Retrovirus-related Pol polyprotein LINE-1;ABSCISIC ACID-INSENSITIVE 5-like protein 6	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-178	ACAGACGGCAAAAGAGCUCUUUGC	noHit	3B				Cstv	miRNA	H_novel
M-179	ACAUGAUUAUCGAAGUUUCAGACU	noHit	1B		XLOC_063827(2.0)		Cstv	miRNA	H_novel
M-180	ACCACGACUGUCAUUAAGCAU	miR9672	5A,5B,5D	mRNA;F-box only protein 7;	XLOC_107815(0.0)	F-box only protein 7	Cstv	miRNA	A_MiRBaseTae
M-181	ACCUGCAGUUGGGCCAAUGAC	miR1847	2A,3A,3B,4A,4B,5A,5B,7A				Cstv	miRNA	A_MiRBaseTae
M-182	ACCUUCAGGAAGGACUGCAUC	noHit	7D	mRNAPutative disease resistance RPP8-like protein 2;Rx_N domain-containing protein;	XLOC_007465(0.0),XLOC_147261(0.0),XLOC_007467(1.5),XLOC_006323(1.5),XLOC_054625(1.5),XLOC_063003(1.5),XLOC_088900(1.5),XLOC_046673(1.5),XLOC_012010(2.0),XLOC_079992(2.0),XLOC_150458(2.0),XLOC_123964(2.0),XLOC_099318(2.0),XLOC_051646(2.0),XLOC_071922(2.0)	Putative disease resistance RPP8-like protein 2;Ubiquitin carboxyl-terminal hydrolase 13;Putative disease resistance protein RXW24L;Disease resistance protein RPP13;Protein VACUOLELESS1;Os09g0517200 protein;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;D-aminoacyl-tRNA deacylase;Disease resistance RPP13-like protein 4;C2 NT-type domain-containing protein	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-183	ACGGCAUAGAGGCACUGCAAA	noHit	1A,1B,1D		XLOC_055832(2.0)	Os01g0621900 protein (cDNA clone:J013038F15, full insert sequence)	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-184	ACGGCAUAGAGGCACUGCAAAC	noHit	1A,1B,1D		XLOC_055832(2.0)	Os01g0621900 protein (cDNA clone:J013038F15, full insert sequence)	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-185	ACUGGAUGACGCGGGAGCUAA	miR319	3A,3B,3D		XLOC_061121(0.0),XLOC_049362(0.0),XLOC_026148(1.5)	NAC domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot
M-186	ACUUUCGUGCCGGAUUAUCAUUUC	noHit	6D	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-29;Protein kinase domain-containing protein;	XLOC_078126(2.0)		Cstv	miRNA	F_PSUSStaeMirSir
M-187	AGAAAGCAAUGUCAGAGGAGU	noHit	7A,7B		XLOC_137230(1.5),XLOC_144876(1.5),XLOC_151052(1.5),XLOC_132776(1.5)	FBD domain-containing protein;DBD_Tnp_Mut domain-containing protein;IPPc domain-containing protein;Threonyl-tRNA synthetase	Cstv	miRNA	F_PSUSStaeMirSir
M-188	AGAAGAUUAGAAGAUUAAGCA	noHit	3B	mRNA;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);(Fragment);Disease resistance protein RPP13;CNL-C (Disease resistance RPP13-like protein 4);NB-ARC domain-containing protein	XLOC_063979(0.0),XLOC_052866(0.0),XLOC_042744(1.5),XLOC_062178(2.0),XLOC_062352(2.0),XLOC_063832(2.0),XLOC_080993(2.0),XLOC_100286(2.0),XLOC_045848(2.0),XLOC_069284(2.0)	Os01g0971900 protein;CNL-C (Disease resistance RPP13-like protein 4);Disease resistance protein RPP13;PUM-HD domain-containing protein;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);(Fragment);Zinc finger CCCH domain-containing protein 24 (OsC3H24);NB-ARC domain-containing protein;Zinc finger CCCH domain-containing protein 24;START domain-containing protein	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-189	AGAGAGGACAGAGCGAGCAGGC	noHit	2A				CRupCRdw	miRNA	F_PSUSStaeMirSir

M-190	AGAGAUCAAUGUGAAGAACAC	noHit	5D		XLOC_061249(2.0),XLOC_086856(2.0)	DUF295 domain-containing protein;Os01g0327900 protein;F-box domain-containing protein	Cstv	miRNA	H_novel
M-191	AGAGGACAGAGCGAGCAGGCAC	noHit	2A				Cstv	miRNA	H_novel
M-192	AGAGUUCUGCAUCAUCGAAGGGCA	noHit	2A,7A		XLOC_002723(1.0),XLOC_010218(1.5),XLOC_005682(2.0)	Protein kinase domain-containing protein;Reverse transcriptase domain-containing protein;Leucine Rich Repeat family protein (Os11g0628000 protein)	Cstv	miRNA	H_novel
M-193	AGAUAGAGAGCCCUGGCGAAAG	noHit	3B				Cstv	miRNA	H_novel
M-194	AGCUGAAUCGGUCCUCAGACC	noHit	7D	DBD_Tnp_Mut domain-containing protein;	XLOC_132759(0.0),XLOC_132776(2.0)	FBD domain-containing protein;USP domain-containing protein;IPPc domain-containing protein;Threonyl-tRNA synthetase	CRdw	miRNA	H_novel
M-195	AGGGUCGAACUGAGAACACAU	miR398	3D	Beta-Casp domain-containing protein	XLOC_064449(0.0),XLOC_106601(1.5),XLOC_099618(1.5),XLOC_071009(1.5),XLOC_027447(2.0),XLOC_038067(2.0)	CLIP-associating protein 2;Beta-Casp domain-containing protein;E3 ubiquitin-protein ligase listerin	Cstv	miRNA	B_MiRBaseMonocot
M-196	AGGUCAUGCUGGAGUUUCAUC	miR167	5A,5B,5D		XLOC_096564(1.5),XLOC_125513(2.0),XLOC_119450(2.0),XLOC_113564(2.0),XLOC_094159(2.0)	Cytochrome P450 97B2, chloroplastic	Cstv	miRNA	B_MiRBaseMonocot
M-197	AGUUCCAGACGAUGCAGGCCU	noHit	7B,7D	;Disease resistance protein RPM1;Rx_N domain-containing protein;Putative disease resistance RPP8-like protein 2			CRup	miRNA	F_PSUSStaeMirSir
M-198	AUACCACGACUGUCAUUAAGCA	miR9672	5B	mRNA	XLOC_107815(0.5)	F-box only protein 7	Cstv	miRNA	A_MiRBaseTae
M-199	AUAGCAUCAUCCAUCCUACCA	miR9674	4D,5A		XLOC_123488(1.0),XLOC_117057(1.0),XLOC_007609(1.5),XLOC_017237(1.5),XLOC_150040(1.5),XLOC_067022(2.0),XLOC_129417(2.0),XLOC_116887(2.0),XLOC_108117(2.0),XLOC_092129(2.0),XLOC_054186(2.0),XLOC_050782(2.0),XLOC_052955(2.0),XLOC_079645(2.0)	Putative bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1;NB-ARC domain-containing protein;PPR_long domain-containing protein;ANK_REP_REGION domain-containing protein;F-box domain-containing protein;Os04g0514500 protein (cDNA clone:J033104I01, full insert sequence);Protein Rf1, mitochondrial	CRdw	miRNA	A_MiRBaseTae
M-200	AUAGCAUCAUCCAUCCUGCCA	noHit	4B,4D		XLOC_123488(0.0),XLOC_117057(0.0),XLOC_128876(1.5),XLOC_128855(1.5),XLOC_117163(1.5),XLOC_116908(1.5),XLOC_007609(2.0),XLOC_017237(2.0),XLOC_150040(2.0),XLOC_127777(2.0),XLOC_123443(2.0),XLOC_128116(2.0),XLOC_128947(2.0),XLOC_127644(2.0),XLOC_128858(2.0),XLOC_117086(2.0),XLOC_115860(2.0),XLOC_097076(2.0),XLOC_093889(2.0),XLOC_054186(2.0)	ANK_REP_REGION domain-containing protein;PPR_long domain-containing protein;Protein Rf1, mitochondrial	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-201	AUAUCAAGCACUCAUCGACAG	noHit	6B				Cstv	miRNA	F_PSUSStaeMirSir
M-202	AUAUUAUGGUGCGGAGGGAGU	noHit	1D,6D		XLOC_000238(0.5),XLOC_031108(1.0),XLOC_055800(1.0),XLOC_137671(1.0),XLOC_111798(1.0),XLOC_119928(1.0),XLOC_044310(1.0),XLOC_007103(1.5),XLOC_021879(1.5),XLOC_033977(1.5),XLOC_063632(1.5),XLOC_141009(1.5),XLOC_123876(1.5),XLOC_130055(1.5),XLOC_132155(1.5),XLOC_121403(1.5),XLOC_112071(1.5),XLOC_102007(1.5),XLOC_093851(1.5),XLOC_043606(1.5),XLOC_033351(1.5),XLOC_000930(2.0),XLOC_006435(2.0),XLOC_002102(2.0),XLOC_002186(2.0),XLOC_002097(2.0),XLOC_009379(2.0),XLOC_012781(2.0),XLOC_019102(2.0),XLOC_017282(2.0),XLOC_013212(2.0),XLOC_014238(2.0),XLOC_021601(2.0),XLOC_012050(2.0),XLOC_012046(2.0),XLOC_011094(2.0),XLOC_013030(2.0),XLOC_017609(2.0),XLOC_019254(2.0),XLOC_020816(2.0),XLOC_014103(2.0),XLOC_013691(2.0),XLOC_015768(2.0),XLOC_016690(2.0),XLOC_016124(2.0),XLOC_013538(2.0),XLOC_029560(2.0),XLOC_029497(2.0),XLOC_020353(2.0),XLOC_020626(2.0),XLOC_030603(2.0),XLOC_029607(2.0),XLOC_024254(2.0),XLOC_023809(2.0),XLOC_022612(2.0),XLOC_023791(2.0),XLOC_031632(2.0),XLOC_025731(2.0),XL	Non-specific serine/threonine protein kinase;Rhodanese domain-containing protein;ABC1 domain-containing protein;M20_dimer domain-containing protein;3-hydroxyacyl-CoA dehydrogenase;Cyclopropane-fatty-acyl-phospholipid synthase;RRM domain-containing protein;Reverse transcriptase domain-containing protein;3-hydroxyisobutyryl-CoA hydrolase-like protein 2, mitochondrial;Cytochrome P450 97B2, chloroplastic;DUF4210 domain-containing protein;Phosphopantothenoylcysteine decarboxylase (Putative phosphopantothenoylcysteine decarboxylase);TPX2 domain-containing protein;tRNA pseudouridine synthase (tRNA pseudouridine synthase 3);Ras-related protein Rab-18;Os09g0386500 protein (cDNA clone:001-125-G12, full insert sequence) (cDNA clone:J023141K03, full insert sequence);Os01g0704200 protein;Putative G protein-coupled receptor;Os01g0679600 protein;Os10g0475900 protein (Ubiquitin family protein, expressed) (cDNA clone:J033069L14, full insert sequence);RNA-binding protein 40;Amino-acid N-acetyltransferase;Os09g0489200 protein;FYVE-type domain-containing protein;Os08g0402600 protein (Fragment);Os01g0355900 protein (cDNA	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-203	AUCAGGAGAGAUGACACCGACGCC	noHit	2B,2D		XLOC_102168(0.0),XLOC_000994(0.5),XLOC_003288(0.5),XLOC_008047(0.5),XLOC_015506(0.5),XLOC_058856(0.5),XLOC_014564(2.0),XLOC_073716(2.0)	Os03g0812800 protein;EF hand family protein, expressed (Os03g0812400 protein) (cDNA clone:002-184-C05, full insert sequence);Tryptophan aminotransferase 1;DNA-directed DNA polymerase;Calmodulin-like protein 5;Alliinase_C domain-containing protein	CRdw	miRNA	H_novel
M-204	AUCAUGCGAUCCUUUUGGAAG	noHit	3A,3D	MIR,microRNA;	XLOC_098953(2.0)	H_novelexosyltransferase	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-205	AUCCAGAUCGAGUGCGGCGCUGUC	noHit	1A		XLOC_069968(2.0)	Os05g0119600 protein	Cstv	miRNA	H_novel
M-206	AUCCCUCAACUCCGCCGCCUC	noHit	5B	DNA-directed RNA polymerase subunit beta	XLOC_000260(2.0),XLOC_006531(2.0),XLOC_013514(2.0)	Protein-serine/threonine phosphatase;Probable protein phosphatase 2C 72 (OsPP2C72)	CRdw	miRNA	H_novel
M-207	AUCGUCCAACAGUAUCGUCUG	noHit	2D				CRdw	miRNA	F_PSUSStaeMirSir

M-208	AUGCCGUGUUGUUCUGAAAGAA	noHit	4A,4D		XLOC_109292(1.5),XLOC_017822(2.0),XLOC_024545(2.0),XLOC_085565(2.0),XLOC_086968(2.0),XLOC_135331(2.0),XLOC_147065(2.0),XLOC_149501(2.0),XLOC_034963(2.0),XLOC_076066(2.0)	Non-specific serine/threonine protein kinase;Protein kinase domain-containing protein;Os07g0639600 protein;Receptor-like serine/threonine-protein kinase;E3 ubiquitin ligase (RING-type E3 ubiquitin transferase) (Serine/threonine-protein kinase);Os09g0396300 protein;Separase;NPH3 domain-containing protein;C2H2-type domain-containing protein;Aldehyde oxygenase (deformylating)	Cstv	miRNA	E_LiteratureTaeMiRNA 2020
M-209	AUGCUAUGGGUAAACUCGACC	noHit	5A		XLOC_079645(2.0)		Cstv	miRNA	E_LiteratureTaeMiRNA 2020
M-210	AUGGACCGCGGACAACGAGGAU	noHit	1B		XLOC_119379(1.5)	AP2-like ethylene-responsive transcription factor AIL5	Cstv	miRNA	F_PSUSStaeMirSir
M-211	AUGGCGUGAAAGGACGGAAGUACC	noHit	2D				CRdw	miRNA	F_PSUSStaeMirSir
M-212	AUGGUCAAACUUAGAUCUCGAGAA	noHit	2A,2D,7A	,mRNA;	XLOC_021273(0.0),XLOC_102847(0.0),XLOC_109893(0.0),XLOC_095086(0.0),XLOC_041207(0.0),XLOC_042456(0.5),XLOC_020000(2.0),XLOC_124931(2.0),XLOC_118938(2.0),XLOC_116567(2.0),XLOC_112140(2.0),XLOC_099846(2.0)	Acetyl-CoA carboxylase;Nucleic acid binding protein;Flowering locus K homology domain;Ankyrin repeat protein SKIP35;Transmembrane protein;Os07g0143600 protein (Fragment);Cysteine-rich receptor-like protein kinase 26;Methyltranfer_dom domain-containing protein;GPI inositol-deacylase;Os02g0666800 protein (cDNA clone:J033039F09, full insert sequence);S-acyltransferase;Purine-uracil permease NCS1;Glutathione reductase (GRase);Retrovirus-related Pol polyprotein LINE-1;Os02g0666700 protein	Cstv	miRNA	F_PSUSStaeMirSir
M-213	AUUCGGUCAAACUUUGAGAUAGUU	noHit	5B,6D,7A	,mRNA;	XLOC_112629(0.5),XLOC_026347(1.5),XLOC_034922(1.5),XLOC_098803(2.0),XLOC_090593(2.0)	Serine/threonine-protein phosphatase;F-box/LRR-repeat protein 14;BED-type domain-containing protein;Naringenin,2-oxoglutarate 3-dioxygenase;Serine/threonine-protein phosphatase BSL3;Fe2OG dioxygenase domain-containing protein;NB-ARC domain-containing protein;F-box domain-containing protein	Cstv	miRNA	H_novel
M-214	AUUUGGGCCUAUUUUCUUGCAAUG	noHit	7A		XLOC_003688(2.0),XLOC_007928(2.0),XLOC_134163(2.0)	Os06g0231100 protein;DYW_deaminase domain-containing protein;Uncharacterized mitochondrial protein AtMg00030 (ORF107a);Nuclear cap-binding protein subunit 1	CRup	miRNA	H_novel
M-215	AUUUGUAGUGUUCGGAUUGAGUUU	miR1878	5A,5B,5D				Cstv	miRNA	B_MiRBaseMonocot
M-216	CAGAACCAGAAUGAGUAGCUC	miR9679	2D	Fe2OG dioxygenase domain-containing protein;Hexosyltransferase;	XLOC_091988(1.5),XLOC_057758(2.0),XLOC_039379(2.0)	Elicitor-responsive protein 1;Fe2OG dioxygenase domain-containing protein;Hexosyltransferase;C2 domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-217	CAGAUCGAGUGCGGCGCUGUCGUC	noHit	1A				Cstv	miRNA	H_novel
M-218	CAGCCAAGGAUGACUUGCCGGC	miR169	2A,2B,2D		XLOC_027763(0.0),XLOC_143891(1.0),XLOC_151645(1.0),XLOC_109128(1.0),XLOC_029946(2.0),XLOC_024581(2.0),XLOC_032331(2.0),XLOC_085898(2.0),XLOC_082116(2.0),XLOC_126522(2.0),XLOC_113661(2.0),XLOC_117670(2.0),XLOC_106316(2.0),XLOC_096308(2.0),XLOC_041602(2.0),XLOC_042710(2.0),XLOC_074766(2.0),XLOC_077948(2.0)	Nuclear transcription factor Y subunit A-5;CCAAT-binding transcription factor B (CCAAT-binding transcription factor family protein);CCAAT-binding transcription factor B;CCAAT-binding transcription factor B (CCAAT-box transcription factor complex WHAP4);Nuclear transcription factor Y subunit A-9;CCAAT-box transcription factor complex WHAP6	CRdw	miRNA	B_MiRBaseMonocot
M-219	CAUCGGCGAGUUGUUCUUGGC	noHit	3D	miRNA,microRNA,MIR	XLOC_000472(1.5),XLOC_007832(1.5),XLOC_015391(1.5),XLOC_002496(2.0),XLOC_001161(2.0),XLOC_017371(2.0),XLOC_027423(2.0),XLOC_060950(2.0),XLOC_142521(2.0),XLOC_145188(2.0),XLOC_114839(2.0),XLOC_100971(2.0),XLOC_034090(2.0),XLOC_079650(2.0)	DDE Tnp4 domain-containing protein;Os09g0122200 protein (Fragment);DUF4005 domain-containing protein;Os10g0468250 protein;Os03g0643100 protein (Fragment);Os10g0468301 protein (Fragment);DUF1084 domain-containing protein;Os07g0690900 protein;Integrase catalytic domain-containing protein;Protein IQ-DOMAIN 14;X8 domain-containing protein;Os03g0643050 protein (Fragment);QWRF motif-containing protein 3;Os10g0460733 protein;Myb_DNA-bind_3 domain-containing protein;Os07g0115800 protein (Fragment);Brittle stalk-2-like protein 8 (COBRA-like protein 7);Tobamovirus multiplication protein 3;Os09g0122100 protein;WAC domain-containing protein;Aluminum-activated malate transporter 12;Os10g0572100 protein;X8 domain-containing protein (Fragment);Os08g0115200 protein;Reverse transcriptase Ty1/copia-type domain-containing protein;Os07g0116050 protein;Carbohydrate-binding X8 domain superfamily protein	Cstv	miRNA	E_LiteratureTaeMiRNA 2020
M-220	CAUGUGCCUUUCUUCUCCACC	miR164	1A,1B,1D		XLOC_034195(1.5),XLOC_032403(2.0),XLOC_025536(2.0),XLOC_060953(2.0),XLOC_136157(2.0),XLOC_048106(2.0),XLOC_033405(2.0)	F-box/kelch-repeat protein SKIP11;Glutaredoxin domain-containing protein;Pentatricopeptide repeat-containing protein CRP1, chloroplastic (Protein CHLOROPLAST RNA PROCESSING 1);Os07g0639600 protein;PPR_long domain-containing protein;C2H2-type domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot

M-221	CCCGUCCUUUUUACUCUGCAUAUU	noHit	4B	mRNA,tRNA	XLOC_023795(1.5),XLOC_140663(1.5),XLOC_145831(1.5),XLOC_112320(1.5),XLOC_116152(1.5),XLOC_122085(1.5),XLOC_090067(1.5),XLOC_094017(1.5),XLOC_048632(1.5),XLOC_074476(1.5),XLOC_077454(1.5),XLOC_070618(1.5),XLOC_046701(2.0),XLOC_046407(2.0)	Eukaryotic translation initiation factor 2 alpha subunit family protein, expressed (Os03g0296400 protein) (cDNA clone:006-208-D01, full insert sequence) (cDNA clone:J033040F15, full insert sequence);DnaJ subfamily C member 7;Protein kinase domain-containing protein;Calmodulin-binding transcription activator 6;ENT domain-containing protein;Eukaryotic translation initiation factor 2 subunit alpha;Magnesium transporter;Meristematic receptor-like kinase;Rx_N domain-containing protein;F-box domain-containing protein;Os02g0704600 protein (cDNA, clone: J065163N21, full insert sequence);DnaJ homolog subfamily C member 3;S1 motif domain-containing protein	Cstv	miRNA	F_PSUSStaeMirSir
M-222	CCUCAGACCAGUAGUUUCCAACUC	noHit	7D	DBD_Tnp_Mut domain-containing protein;	XLOC_132759(0.0),XLOC_137230(2.0)	USP domain-containing protein	Cstv	miRNA	F_PSUSStaeMirSir
M-223	CCUGUUUGUCAUUAAGUUUCUU	miR9652	2A,2B,2D		XLOC_058622(1.5),XLOC_105554(1.5),XLOC_107844(1.5),XLOC_096840(1.5),XLOC_096836(1.5),XLOC_093085(1.5),XLOC_092193(1.5),XLOC_055354(2.0),XLOC_036863(2.0)	Protein argonaute MEL1	Cstv	miRNA	A_MiRBaseTae
M-224	CGCGGACAACGAGGAUGAAGACGA	noHit	1B				Cstv	miRNA	F_PSUSStaeMirSir
M-225	CGGACAACGAGGAUGAAGACGAU	noHit	1B		XLOC_056102(2.0)	Os09g0419500 protein (cDNA clone:J033067G01, full insert sequence);RING/U-box superfamily protein;Os09g0419500 protein;RBR-type E3 ubiquitin transferase	Cstv	miRNA	H_novel
M-226	CGGAUAGGCAUGGUGCUGGACGCU	noHit	3B		XLOC_060894(1.0),XLOC_120991(2.0)	Putative protein phosphatase 2C 23 (OsPP2C23);Protein phosphatase	Cstv	miRNA	H_novel
M-227	CGUUGAGAUGAGAUUACCCCA	noHit	4A,5A,5B,5D		XLOC_102302(0.0),XLOC_101813(0.0),XLOC_121718(1.5),XLOC_120961(1.5),XLOC_115019(1.5),XLOC_003091(2.0),XLOC_122610(2.0),XLOC_115971(2.0)	F-box domain-containing protein;	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-228	CUCAGGAGAGAUGACACCGAC	miR1432	2A		XLOC_115353(0.0),XLOC_102168(1.0),XLOC_003288(1.5),XLOC_008047(1.5),XLOC_000994(1.5),XLOC_015506(1.5),XLOC_058856(1.5),XLOC_134742(1.5),XLOC_119226(2.0),XLOC_103639(2.0)	VWFA domain-containing protein;Os03g0812800 protein;Calcium-transporting ATPase;Inter-alpha-trypsin inhibitor heavy chain-related;EF hand family protein, expressed (Os03g0812400 protein) (cDNA clone:002-184-C05, full insert sequence);Tryptophan aminotransferase 1;Calmodulin-like protein 5;Alliinase_C domain-containing protein;Fucosyltransferase	Cstv	miRNA	B_MiRBaseMonocot
M-229	CUCCACAGGCUUUCUUGAACU	miR396	2A,2B,2D,6A,6D		XLOC_114027(0.0),XLOC_022666(1.0),XLOC_030908(1.0),XLOC_087793(1.0),XLOC_080318(1.0),XLOC_133553(1.0),XLOC_148284(1.0),XLOC_124078(1.0),XLOC_126012(1.0),XLOC_112641(1.0),XLOC_112330(1.0),XLOC_119737(1.0),XLOC_036899(1.0),XLOC_069277(1.0),XLOC_070618(1.0),XLOC_027870(2.0),XLOC_022258(2.0),XLOC_110439(2.0)	Growth regulating factor 4-D1;Growth regulating factor 5-A1;Growth-regulating factor 4 protein (Fragment);Growth regulating factor 1-D1;Growth regulating factor 4a-B1;Importin N-terminal domain-containing protein;Growth regulating factor 9-A1;Growth regulating factor 1-B1;Importin-beta1 (Os05g0353400 protein) (cDNA clone:J013029L11, full insert sequence);Growth regulating factor 1-A1;Importin subunit beta-1;QLQ domain-containing protein;Growth regulating factor 5-D1;GRF transcription factor (Growth-regulating factor);Growth-regulating factor 4 (OsGRF4) (Protein GRAIN LENGTH AND WIDTH 2) (Protein GRAIN SIZE ON CHROMOSOME 2) (Protein PANICLE TRAITS 2) (Transcription activator GRF4);WRC domain-containing protein;Growth regulating factor 5-4A;Os04g0600900 protein;Growth-regulating factor 1 (OsGRF1) (Transcription activator GRF1);Growth regulating factor 4-A1	Cstv	miRNA	B_MiRBaseMonocot
M-230	CUCGGACCAGGCUUCAAUCCC	miR166	5A,5B,5D		XLOC_109041(0.0),XLOC_003329(1.0),XLOC_010176(1.0),XLOC_140819(1.0),XLOC_147220(1.0),XLOC_132764(1.0)		Cstv	miRNA	B_MiRBaseMonocot
M-231	CUCGGACCAGGCUUCAUUCCCC	miR166	1A,1B,1D,7A,7B,7D		XLOC_003329(0.0),XLOC_010176(0.0),XLOC_140819(0.0),XLOC_147220(0.0),XLOC_132764(0.0),XLOC_109041(1.0),XLOC_110868(1.5),XLOC_103056(1.5),XLOC_095467(1.5),XLOC_001247(2.0),XLOC_008070(2.0),XLOC_015254(2.0),XLOC_059246(2.0),XLOC_084967(2.0),XLOC_105100(2.0),XLOC_098448(2.0),XLOC_092135(2.0),XLOC_090000(2.0),XLOC_051179(2.0),XLOC_076905(2.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot

M-232	CUGACAGAAGAGAGUGAGCAC	miR156	3A,3B,3D,5A,5B,5D		XLOC_006637(0.5),XLOC_135366(1.0),XLOC_142136(1.0),XLOC_148120(1.0),XLOC_128905(1.0),XLOC_121975(1.0),XLOC_115110(1.0),XLOC_060958(1.5),XLOC_079815(1.5),XLOC_031786(2.0),XLOC_084330(2.0),XLOC_127342(2.0),XLOC_122219(2.0),XLOC_116249(2.0),XLOC_119631(2.0),XLOC_101060(2.0),XLOC_104831(2.0),XLOC_100923(2.0),XLOC_032645(2.0),XLOC_041706(2.0)	Squamosa promoter-binding-like protein 16;Pumilio-like protein 1;MND1-interacting protein 1;STAS domain-containing protein;Nucleolin;P-type H(+)-exporting transporter;L-ascorbate oxidase-like protein;Squamosa promoter-binding-like protein 3;Os09g0491532 protein;Plasma membrane ATPase;SBP-type domain-containing protein;Hydrolase_4 domain-containing protein;GLW7;Squamosa promoter-binding-like protein 2;NAB domain-containing protein;Squamosa promoter-binding-like protein 13;GW8-B1;SBP transcription factor (Squamosa promoter-binding-like protein 2);SBP protein	CRdw	miRNA	A_MiRBaseTae
M-233	CUGGCUGACCUUGCUAUUGGCAUG	noHit	1A		XLOC_016595(1.0),XLOC_081804(1.0),XLOC_006831(1.5),XLOC_021064(2.0),XLOC_060047(2.0)	WD repeat-containing protein 44;WD_REPEATS_REGION domain-containing protein	Cstv	miRNA	H_novel
M-234	CUUAAUGACAGUCGUGGUGUC	miR9672	5A	mRNA	XLOC_107815(0.5)	F-box only protein 7	Cstv	miRNA	C_MiRBaseAll
M-235	CUUGGACUGAAGGGAGCUCCCU	miR319	4A,4B,4D	MIR,microRNA;	XLOC_060579(1.0),XLOC_022594(1.5),XLOC_063612(1.5),XLOC_054670(1.5),XLOC_110530(1.5),XLOC_094745(1.5),XLOC_060535(1.5),XLOC_046151(1.5),XLOC_005608(2.0),XLOC_061121(2.0),XLOC_088604(2.0),XLOC_125777(2.0),XLOC_049362(2.0)	Transcription factor;Transcription factor GAMYB (OsGAMyb);Transcription factor GAMYB;Putative membrane lipoprotein; Phosphatidylserine decarboxylase 1 alpha chain];Transcription factor (Fragment);Phenylalanine ammonia-lyase;DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;Os11g0703600 protein (Fragment);Expressed protein (Os11g0703600 protein) (cDNA clone:001-008-G09, full insert sequence)	Cstv	miRNA	A_MiRBaseTae
M-236	CUUGGACUGAAGGGUGCUCCC	miR319	3A,3B,3D	MIR,mRNA,microRNA;Os01g0222001 protein (Fragment);	XLOC_060579(0.0),XLOC_061121(1.0),XLOC_049362(1.0),XLOC_001904(2.0),XLOC_006959(2.0),XLOC_012831(2.0),XLOC_145942(2.0)	Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase;Phospholipase D;Os01g0222001 protein (Fragment)	Cstv	miRNA	A_MiRBaseTae
M-237	CUUGGACUGAAGGGUGCUCCCUC	miR319	3A,3B,3D	MIR,mRNA,microRNA;Os01g0222001 protein (Fragment);	XLOC_060579(0.0),XLOC_061121(1.0),XLOC_049362(1.0),XLOC_001904(2.0),XLOC_006959(2.0),XLOC_012831(2.0),XLOC_145942(2.0)	Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase;Phospholipase D;Os01g0222001 protein (Fragment)	Cstv	miRNA	A_MiRBaseTae
M-238	CUUGGAUUGAAGGGAGCUCUC	miR159	3B,3D,7A		XLOC_060535(0.0),XLOC_068598(1.0),XLOC_061039(1.0),XLOC_052503(1.0),XLOC_018519(1.5),XLOC_063612(1.5),XLOC_054670(1.5),XLOC_115896(1.5),XLOC_060579(1.5),XLOC_046151(1.5),XLOC_022594(2.0),XLOC_110530(2.0),XLOC_094745(2.0),XLOC_095968(2.0)	Transcription factor;Os11g0656500 protein (Retrotransposon protein, putative, unclassified, expressed);Transcription factor GAMYB (OsGAMyb);Transcription factor GAMYB;Reverse transcriptase Ty1/copia-type domain-containing protein; Phosphatidylserine decarboxylase 1 alpha chain];Transcription factor (Fragment);DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;Integrase catalytic domain-containing protein	Cstv	miRNA	A_MiRBaseTae
M-239	CUUGGGACAAAGCUUGAACUU	noHit	1A,1D		XLOC_108112(1.5),XLOC_096349(1.5),XLOC_090523(1.5),XLOC_043758(1.5),XLOC_012011(2.0),XLOC_031672(2.0),XLOC_026145(2.0),XLOC_020211(2.0),XLOC_085018(2.0),XLOC_049345(2.0),XLOC_038230(2.0)	Thiolase_N domain-containing protein;CBL-interacting protein kinase 7 (Non-specific serine/threonine protein kinase);GATA transcription factor;Non-specific serine/threonine protein kinase;Putative acetyl-CoA acetyltransferase, cytosolic 2;Os04g0539500 protein;GATA zinc finger family protein;RNA helicase;DExH-box ATP-dependent RNA helicase DExH16 mitochondrial;Protein RETICULATA chloroplastic;C2C2-GATA transcription factor (GATA zinc finger family protein) (Fragment);GATA-type domain-containing protein;ATP-dependent RNA helicase SUV3, mitochondrial (OsSUV3);Retrovirus-related Pol polyprotein from transposon TNT 1-94	CRupCRdw	miRNA	H_novel
M-240	CUUGUUUGUCAUUAAGCUUCUG	noHit	1D,2B		XLOC_017179(2.0),XLOC_055354(2.0),XLOC_085766(2.0),XLOC_135146(2.0),XLOC_108585(2.0),XLOC_093056(2.0),XLOC_095002(2.0),XLOC_089896(2.0),XLOC_036863(2.0),XLOC_073062(2.0)	Fatty acyl-CoA reductase;Carboxypeptidase SOL1	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-241	GAAAGCAAUGUCAGAGGAGUU	noHit	7A,7B		XLOC_137230(1.5),XLOC_144876(1.5),XLOC_151052(1.5),XLOC_132776(1.5),XLOC_019705(2.0),XLOC_030271(2.0),XLOC_113453(2.0)	FBD domain-containing protein;DBD_Tnp_Mut domain-containing protein;IPPc domain-containing protein;Threonyl-tRNA synthetase;ABC transporter C family member 9;Nucleoporin GLE1	Cstv	miRNA	F_PSUSStaeMirSir
M-242	GAAAUUAGCGUGGGAUAAAACUCU	noHit	3D		XLOC_065007(2.0),XLOC_057209(2.0),XLOC_041067(2.0)	Protein-serine/threonine phosphatase	Cstv	miRNA	H_novel
M-243	GAGAAAGCAAUGUCAGAGGAG	noHit	7A,7B		XLOC_137230(1.5),XLOC_144876(1.5),XLOC_132776(1.5),XLOC_005482(2.0),XLOC_010507(2.0),XLOC_018318(2.0),XLOC_130057(2.0)	FBD domain-containing protein;Os06g0530300 protein (Fragment);DBD_Tnp_Mut domain-containing protein;IPPc domain-containing protein;Threonyl-tRNA synthetase;Os05g0163400 protein (cDNA clone:J033030D14, full insert sequence);RING-type domain-containing protein	Cstv	miRNA	E_LiteratureTaeMiRNA2020
M-244	GAGAAAUUGGAGCUCGGUGCAGGC	noHit	3A,3D		XLOC_003012(2.0)	Os05g0427300 protein (cDNA clone:001-038-E05, full insert sequence) (cDNA clone:001-041-A02, full insert sequence) (cDNA clone:006-309-A11, full insert sequence) (cDNA clone:J013095G08, full insert sequence);RNA recognition motif-containing protein;RRM domain-containing protein;Splicing factor 3B subunit 4	CRdw	miRNA	H_novel
M-245	GAGAGGACAGAGCGAGCAGGCA	noHit	2A		XLOC_025967(2.0)		Cstv	miRNA	F_PSUSStaeMirSir
M-246	GAGUAGCGUUGAUACACCGUCGA	noHit	4B,5A	mRNA;			Cstv	miRNA	H_novel

M-247	GCGGACAACGAGGAUGAAGACGAU	noHit	1B		XLOC_014480(2.0),XLOC_056102(2.0)	Os09g0419500 protein (cDNA clone:J033067G01, full insert sequence);RING/U-box superfamily protein;Os09g0419500 protein;RBR-type E3 ubiquitin transferase	Cstv	miRNA	F_PSUSStaeMirSir
M-248	GCGGAUUUUUCACCAAGAUUCAAG	miR5062	5A,5D	H_noveleat shock cognate 70 kDa protein 2, putative, expressed (Os03g0821100 protein) (cDNA clone:J023030D03, full insert sequence);Heat shock 70 kDa protein 3;Heat shock cognate 70 kDa protein 1	XLOC_106053(0.0),XLOC_107103(1.5),XLOC_101542(1.5)	Carboxypeptidase	CRup	miRNA	C_MiRBaseAll
M-249	GCGUGCAAGGAGCCAAGCAUG	miR160	7A,7B,7D	MIR,mRNA,miRNA,microRNA;	XLOC_140630(0.0),XLOC_133015(0.0),XLOC_108631(1.5),XLOC_099427(2.0),XLOC_052955(2.0)	Reverse transcriptase domain-containing protein;Pectin acetylesterase;OSJNBa0059D20.8-like protein;Putative bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1;Integrase catalytic domain-containing protein	Cstv	miRNA	B_MiRBaseMonocot
M-250	GGCAAGUCCGUCCUUGGCUAC	miR169	2A,2B,2D		XLOC_101840(2.0)		Cstv	miRNA	B_MiRBaseMonocot
M-251	GGCAAGUCCGUCCUUGGCUACA	miR169	2A,2B,2D		XLOC_101840(2.0)		Cstv	miRNA	B_MiRBaseMonocot
M-252	GGGGUCGAACUGGGAACACAU	noHit	3B	Beta-Casp domain-containing protein	XLOC_057863(0.0),XLOC_064449(1.0),XLOC_065563(1.5),XLOC_046804(1.5)	Beta-Casp domain-containing protein	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-253	GUAGGAUGGCUGGUGCUAUGG	miR9674	4B,4D		XLOC_129417(2.0),XLOC_092129(2.0),XLOC_079645(2.0)		Cstv	miRNA	C_MiRBaseAll
M-254	GUCGGACCAGGCUUCAUUCCC	miR166	6A,6B,6D		XLOC_003329(1.0),XLOC_010176(1.0),XLOC_140819(1.0),XLOC_147220(1.0),XLOC_132764(1.0),XLOC_110868(1.5),XLOC_103056(1.5),XLOC_095467(1.5),XLOC_001247(2.0),XLOC_008070(2.0),XLOC_015254(2.0),XLOC_059246(2.0),XLOC_084967(2.0),XLOC_109041(2.0),XLOC_105100(2.0),XLOC_098448(2.0),XLOC_092135(2.0),XLOC_090000(2.0),XLOC_051179(2.0),XLOC_076905(2.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	CRdw	miRNA	B_MiRBaseMonocot
M-255	GUCUCUGCCAAUUCUUCGUGU	noHit	7B,7D		XLOC_019968(2.0)	Oxygen-evolving enhancer protein 3-1, chloroplastic;Os07g0105600 protein (PsbQ domain protein family, putative-like protein) (cDNA clone:J013148M09, full insert sequence)	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-256	GUGCUAUGGAUAAAUUCAACC	miR9674	3A,3B,3D,4D,6B,6D		XLOC_117006(0.0),XLOC_123290(0.0),XLOC_061178(1.0),XLOC_129417(1.0),XLOC_092129(1.0),XLOC_079645(1.0),XLOC_065748(1.5),XLOC_129409(2.0)		Cstv	miRNA	C_MiRBaseAll
M-257	GUGUUCUCAGGUCGCCCCCGC	miR398	3A,3B,3D	Beta-Casp domain-containing protein	XLOC_064449(0.0),XLOC_065563(0.0),XLOC_057863(0.0),XLOC_046804(0.0),XLOC_083186(2.0),XLOC_103675(2.0)	Beta-Casp domain-containing protein;Expressed protein (Os12g0580600 protein)	CRdw	miRNA	A_MiRBaseTae
M-258	GUUGCCGUAGCGCUGUGGGCC	noHit	7A	mRNA	XLOC_138680(0.0),XLOC_144566(0.0),XLOC_003442(2.0),XLOC_064213(2.0)	TFIIS N-terminal domain-containing protein;MATH domain-containing protein	CRdw	miRNA	F_PSUSStaeMirSir
M-259	UAAAAAAUCUGUAGGCACUAC	noHit	5B,6D,7A		XLOC_122492(1.0),XLOC_121403(1.0),XLOC_036886(1.5),XLOC_030899(2.0),XLOC_127658(2.0),XLOC_047682(2.0),XLOC_078140(2.0)	Aminotran_1_2 domain-containing protein;Putative homeobox-leucine zipper protein HOX26;Os01g0704200 protein;Pentatricopeptide repeat-containing protein ELI1 chloroplastic;NB-ARC domain-containing protein;Homeobox domain-containing protein;Alanine aminotransferase 2	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-260	UAAAAAUUGCGGAUUUUUGUCU	noHit	7A,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_081366(2.0),XLOC_088588(2.0),XLOC_139134(2.0),XLOC_133505(2.0),XLOC_139391(2.0),XLOC_145885(2.0),XLOC_130378(2.0),XLOC_130100(2.0),XLOC_048529(2.0),XLOC_071855(2.0)	Cysteine-rich receptor-like protein kinase 15;1-phosphatidylinositol-3-phosphate 5-kinase;GRF-type domain-containing protein;Protein kinase domain-containing protein	Cstv	miRNA	H_novel
M-261	UAAGGGAUUUCAAAAUGGACU	noHit	4A	mRNA	XLOC_079700(0.5),XLOC_127874(2.0),XLOC_132175(2.0),XLOC_114907(2.0),XLOC_101548(2.0)	Protein kinase domain-containing protein;Tyrosine-sulfated glycopeptide receptor 1;Receptor-like serine/threonine-protein kinase;Serine/threonine-protein kinase	Cstv	miRNA	H_novel
M-262	UAAUCUUCUGGAAAUAUGCUU	noHit	3B	mRNA;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);(Fragment);Disease resistance protein RPP13;CNL-C (Disease resistance RPP13-like protein 4);NB-ARC domain-containing protein	XLOC_063979(0.0),XLOC_052866(0.0),XLOC_004306(1.5),XLOC_012143(1.5),XLOC_017173(1.5),XLOC_059931(1.5),XLOC_059932(1.5),XLOC_058395(1.5),XLOC_062352(1.5),XLOC_064331(1.5),XLOC_134203(1.5),XLOC_139841(1.5),XLOC_134281(1.5),XLOC_136114(1.5),XLOC_118267(1.5),XLOC_118985(1.5),XLOC_107834(1.5),XLOC_057871(1.5),XLOC_047744(1.5),XLOC_047426(1.5),XLOC_056601(1.5),XLOC_046410(1.5),XLOC_050806(1.5),XLOC_072265(1.5),XLOC_011485(2.0),XLOC_004769(2.0),XLOC_011614(2.0),XLOC_011706(2.0),XLOC_032245(2.0),XLOC_064275(2.0),XLOC_062178(2.0),XLOC_121541(2.0),XLOC_118591(2.0),XLOC_102971(2.0),XLOC_047371(2.0)	TTF-type domain-containing protein;Putative disease resistance RPP13-like protein 3;CNL-C (Disease resistance RPP13-like protein 4);Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13;WRKY domain-containing protein;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);Rx_N domain-containing protein;(Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;Disease resistance RPP13-like protein 4	Cstv	miRNA	D_LiteratureTaeMiRNA2015

M-263	UAAUCUUCUGGAAAUAUGCUUA	noHit	3B	mRNA;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence); (Fragment);Disease resistance protein RPP13;CNL-C (Disease resistance RPP13-like protein 4);NB-ARC domain-containing protein	XLOC_063979(0.0),XLOC_052866(0.0),XLOC_004306(1.5),XLOC_012143(1.5),XLOC_017173(1.5),XLOC_059932(1.5),XLOC_059931(1.5),XLOC_058395(1.5),XLOC_062352(1.5),XLOC_064331(1.5),XLOC_134203(1.5),XLOC_139841(1.5),XLOC_136114(1.5),XLOC_134281(1.5),XLOC_118985(1.5),XLOC_118267(1.5),XLOC_107834(1.5),XLOC_057871(1.5),XLOC_047744(1.5),XLOC_047426(1.5),XLOC_056601(1.5),XLOC_046410(1.5),XLOC_050806(1.5),XLOC_072265(1.5),XLOC_011485(2.0),XLOC_004769(2.0),XLOC_011614(2.0),XLOC_011706(2.0),XLOC_032245(2.0),XLOC_064275(2.0),XLOC_062178(2.0),XLOC_121541(2.0),XLOC_118591(2.0),XLOC_102971(2.0),XLOC_047371(2.0)	TTF-type domain-containing protein;Putative disease resistance RPP13-like protein 3;CNL-C (Disease resistance RPP13-like protein 4);Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13;WRKY domain-containing protein;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence);Rx_N domain-containing protein; (Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;Disease resistance RPP13-like protein 4	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-264	UACGGCCUGAUGACAUCCACA	noHit	3D				CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-265	UAGAAGAUUAGAAGAUUAAGC	noHit	3B,3D	mRNA;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence); (Fragment);Disease resistance protein RPP13;CNL-C (Disease resistance RPP13-like protein 4);NB-ARC domain-containing protein	XLOC_063979(0.0),XLOC_052866(0.0),XLOC_100286(1.5),XLOC_063832(2.0),XLOC_111698(2.0),XLOC_104645(2.0)	Calcium-binding EF-hand family like-protein;CNL-C (Disease resistance RPP13-like protein 4);Disease resistance protein RPP13;Os09g0483100 protein;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence); (Fragment);NB-ARC domain-containing protein;START domain-containing protein;ANK_REP_REGION domain-containing protein;Protein binding protein;Os02g0735700 protein	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-266	UAGAUUGAUUGGCAAAGUACG	noHit	7D	40S ribosomal protein S27	XLOC_146251(1.0),XLOC_125182(1.5),XLOC_119717(1.5),XLOC_112401(1.5)	Luminal-binding protein 1;Helicase C-terminal domain-containing protein;40S ribosomal protein S27;Os02g0710900 protein	CRdw	miRNA	H_novel
M-267	UAGCCAAGAAUGACUUGCCUG	miR169	2A,2B		XLOC_143891(1.5),XLOC_151645(1.5),XLOC_109128(1.5),XLOC_027763(2.0),XLOC_088522(2.0),XLOC_079950(2.0),XLOC_110064(2.0),XLOC_107389(2.0),XLOC_103660(2.0),XLOC_097529(2.0),XLOC_069195(2.0)	Nuclear transcription factor Y subunit A-2;Nuclear transcription factor Y subunit A-1;Os03g0696300 protein (Fragment)	CRdw	miRNA	B_MiRBaseMonocot
M-268	UAGGGUGUAGAAUAAGCUAUU	noHit	4B,4D,5D,7B		XLOC_083577(1.5),XLOC_088897(2.0),XLOC_085890(2.0),XLOC_086533(2.0),XLOC_146455(2.0),XLOC_103491(2.0),XLOC_109763(2.0),XLOC_077929(2.0)	Polypyrimidine tract-binding protein-like protein 1;Os03g0376900 protein (Polypyrimidine tract-binding protein, putative, expressed) (cDNA clone:001-045-E01, full insert sequence);Hydrolase_4 domain-containing protein;RRM_8 domain-containing protein;Os03g0359000 protein (Fragment);RRM domain-containing protein;ANK_REP_REGION domain-containing protein;C2H2-type domain-containing protein	Cstv	miRNA	F_PSUSStaeMirSir
M-269	UAUGGACCGCGGACAACGAGGA	noHit	1B				Cstv	miRNA	H_novel
M-270	UAUUAUGGUGCGGAGGGAGUA	noHit	1D,6D		XLOC_000238(0.5),XLOC_031108(1.0),XLOC_055800(1.0),XLOC_081215(1.0),XLOC_137671(1.0),XLOC_111798(1.0),XLOC_119928(1.0),XLOC_044310(1.0),XLOC_007103(1.5),XLOC_033977(1.5),XLOC_063632(1.5),XLOC_141009(1.5),XLOC_123876(1.5),XLOC_130055(1.5),XLOC_132155(1.5),XLOC_121403(1.5),XLOC_112071(1.5),XLOC_102007(1.5),XLOC_099824(1.5),XLOC_093851(1.5),XLOC_043606(1.5),XLOC_033351(1.5),XLOC_000930(2.0),XLOC_007239(2.0),XLOC_006435(2.0),XLOC_003060(2.0),XLOC_002102(2.0),XLOC_002186(2.0),XLOC_002097(2.0),XLOC_012781(2.0),XLOC_019102(2.0),XLOC_017282(2.0),XLOC_013212(2.0),XLOC_017495(2.0),XLOC_014238(2.0),XLOC_021601(2.0),XLOC_014419(2.0),XLOC_014616(2.0),XLOC_012050(2.0),XLOC_012046(2.0),XLOC_011094(2.0),XLOC_011059(2.0),XLOC_013030(2.0),XLOC_017609(2.0),XLOC_019254(2.0),XLOC_020816(2.0),XLOC_014103(2.0),XLOC_013691(2.0),XLOC_015768(2.0),XLOC_016690(2.0),XLOC_016124(2.0),XLOC_013538(2.0),XLOC_014963(2.0),XLOC_029560(2.0),XLOC_029497(2.0),XLOC_020353(2.0),XLOC_020626(2.0),XLOC_030603(2.0),XL	Non-specific serine/threonine protein kinase;Rhodanese domain-containing protein;ABC1 domain-containing protein;M20_dimer domain-containing protein;3-hydroxyacyl-CoA dehydrogenase;Cyclopropane-fatty-acyl-phospholipid synthase;RRM domain-containing protein;Reverse transcriptase domain-containing protein;3-hydroxyisobutyryl-CoA hydrolase-like protein 2, mitochondrial;Cytochrome P450 97B2, chloroplastic;DUF4210 domain-containing protein;Chlorophyllide a oxygenase, chloroplastic (Chlorophyll a oxygenase);Phosphopantothenoylcysteine decarboxylase (Putative phosphopantothenoylcysteine decarboxylase);TPX2 domain-containing protein;tRNA pseudouridine synthase (tRNA pseudouridine synthase 3);Os09g0386500 protein (cDNA clone:001-125-G12, full insert sequence) (cDNA clone:J023141K03, full insert sequence);Ras-related protein Rab-18;12-oxophytodienoate reductase 1;Os01g0704200 protein;Putative G protein-coupled receptor;Os01g0679600 protein;Os10g0475900 protein (Ubiquitin family protein, expressed) (cDNA clone:J033069L14, full insert sequence);RNA-binding protein 40;Amino-acid N-acetyltransferase;FYVE-type domain-containing	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-271	UCAAGAAGAUAGAGAGCCCUGGCG	noHit	3A,3B		XLOC_030705(2.0),XLOC_110739(2.0),XLOC_039639(2.0)	Calcium-dependent protein kinase 2	Cstv	miRNA	H_novel
M-272	UCACCGGCGCUGCACACAAUG	miR397	6A,6B	MIR,mRNA,miRNA,microRNA;Os02g0160900 protein;	XLOC_115042(0.0),XLOC_129165(1.0),XLOC_148265(1.5),XLOC_039333(1.5)	Os02g0160900 protein;RNase H domain-containing protein;Peroxisomal membrane protein PMP22;HMA domain-containing protein;Retrovirus-related Pol polyprotein from transposon TNT 1-94	CRdw	miRNA	A_MiRBaseTae
M-273	UCAGGAAGGACCGCAUCAUCU	noHit	7B		XLOC_088900(1.5),XLOC_147261(1.5),XLOC_137853(1.5),XLOC_150570(1.5),XLOC_127616(1.5),XLOC_079992(2.0),XLOC_150458(2.0),XLOC_127758(2.0),XLOC_112307(2.0),XLOC_071922(2.0)	Putative disease resistance RPP8-like protein 2;Putative disease resistance protein RXW24L;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-274	UCAGUGCAAUCCCUCUGGAAU	miR393	2A,2B,2D		XLOC_019712(2.0),XLOC_030969(2.0),XLOC_037415(2.0)	Os07g0105900 protein;NB-ARC domain-containing protein;Calcineurin-like metallo-phosphoesterase superfamily protein;Metallophos domain-containing protein;Os07g0106000 protein (cDNA clone:J033024P06, full insert sequence)	Cstv	miRNA	B_MiRBaseMonocot

M-275	UCCACAGGCUUUCUUGAACUGU	miR396	2A,2B,2D,6A,6D		XLOC_022666(0.0),XLOC_030908(0.0),XLOC_087793(0.0),XLOC_080318(0.0),XLOC_133553(0.0),XLOC_148284(0.0),XLOC_124078(0.0),XLOC_126012(0.0),XLOC_114027(0.0),XLOC_112641(0.0),XLOC_112330(0.0),XLOC_119737(0.0),XLOC_036899(0.0),XLOC_069277(0.0),XLOC_070618(0.0),XLOC_027870(1.0),XLOC_022258(1.0),XLOC_088721(1.5),XLOC_087284(1.5),XLOC_080810(1.5),XLOC_070536(1.5),XLOC_110439(2.0)	Growth regulating factor 4-D1;Os03g0729500 protein;Growth regulating factor 5-A1;Growth regulating factor 6-B1;Growth-regulating factor 4 protein (Fragment);Growth regulating factor 1-D1;Growth regulating factor 4a-B1;Importin N-terminal domain-containing protein;Growth regulating factor 6-D1;Growth regulating factor 9-A1;Growth regulating factor 1-B1;Importin-beta1 (Os05g0353400 protein) (cDNA clone:J013029L11, full insert sequence);Growth regulating factor 1-A1;Importin subunit beta-1;QLQ domain-containing protein;Growth regulating factor 5-D1;Growth regulating factor 6-A1;GRF transcription factor (Growth-regulating factor);Growth-regulating factor 4 (OsGRF4) (Protein GRAIN LENGTH AND WIDTH 2) (Protein GRAIN SIZE ON CHROMOSOME 2) (Protein PANICLE TRAITS 2) (Transcription activator GRF4);WRC domain-containing protein;Growth regulating factor 5-4A;Os04g0600900 protein;Growth-regulating factor 1 (OsGRF1) (Transcription activator GRF1);Growth regulating factor 4-A1	Cstv	miRNA	B_MiRBaseMonocot
M-276	UCCAGGACUAGAAUGAGCAGC	noHit	5A,7A		XLOC_139608(2.0)	Tyrosyl-tRNA synthetase	Cstv	miRNA	F_PSUSStaeMirSir
M-277	UCCGCGAUCAUCACGACCAAA	noHit	2B				Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-278	UCCGGACAGCUGUGGUAGUGUUAC	noHit	2B,3B,4B,7B				CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-279	UCCGUCCCAUACUAUAAGAGC	noHit	1B,7A	,mRNA;	XLOC_021601(0.0),XLOC_126251(0.0),XLOC_093405(1.0),XLOC_002097(1.5),XLOC_008883(1.5),XLOC_007103(1.5),XLOC_008341(1.5),XLOC_008860(1.5),XLOC_004153(1.5),XLOC_007545(1.5),XLOC_001721(1.5),XLOC_001073(1.5),XLOC_002186(1.5),XLOC_004748(1.5),XLOC_016690(1.5),XLOC_015768(1.5),XLOC_023206(1.5),XLOC_019699(1.5),XLOC_013219(1.5),XLOC_013141(1.5),XLOC_021271(1.5),XLOC_018088(1.5),XLOC_013247(1.5),XLOC_014419(1.5),XLOC_021617(1.5),XLOC_017574(1.5),XLOC_028858(1.5),XLOC_028499(1.5),XLOC_029754(1.5),XLOC_020626(1.5),XLOC_025731(1.5),XLOC_020747(1.5),XLOC_032230(1.5),XLOC_024610(1.5),XLOC_062952(1.5),XLOC_068471(1.5),XLOC_062131(1.5),XLOC_055499(1.5),XLOC_064327(1.5),XLOC_066779(1.5),XLOC_063621(1.5),XLOC_068954(1.5),XLOC_062217(1.5),XLOC_059794(1.5),XLOC_059824(1.5),XLOC_066432(1.5),XLOC_055841(1.5),XLOC_055619(1.5),XLOC_080105(1.5),XLOC_085438(1.5),XLOC_088976(1.5),XLOC_084794(1.5),XLOC_082945(1.5),XLOC_080484(1.5),XLOC_081384(1.5),XLOC_080386(1.5),XLOC_080605(1.5),XLOC_080615(1.5),XL	Os11g0275500 protein;Non-specific serine/threonine protein kinase;CBS domain-containing protein;CULLIN_2 domain-containing protein;ABC1 domain-containing protein;M20_dimer domain-containing protein;Cyclopropane-fatty-acyl-phospholipid synthase;Putative copper-transporting ATPase 3;RRM domain-containing protein;Os03g0679700 protein (Thiamine biosynthesis protein thiC, putative, expressed) (cDNA clone:J013044A21, full insert sequence) (cDNA clone:J013159A12, full insert sequence);Reverse transcriptase domain-containing protein;Calcium-dependent protein kinase 32;Os07g0513200 protein (cDNA clone:J033037I20, full insert sequence);Cytochrome P450 97B2, chloroplastic;DUF4210 domain-containing protein;Calcineurin subunit B;Os02g0666700 protein;Protein HASTY 1;Delta-1-pyrroline-5-carboxylate synthase 2 (OsP5CS2) [Includes: Glutamate 5-kinase (GK);Putative G protein-coupled receptor;Os08g0387500 protein (cDNA clone:001-100-E10, full insert sequence);RNA-binding protein 40;Os09g0489200 protein;Os08g0238200 protein (cDNA clone:J013096H10, full insert sequence);Os01g0355900 protein (cDNA clone:J023041L20, full insert sequence);UDP-glucuronic acid decarboxylase 1-Delta-1	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-280	UCGGACCAGGCUUCAUUCCCCC	miR166	4A,4D,7A,7B,7D		XLOC_003329(0.0),XLOC_010176(0.0),XLOC_140819(0.0),XLOC_147220(0.0),XLOC_132764(0.0),XLOC_110868(0.5),XLOC_103056(0.5),XLOC_095467(0.5),XLOC_001247(1.0),XLOC_008070(1.0),XLOC_015254(1.0),XLOC_059246(1.0),XLOC_084967(1.0),XLOC_109041(1.0),XLOC_105100(1.0),XLOC_098448(1.0),XLOC_092135(1.0),XLOC_090000(1.0),XLOC_051179(1.0),XLOC_076905(1.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot
M-281	UCGGACCAGGCUUCAUUCCCCU	miR166	1A,1B,1D		XLOC_003329(0.0),XLOC_010176(0.0),XLOC_140819(0.0),XLOC_147220(0.0),XLOC_132764(0.0),XLOC_110868(0.5),XLOC_103056(0.5),XLOC_095467(0.5),XLOC_001247(1.0),XLOC_008070(1.0),XLOC_015254(1.0),XLOC_059246(1.0),XLOC_084967(1.0),XLOC_109041(1.0),XLOC_105100(1.0),XLOC_098448(1.0),XLOC_092135(1.0),XLOC_090000(1.0),XLOC_051179(1.0),XLOC_076905(1.0)	Os03g0109400 protein;Homeobox-leucine zipper protein HOX32 (HD-ZIP protein HOX32) (Homeodomain transcription factor HOX32) (OsHox32);Homeobox-leucine zipper protein HOX10;Homeobox-leucine zipper protein HOX33;Homeobox-leucine zipper protein REVOLUTA;Homeobox-leucine zipper protein HOX9 (HD-ZIP protein HOX9) (Homeodomain transcription factor HOX9) (OsHB2) (OsHox9);START domain-containing protein;Homeobox-leucine zipper protein HOX9;Homeobox-leucine zipper protein HOX10 (HD-ZIP protein HOX10) (Homeodomain transcription factor HOX10) (OsHB1) (OsHox10);Homeobox-leucine zipper protein HOX33 (HD-ZIP protein HOX33) (Homeodomain transcription factor HOX33) (OsHox33)	Cstv	miRNA	B_MiRBaseMonocot

M-282	UCUGAAAUUUUGGAGUGCACC	noHit	4A,4D		XLOC_135981(1.5),XLOC_133878(1.5),XLOC_090140(1.5),XLOC_036475(1.5),XLOC_029272(2.0),XLOC_063807(2.0),XLOC_084276(2.0),XLOC_116823(2.0),XLOC_121101(2.0),XLOC_046520(2.0),XLOC_071369(2.0),XLOC_071957(2.0)	Choline/ethanolamine kinase;USP domain-containing protein;Protein kinase domain-containing protein;ATP-citrate synthase beta chain protein 1 (ATP-citrate synthase B-1);Protein FAR1-RELATED SEQUENCE;ATP-citrate synthase beta chain protein 2;Wall-associated receptor kinase 1;Superoxide dismutase;SWIM-type domain-containing protein;Ubiquitin carboxyl-terminal hydrolase;ATP citrate synthase;F-box domain-containing protein	Cstv	miRNA	H_novel
M-283	UCUGGCGAGGGACAUACACUGU	miR9678	7A,7B,7D		XLOC_123892(2.0)	Serine protease SPPA chloroplastic;Protease 4;Peptidase_S49 domain-containing protein;Os02g0728100 protein	CRdw	miRNA	A_MiRBaseTae
M-284	UCUGUCCCAAAUUACAAGAUG	noHit	2A,2D,6A,7A		XLOC_088934(1.0),XLOC_088111(1.0),XLOC_022714(1.5),XLOC_080147(1.5),XLOC_043556(1.5),XLOC_071493(1.5),XLOC_021066(2.0),XLOC_029494(2.0),XLOC_038046(2.0)	Os03g0452300 protein (Ribosomal protein S5 containing protein, expressed) (cDNA clone:001-035-H10, full insert sequence) (cDNA clone:001-037-G07, full insert sequence) (cDNA clone:J013042D13, full insert sequence);Exocyst subunit Exo70 family protein;Ribosomal_S5_C domain-containing protein;Mediator of RNA polymerase II transcription subunit 30;S5 DRBM domain-containing protein;30S ribosomal protein S5 chloroplastic;(3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase	Cstv	miRNA	F_PSUSStaeMirSir
M-285	UCUGUCCCAUAGUAUAAGAGC	noHit	4A	mRNA	XLOC_001215(0.5),XLOC_005739(0.5),XLOC_000930(0.5),XLOC_005175(0.5),XLOC_021781(0.5),XLOC_033891(0.5),XLOC_029560(0.5),XLOC_062108(0.5),XLOC_064559(0.5),XLOC_054479(0.5),XLOC_059901(0.5),XLOC_088770(0.5),XLOC_084009(0.5),XLOC_081599(0.5),XLOC_141415(0.5),XLOC_144231(0.5),XLOC_143748(0.5),XLOC_142226(0.5),XLOC_136528(0.5),XLOC_146894(0.5),XLOC_148810(0.5),XLOC_148583(0.5),XLOC_134846(0.5),XLOC_131116(0.5),XLOC_110029(0.5),XLOC_119146(0.5),XLOC_118688(0.5),XLOC_104179(0.5),XLOC_098110(0.5),XLOC_100561(0.5),XLOC_100793(0.5),XLOC_091086(0.5),XLOC_094756(0.5),XLOC_095336(0.5),XLOC_045875(0.5),XLOC_053145(0.5),XLOC_044145(0.5),XLOC_034533(0.5),XLOC_040075(0.5),XLOC_032720(0.5),XLOC_036453(0.5),XLOC_078054(0.5),XLOC_075145(0.5),XLOC_078967(0.5),XLOC_069723(0.5),XLOC_002097(1.0),XLOC_008883(1.0),XLOC_007103(1.0),XLOC_008341(1.0),XLOC_008860(1.0),XLOC_004153(1.0),XLOC_007545(1.0),XLOC_001721(1.0),XLOC_001073(1.0),XLOC_002186(1.0),XLOC_004748(1.0),XLOC_015609(1.0),XLOC_015768(1.0),XL	Os11g0275500 protein;Putative WRKY transcription factor 14;Non-specific serine/threonine protein kinase;CBS domain-containing protein;RNA-binding protein CP31B chloroplastic;Vacuolar fusion protein MON1 homolog;Putative copper-transporting ATPase 3;Putative D-2-hydroxyglutarate dehydrogenase, mitochondrial;Putative G protein-coupled receptor;RNA-binding protein 40;Cullin-associated NEDD8-dissociated protein 1; (Fragment);Os01g0355900 protein (cDNA clone:J023041L20, full insert sequence);RNase H domain-containing protein;SCARECROW-LIKE protein 7 (Os-SCL7);Acyltransferase family protein, expressed (cDNA clone:J013152C07, full insert sequence);ATPase 100191268;Putative WRKY DNA-binding domain superfamily protein;DUF1338 domain-containing protein;Alpha/beta-Hydrolases superfamily protein;RAI1 domain-containing protein;BHLH domain-containing protein;Cytochrome b5 heme-binding domain-containing protein;EPIDERMAL PATTERNING FACTOR-like protein 6;Os06g0548100 protein;Coatomer subunit delta-1 (Delta-coat protein 1) (Delta-COP 1);Peptidylprolyl isomerase;Os03g0290400 protein (Fragment);Os07g0121200 protein;Metal tolerance protein 4 (OsMTP4);Putative phenylpropathoxamlcystaine decarboxylase-TiS;	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-286	UCUGUUCCUAAAUGUAAGCCU	noHit	3B,5A,6D	,MIR,mRNA,miRNA,microRNA;	XLOC_141063(0.0),XLOC_022359(0.5),XLOC_062382(0.5),XLOC_142817(0.5),XLOC_042150(0.5),XLOC_076716(0.5),XLOC_006930(1.0),XLOC_015773(1.0),XLOC_029097(1.0),XLOC_056398(1.0),XLOC_149817(1.0),XLOC_132692(1.0),XLOC_110167(1.0),XLOC_101548(1.0),XLOC_098387(1.0),XLOC_093854(1.0),XLOC_099081(1.0),XLOC_097862(1.0),XLOC_097859(1.0),XLOC_050133(1.0),XLOC_042489(1.0),XLOC_032691(1.0),XLOC_039298(1.0),XLOC_076687(1.0),XLOC_071127(1.0),XLOC_010175(1.5),XLOC_003666(1.5),XLOC_008510(1.5),XLOC_009318(1.5),XLOC_003640(1.5),XLOC_006854(1.5),XLOC_001498(1.5),XLOC_017441(1.5),XLOC_011618(1.5),XLOC_012143(1.5),XLOC_012465(1.5),XLOC_018957(1.5),XLOC_012193(1.5),XLOC_017350(1.5),XLOC_030423(1.5),XLOC_031134(1.5),XLOC_026857(1.5),XLOC_031338(1.5),XLOC_065416(1.5),XLOC_058986(1.5),XLOC_063725(1.5),XLOC_088684(1.5),XLOC_081166(1.5),XLOC_082362(1.5),XLOC_084210(1.5),XLOC_083491(1.5),XLOC_093330(1.5),XLOC_139983(1.5),XLOC_143987(1.5),XLOC_142518(1.5),XLOC_136027(1.5),XL	E3 ubiquitin-protein ligase RNF170 (RING finger protein 170) (RING-type E3 ubiquitin transferase RNF170);Non-specific serine/threonine protein kinase;DEAD-box ATP-dependent RNA helicase 32;ABC1 domain-containing protein;Pumilio-like protein 1;Os02g0586400 protein (cDNA clone:006-207-A09, full insert sequence) (cDNA clone:J023004J08, full insert sequence);Glutamate dehydrogenase;Ubiquinone biosynthesis protein COQ4 homolog, mitochondrial (Coenzyme Q biosynthesis protein 4 homolog);Malate dehydrogenase (NADP(+));Tyrosyl-tRNA synthetase;Transposase_23 domain-containing protein;Cyclopropane-fatty-acyl-phospholipid synthase;CCT domain-containing protein;RRM domain-containing protein;RING/U-box superfamily protein;Beta-glucuronosyltransferase GlcAT14B;TPR repeat-containing thioredoxin TDX;Ypt homolog3;Putative UDP-arabinopyranose mutase 5;Os03g0292200 protein;Eukaryotic translation initiation factor 2A;Serine/threonine-protein kinase Nek3;PseudoU_synth_1 domain-containing protein;Amino-acid N-acetyltransferase;Disease resistance protein RGA2;Pyrophosphate--fructose 6-phosphate 1-phosphotransferase subunit beta (PFP); (Fragment);FYVE-type domain-containing	Cstv	miRNA	E_LiteratureTaeMiRNA2020

M-287	UGAAAUGAGUGAGCAAAGACA	noHit	4B		XLOC_078369(0.0),XLOC_073971(1.0),XLOC_012734(1.5),XLOC_018413(1.5),XLOC_023213(1.5),XLOC_016807(1.5),XLOC_027825(1.5),XLOC_068449(1.5),XLOC_059204(1.5),XLOC_064666(1.5),XLOC_083265(1.5),XLOC_091275(1.5),XLOC_138918(1.5),XLOC_142314(1.5),XLOC_134035(1.5),XLOC_150791(1.5),XLOC_122813(1.5),XLOC_122210(1.5),XLOC_112295(1.5),XLOC_119854(1.5),XLOC_102928(1.5),XLOC_109359(1.5),XLOC_106613(1.5),XLOC_104316(1.5),XLOC_096637(1.5),XLOC_092529(1.5),XLOC_100631(1.5),XLOC_092858(1.5),XLOC_053085(1.5),XLOC_051472(1.5),XLOC_057004(1.5),XLOC_057961(1.5),XLOC_036724(1.5),XLOC_044880(1.5),XLOC_038073(1.5),XLOC_036827(1.5),XLOC_078577(1.5),XLOC_076636(1.5),XLOC_076786(1.5),XLOC_070618(1.5),XLOC_002881(2.0),XLOC_026425(2.0),XLOC_028362(2.0),XLOC_029646(2.0),XLOC_022110(2.0),XLOC_022139(2.0),XLOC_066292(2.0),XLOC_140065(2.0),XLOC_131614(2.0),XLOC_121569(2.0),XLOC_124037(2.0),XLOC_129064(2.0),XLOC_123994(2.0),XLOC_115375(2.0),XLOC_100929(2.0),XLOC_107383(2.0),XLOC_097887(2.0),XLOC_057174(2.0),XL	Os09g0482720 protein;Mitotic spindle checkpoint protein MAD2;2Fe-2S ferredoxin-type domain-containing protein;Glyco_trans_4-like_N domain-containing protein;Sulfoquinovosyl transferase SQD2;STAS domain-containing protein;CCT domain-containing protein;Zinc finger protein CONSTANS;Os04g0486500 protein (cDNA clone:001-007-B02, full insert sequence);F-box domain-containing protein;2Fe-2S ferredoxin;Phosphate transporter PHO1-3;Sulfate transporter 1.2;PCI domain-containing protein;B2-adapt-app_C domain-containing protein;FYVE-type domain-containing protein;COP9 signalosome complex subunit 1, putative, expressed (Os03g0116500 protein) (cDNA clone:J033070E21, full insert sequence);ADP-ribosylation factor-like protein 1;Retrovirus-related Pol polyprotein LINE-1;AMP-binding domain-containing protein;BPI2 domain-containing protein;RNase H domain-containing protein;Receptor-like serine/threonine-protein kinase;Expressed protein (Os10g0573100 protein) (cDNA clone:J023001I03, full insert sequence);Os04g0348000 protein (cDNA clone:J033138G07, full insert sequence);DEAD-box ATP-dependent RNA helicase 8;Long chain acyl-CoA synthetase 9, chloroplastic;High affinity sulfate transporter 1;Nuclear pore complex protein NUP205;AMP-binding enzyme family	Cstv	miRNA	F_PSUSStaeMirSir
M-288	UGAACCUUAGGGAACAGCCGCAU	miR5062	5B		XLOC_058622(0.5),XLOC_106077(0.5),XLOC_092193(0.5),XLOC_106053(1.0),XLOC_105192(1.0),XLOC_092721(1.5),XLOC_104944(2.0),XLOC_098330(2.0)	Protein argonaute MEL1	Cstv	miRNA	A_MiRBaseTae
M-289	UGAAGCUGCCAGCAUGAUCUG	miR167	4A,4D,5A,5B,5D,6A				CRdw	miRNA	A_MiRBaseTae
M-290	UGAGAUGAGAUUACCCCAUAC	miR9772	4A,5A,5B,5D		XLOC_102302(0.0),XLOC_101813(0.0),XLOC_122610(1.0),XLOC_121718(1.5),XLOC_114720(1.5),XLOC_122290(1.5),XLOC_120961(1.5),XLOC_115019(1.5),XLOC_073377(1.5),XLOC_017413(2.0),XLOC_012082(2.0),XLOC_121450(2.0),XLOC_115971(2.0)	F-box domain-containing protein;	Cstv	miRNA	A_MiRBaseTae
M-291	UGAGCAUCAACUAAUUCCGGA	noHit	1A,1B,1D,2B,3D,5D	,mRNA;	XLOC_102825(0.0),XLOC_011476(0.5),XLOC_150530(1.0),XLOC_082288(1.5),XLOC_150000(1.5),XLOC_109693(1.5),XLOC_102482(1.5),XLOC_049307(1.5),XLOC_036792(1.5),XLOC_122163(2.0),XLOC_052005(2.0),XLOC_037092(2.0)	ULP_PROTEASE domain-containing protein;Protein kinase domain-containing protein;DUF1618 domain-containing protein;Receptor-like serine/threonine-protein kinase;Serine/threonine-protein kinase;DUF4210 domain-containing protein	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-292	UGAGCCGAACCAAUAUCACUC	miR171	2A,2B,2D	MIR,mRNA,microRNA;	XLOC_083692(0.0),XLOC_079087(0.0),XLOC_073932(0.0)	Protein NODULATION SIGNALING PATHWAY 2 (OsNSP2) (GRAS family protein 13) (OsGRAS-13);GRAS domain-containing protein;GRAS transcription factor (Scarecrow-like protein 26) (Fragment)	CRdw	miRNA	B_MiRBaseMonocot
M-293	UGAGGAACGAAAGAAUUGUGUAUU	noHit	1B		XLOC_080517(1.5)	E3 ubiquitin-protein ligase SRFP1 (OsSRFP1)	Cstv	miRNA	H_novel
M-294	UGAUUAUCGAAGUUUCAGACUU	noHit	1B		XLOC_030492(2.0),XLOC_030982(2.0),XLOC_063827(2.0),XLOC_059407(2.0),XLOC_054913(2.0)	SSD domain-containing protein;Beta-galactosidase;GRF-type domain-containing protein;RING-type E3 ubiquitin transferase;Niemann-Pick C1 protein	Cstv	miRNA	H_novel
M-295	UGAUUGAGCCGUGCCAAUAUC	miR171	4A,4B,4D,7A,7B,7D	MIR,mRNA,microRNA;	XLOC_094074(0.0),XLOC_123950(0.5),XLOC_004052(2.0)	Os02g0662700 protein;Os02g0663100 protein (cDNA clone:J023150P14, full insert sequence);GRAS domain-containing protein;Phosphomannomutase	Cstv	miRNA	A_MiRBaseTae
M-296	UGCAUCAUCUCGAACUCGUCGG	miR9778	7A,7B,7D	,mRNA;Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein;	XLOC_147261(0.0),XLOC_004785(0.5),XLOC_004095(1.0),XLOC_017292(1.0),XLOC_058711(1.0),XLOC_060090(1.0),XLOC_136162(1.0),XLOC_146010(1.0),XLOC_150570(1.0),XLOC_146796(1.0),XLOC_095247(1.0),XLOC_039460(1.0),XLOC_004987(1.5),XLOC_012250(1.5),XLOC_010815(1.5),XLOC_043149(1.5),XLOC_007467(2.0),XLOC_025363(2.0),XLOC_025542(2.0),XLOC_088900(2.0),XLOC_089124(2.0),XLOC_135317(2.0),XLOC_149072(2.0),XLOC_122763(2.0),XLOC_121477(2.0),XLOC_109872(2.0),XLOC_043884(2.0),XLOC_034177(2.0),XLOC_071922(2.0)	Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein (Fragment);Cysteine-rich receptor-like protein kinase 10;Putative disease resistance protein RXW24L;Disease resistance protein RPM1 (Resistance protein RGA2);Protein kinase domain-containing protein;Disease resistance protein RPP13;Resistance protein RGA2;AAA domain-containing protein;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;DUF676 domain-containing protein;Os08g0447200 protein	Cstv	miRNA	A_MiRBaseTae
M-297	UGCAUCAUCUCGAACUCGUCGGU	miR9778	7A,7B,7D	,mRNA;Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein;	XLOC_147261(0.0),XLOC_004785(0.5),XLOC_004095(1.0),XLOC_017292(1.0),XLOC_058711(1.0),XLOC_060090(1.0),XLOC_146010(1.0),XLOC_136162(1.0),XLOC_146796(1.0),XLOC_150570(1.0),XLOC_095247(1.0),XLOC_039460(1.0),XLOC_004987(1.5),XLOC_012250(1.5),XLOC_010815(1.5),XLOC_043149(1.5),XLOC_007467(2.0),XLOC_025363(2.0),XLOC_025542(2.0),XLOC_088900(2.0),XLOC_089124(2.0),XLOC_135317(2.0),XLOC_149072(2.0),XLOC_122763(2.0),XLOC_121477(2.0),XLOC_109872(2.0),XLOC_043884(2.0),XLOC_034177(2.0),XLOC_071922(2.0)	Putative disease resistance RPP8-like protein 2;Rx_N domain-containing protein (Fragment);Cysteine-rich receptor-like protein kinase 10;Putative disease resistance protein RXW24L;Disease resistance protein RPM1 (Resistance protein RGA2);Protein kinase domain-containing protein;Disease resistance protein RPP13;Resistance protein RGA2;AAA domain-containing protein;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;DUF676 domain-containing protein;Os08g0447200 protein	Cstv	miRNA	A_MiRBaseTae

M-298	UGCAUCAUUUGGAACUCGCCG	noHit	7B,7D	;Disease resistance protein RPM1;Rx_N domain-containing protein;Putative disease resistance RPP8-like protein 2	XLOC_007467(0.5),XLOC_025542(0.5),XLOC_025363(0.5),XLOC_088900(0.5),XLOC_034177(0.5),XLOC_043884(0.5),XLOC_071922(0.5),XLOC_079992(1.0),XLOC_043881(1.0),XLOC_005193(1.5),XLOC_128720(1.5),XLOC_006323(2.0),XLOC_033991(2.0),XLOC_066749(2.0),XLOC_144367(2.0),XLOC_147504(2.0),XLOC_049967(2.0)	Putative disease resistance RPP8-like protein 2;Cysteine-rich receptor-like protein kinase 10;Putative disease resistance protein RXW24L;Protein kinase domain-containing protein;LRR receptor-like serine/threonine-protein kinase GSO1;LRRNT_2 domain-containing protein;Rx_N domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1;Disease resistance RPP13-like protein 4	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-299	UGCCAAAGGAGAGUUGCCCUG	miR399	2A,2B,2D	MIR,mRNA,miRNA,microRNA,ncRNA;	XLOC_063641(1.0),XLOC_055505(1.0),XLOC_000410(1.5),XLOC_007992(1.5),XLOC_015424(1.5),XLOC_028321(1.5),XLOC_096275(1.5),XLOC_037340(1.5),XLOC_001946(2.0),XLOC_047140(2.0)	Putative ubiquitin-conjugating enzyme E2 24;Protein tumorous imaginal discs, mitochondrial;UBC core domain-containing protein	CRupCRdw	miRNA	A_MiRBaseTae
M-300	UGCCUGGCUCCCUGUAUGCCA	miR160	5B,5D,6A,6B,6D,7A,7B,7D	MIR,mRNA,miRNA,microRNA;	XLOC_002213(0.0),XLOC_009428(0.0),XLOC_014749(0.0),XLOC_137979(0.0),XLOC_144414(0.0),XLOC_133015(0.0),XLOC_130942(0.0),XLOC_112706(0.0),XLOC_021003(0.5),XLOC_039108(0.5),XLOC_030294(1.0)	Auxin response factor 8;PB1 domain-containing protein;Auxin response factor 13;Auxin response factor;Auxin response factor 18 (OsARF10);Auxin response factor 22	Cstv	miRNA	A_MiRBaseTae
M-301	UGGAGAAGCAGGGCACGUGCA	miR164	1A,1B,1D,2A,2B,2D		XLOC_145709(1.0),XLOC_137408(1.0),XLOC_146890(1.0),XLOC_130711(1.0),XLOC_132167(1.0),XLOC_109665(1.0),XLOC_103023(1.0),XLOC_095159(1.0),XLOC_008857(1.5),XLOC_021683(1.5),XLOC_030035(1.5),XLOC_087568(1.5),XLOC_125527(1.5),XLOC_031215(2.0),XLOC_087287(2.0),XLOC_085120(2.0),XLOC_141578(2.0),XLOC_146583(2.0),XLOC_148714(2.0),XLOC_130800(2.0),XLOC_103379(2.0),XLOC_053344(2.0)	NAC domain-containing protein;Spastin;NAC domain-containing protein 21/22;Deoxyribodipyrimidine photo-lyase;Response regulatory domain-containing protein;DNA photolyase;Helix-loop-helix DNA-binding domain containing protein, expressed (Os12g0610200 protein) (cDNA clone:J013074J19, full insert sequence);UDP-glucose-fructose-phosphate glucosyltransferase;NAC transcription factor 6A;Harpin-induced protein 1 containing protein, expressed (Os12g0159600 protein) (cDNA clone:006-302-D07, full insert sequence) (cDNA clone:J023048L05, full insert sequence);Two-component response regulator ORR6 (OsRR6) (OsRRA6);Cytokinin response regulator2 (Orphans transcription factor);Os02g0639500 protein;NAC transcription factor;LEA_2 domain-containing protein;Chloroplast protein-transporting ATPase;Harpin-induced protein 1 containing protein, expressed (Os11g0157200 protein) (cDNA clone:001-207-B08, full insert sequence) (cDNA clone:002-173-F02, full insert sequence) (cDNA clone:006-301-C09, full insert sequence) (cDNA clone:006-301-F10, full insert sequence) (cDNA clone:006-306-C11, full insert sequence) (cDNA clone:J023052I16, full insert sequence);Retrovirus-related Pol polyprotein from transposon TNT 1-94;Two-component	CRdw	miRNA	A_MiRBaseTae
M-302	UGGGCAAGUCACCCUGGCUACC	miR169	4A,5B,5D		XLOC_106316(0.0),XLOC_096308(0.0),XLOC_070187(0.0)		CRdw	miRNA	B_MiRBaseMonocot
M-303	UGGGGAUCAUCGACGUCAAGC	noHit	2D		XLOC_075933(1.5),XLOC_136338(2.0),XLOC_131888(2.0)	AP2/ERF domain-containing protein	CRdw	miRNA	F_PSUSStaeMirSir
M-304	UGUAGAUACUCCCUAAGGCUU	miR5200	7A,7B,7D				Cstv	miRNA	A_MiRBaseTae
M-305	UGUCAAGAAGAUAGAGAGCCC	noHit	3A,3B		XLOC_141759(1.5),XLOC_108997(1.5),XLOC_054460(1.5),XLOC_119110(2.0),XLOC_077029(2.0)	Os03g0269100 protein (Tropinone reductase, putative, expressed);Os03g0269100 protein (Tropinone reductase, putative, expressed) (cDNA clone:001-018-F07, full insert sequence)	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-306	UGUCUGAAAAAACUUGUCAUC	noHit	6A,6D	mRNA;UBC core domain-containing protein;Ubiquitin-conjugating enzyme E2 22	XLOC_146822(1.0),XLOC_122787(1.0),XLOC_129184(1.0),XLOC_087427(1.5),XLOC_082228(1.5),XLOC_080000(1.5),XLOC_096560(1.5),XLOC_057251(1.5),XLOC_031574(2.0),XLOC_062835(2.0),XLOC_055815(2.0),XLOC_068752(2.0),XLOC_065313(2.0),XLOC_060969(2.0),XLOC_087359(2.0),XLOC_083903(2.0),XLOC_141070(2.0),XLOC_137830(2.0),XLOC_135481(2.0),XLOC_141197(2.0),XLOC_134299(2.0),XLOC_125095(2.0),XLOC_118924(2.0),XLOC_103996(2.0),XLOC_104405(2.0),XLOC_054283(2.0),XLOC_046930(2.0),XLOC_057240(2.0),XLOC_041626(2.0),XLOC_045111(2.0),XLOC_077947(2.0)	NAC domain-containing protein;DUF3511 domain protein;Fructose-bisphosphate aldolase;TF-B3 domain-containing protein;Os08g0108700 protein;Transcription factor MYB3R-2 (Myb-related protein MYB3R-2) (OsMYB3R-2);Peroxidase;Myb-related protein 3R-1;Expansin-B16;Pleckstrin homology domain-containing family A member 8;Os02g0639500 protein;Hydrolase_4 domain-containing protein;Os05g0110600 protein;Glycosyltransferase;ABC transporter C family member 3; yqxC;MFS domain-containing protein;Alpha/beta-Hydrolases superfamily protein;Multidrug resistance associated protein MRP2;S4 RNA-binding domain-containing protein;FtsJ domain-containing protein;GLTP domain-containing protein;S-acyltransferase	Cstv	miRNA	H_novel
M-307	UGUUCUGAAGAAACUGUCACC	noHit	6B		XLOC_017487(0.0),XLOC_091277(1.5),XLOC_114618(1.5),XLOC_034922(1.5),XLOC_046691(2.0)	Serine/threonine-protein phosphatase;Inactive leucine-rich repeat receptor-like protein kinase CORYNE;Protein kinase domain-containing protein;Serine/threonine-protein phosphatase BSL3;Os01g0927500 protein (cDNA clone:J013159J12, full insert sequence) (cDNA clone:J033033P12, full insert sequence)	Cstv	miRNA	E_LiteratureTaeMiRNA2020
M-308	UUAAUUUGUCCAUAGCAUCCG	noHit	3B,3D		XLOC_017227(1.0),XLOC_017089(1.0),XLOC_117006(1.5),XLOC_123290(1.5),XLOC_016763(2.0),XLOC_018015(2.0),XLOC_124403(2.0),XLOC_112916(2.0),XLOC_119566(2.0)	Protein Rf1, mitochondrial	Cstv	miRNA	D_LiteratureTaeMiRNA2015
M-309	UUAGAUGACCAUCAGCAAACA	miR827	2B,2D		XLOC_031123(0.0),XLOC_018893(1.5),XLOC_029708(1.5),XLOC_124642(1.5),XLOC_118258(1.5),XLOC_113366(1.5),XLOC_099902(1.5),XLOC_039889(1.5),XLOC_145987(2.0),XLOC_051206(2.0)	SPX domain-containing protein;Patatin group A-3;SPX domain-containing membrane protein Os04g0573000;DYW_deaminase domain-containing protein;Patatin;Os12g0163600 protein (Pentatricopeptide, putative, expressed) (cDNA clone:J013108L03, full insert sequence);SPX domain-containing membrane protein Osl_17046	CRupCRdw	miRNA	B_MiRBaseMonocot

M-310	UUCCAAAGGGAUCGCAUUGAUC	miR393	2A,2B,2D		XLOC_003903(1.0),XLOC_011568(1.0),XLOC_017972(1.0),XLOC_022116(1.0),XLOC_030681(1.0),XLOC_037090(1.0),XLOC_124694(2.0),XLOC_118338(2.0)	Transport inhibitor response 1-like protein Os05g0150500 (TIR1-like protein);Transp_inhibit domain-containing protein;Transport inhibitor response 1-like protein Os04g0395600 (TIR1-like protein);F-box domain-containing protein;Transport inhibitor response 1;Os02g0529900 protein (cDNA clone:J013154H21, full insert sequence)	Cstv	miRNA	B_MiRBaseMonocot
M-311	UUCGUGCCGGAUUAUCAUUUCA	noHit	6D	G-type lectin S-receptor-like serine/threonine-protein kinase SD1-29;Protein kinase domain-containing protein;			Cstv	miRNA	F_PSUSStaeMirSir
M-312	UUCUCCCAGAAUCUUGAACAC	noHit	1B,2B,3B,5B,6A,7B		XLOC_062921(1.0),XLOC_106165(1.0),XLOC_063479(1.5),XLOC_065927(1.5),XLOC_054598(1.5),XLOC_048737(1.5),XLOC_053333(1.5),XLOC_082176(2.0)	Lactase;Beta-galactosidase;General transcription factor 3C polypeptide 2;tRNA-specific adenosine deaminase-like protein 3;CMP/dCMP-type deaminase domain-containing protein;RING-type domain-containing protein	Cstv	miRNA	E_LiteratureTaeMiRNA2020
M-313	UUGAGACGAAAACAGACCAAC	noHit	5D		XLOC_084896(2.0),XLOC_085007(2.0),XLOC_100921(2.0)	COBRA-like protein;Os03g0292200 protein;Mitochondrial succinate-fumarate transporter 1;Expressed protein (Os03g0322600 protein) (cDNA clone:J013099A05, full insert sequence)	CRdw	miRNA	D_LiteratureTaeMiRNA2015
M-314	UUGAGUGCAGCGUUGAUGAAC	miR397	6A,6B	MIR,mRNA,miRNA,microRNA;Os02g0160900 protein;	XLOC_115042(0.0),XLOC_005767(0.5),XLOC_012268(0.5),XLOC_017797(0.5),XLOC_028715(1.5),XLOC_063411(1.5),XLOC_055788(1.5),XLOC_085889(1.5),XLOC_047175(1.5),XLOC_044403(1.5),XLOC_033279(1.5),XLOC_075021(1.5),XLOC_000475(2.0),XLOC_006321(2.0),XLOC_031122(2.0),XLOC_025201(2.0),XLOC_061669(2.0),XLOC_133647(2.0),XLOC_108606(2.0),XLOC_100583(2.0),XLOC_053933(2.0),XLOC_045825(2.0),XLOC_077486(2.0)	Os02g0160900 protein;Sugar transport protein MST3 (Monosaccharide transporter 3) (OsMST3) (Sugar:proton symporter MST3);MFS domain-containing protein;Plastocyanin-like domain-containing protein;Laccase-3;Laccase-10;Laccase-14;Os11g0108650 protein;Laccase	CRdw	miRNA	B_MiRBaseMonocot
M-315	UUGGACUGAAGGGAGCUCCCU	miR319	4A,4B,4D	MIR,microRNA;	XLOC_022594(0.5),XLOC_110530(0.5),XLOC_094745(0.5),XLOC_061121(1.0),XLOC_049362(1.0),XLOC_060579(1.0),XLOC_000634(1.5),XLOC_006325(1.5),XLOC_012922(1.5),XLOC_063612(1.5),XLOC_054670(1.5),XLOC_046151(1.5),XLOC_032275(2.0),XLOC_056481(2.0),XLOC_088604(2.0),XLOC_125777(2.0)	Transcription factor;Transcription factor GAMYB (OsGAMyb);Lipase_GDSL domain-containing protein;Transcription factor GAMYB;Putative membrane lipoprotein; Phosphatidylserine decarboxylase 1 alpha chain];Transcription factor (Fragment);DUF614 containing protein;Os01g0222001 protein (Fragment);Phosphatidylserine decarboxylase proenzyme 1, mitochondrial;Os11g0703600 protein (Fragment);Expressed protein (Os11g0703600 protein) (cDNA clone:001-008-G09, full insert sequence)	Cstv	miRNA	A_MiRBaseTae
M-316	UUGUGAAGGAAGUAGUAGUAUU	noHit	7B		XLOC_061354(0.5),XLOC_134203(2.0),XLOC_149868(2.0),XLOC_125526(2.0),XLOC_039956(2.0)	NB-ARC domain-containing protein;Disease resistance protein RPM1;Sucrose synthase	CRdw	miRNA	H_novel
M-317	UUUAAAACUUUGACUUGGAAC	noHit	3A,4D,5A,7A,7B	VRN,mRNA;	XLOC_086515(0.0),XLOC_096043(1.0),XLOC_015638(1.5),XLOC_065502(1.5),XLOC_056972(1.5),XLOC_148360(2.0),XLOC_099453(2.0)	MADS-box protein TaVRT-1 (VRN-B1) (Vernalization protein);(S)-2-hydroxy-acid oxidase;Vernalization protein (Vrn-A1) (Fragment);Breast carcinoma-amplified sequence 3-like protein;MADS box transcription factor (VRN-D1);BCAS3 domain-containing protein;MIKC-type MADS-box transcription factor WM6 (VRN-A1) (Vrd1)	Cstv	miRNA	H_novel
M-318	UUUAUGGACCGCGGACAACGAGGA	noHit	1B		XLOC_134213(2.0),XLOC_143178(2.0),XLOC_148092(2.0)	G_PROTEIN_RECEP_F2_4 domain-containing protein;Os06g0199800 protein (cDNA clone:J033046K12, full insert sequence);G-protein coupled receptor 1	Cstv	miRNA	H_novel
M-319	UUUCGAGAUCUAACUUUGACCAUU	noHit	5D		XLOC_021925(1.5),XLOC_024508(1.5),XLOC_118997(1.5),XLOC_119903(1.5),XLOC_099846(1.5),XLOC_042456(1.5),XLOC_011556(2.0)	GPI inositol-deacylase;Methyltranfer_dom domain-containing protein; (Fragment)	Cstv	miRNA	E_LiteratureTaeMiRNA2020
M-320	UUUGUGUCACUGACAUGUGGGGCC	noHit	6B		XLOC_111594(1.0),XLOC_105186(1.0),XLOC_032531(1.0),XLOC_130956(1.5),XLOC_099834(1.5),XLOC_077227(2.0)	Omega-hydroxypalmitate O-feruloyl transferase;MADF domain-containing protein;Coatomer subunit beta'-3;Myb-like domain-containing protein;Protein kinase domain-containing protein;OTU domain-containing protein;Lipase_GDSL domain-containing protein;Putative coatomer beta subunit family protein;Homeodomain-like superfamily protein (Trihelix transcription factor)	CRdw	miRNA	H_novel
M-321	UUUUGUUGGUUGUCAUCUAACC	miR827	2B,2D		XLOC_031123(0.0),XLOC_055067(1.5),XLOC_128210(1.5),XLOC_095885(1.5),XLOC_048298(1.5),XLOC_057939(1.5),XLOC_009391(2.0),XLOC_009083(2.0),XLOC_008029(2.0),XLOC_025471(2.0),XLOC_028404(2.0),XLOC_063451(2.0),XLOC_143427(2.0),XLOC_109013(2.0),XLOC_092724(2.0),XLOC_090977(2.0),XLOC_044622(2.0),XLOC_058085(2.0),XLOC_041013(2.0)	Methyltransferase-like protein;Protein kinase domain-containing protein;Adenosinetriphosphatase;Histone acetyltransferase;Putative disease resistance protein RGA1;Receptor-like serine/threonine-protein kinase;Os01g0694200 protein;Os04g0485800 protein (cDNA clone:J023077K22, full insert sequence); (Fragment);F-box domain-containing protein	CRup	miRNA	B_MiRBaseMonocot
M-322	UUUUUGUUUCCGCCGUAGUUG	noHit	4B		XLOC_118938(1.5),XLOC_109893(1.5),XLOC_042456(1.5),XLOC_004262(2.0),XLOC_020728(2.0),XLOC_023902(2.0),XLOC_134635(2.0),XLOC_150584(2.0),XLOC_150826(2.0),XLOC_131938(2.0),XLOC_122420(2.0),XLOC_052574(2.0),XLOC_033677(2.0)	Nucleic acid binding protein;Flowering locus K homology domain;Os02g0196000 protein;Protein kinase domain-containing protein;Methyltranfer_dom domain-containing protein;Disease resistance protein RGA2;Zinc transporter ZTP29;NB-ARC domain-containing protein;F-box domain-containing protein	Cstv	miRNA	F_PSUSStaeMirSir

S-007	AAAUCCGUAAUUUUUGGCACC	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_081366(1.0),XLOC_088588(1.0),XLOC_071855(1.0),XLOC_139134(1.5),XLOC_130378(1.5),XLOC_102969(1.5),XLOC_009238(2.0),XLOC_063532(2.0),XLOC_067501(2.0),XLOC_087395(2.0),XLOC_138350(2.0),XLOC_139895(2.0),XLOC_148653(2.0),XLOC_146803(2.0),XLOC_149916(2.0),XLOC_130227(2.0),XLOC_114627(2.0),XLOC_047372(2.0),XLOC_072022(2.0),XLOC_072874(2.0),XLOC_069829(2.0)	Cysteine-rich receptor-like protein kinase 15;GUN4-like family protein, expressed (Os11g0267000 protein) (cDNA clone:006-204-B07, full insert sequence) (cDNA clone:006-302-H05, full insert sequence) (cDNA clone:006-306-B02, full insert sequence) (cDNA clone:J023122C05, full insert sequence);Protein kinase domain-containing protein;GUN4 domain-containing protein;Glutaredoxin domain-containing protein;Calcyclin-binding protein;Cysteine-rich receptor-like protein kinase 26;Cysteine-rich receptor-like protein kinase 7;Putative late blight resistance protein-like protein R1C-3;Disease resistance protein RPM1;C2H2-type domain-containing protein;Disease resistance RPP13-like protein 4;Os01g0757500 protein (Fragment)	Cstv	siRNA	F_PSUSStaeMirSir
S-016	CGAGCACCACCUCUCAGGCUG	noHit	6D	siRNA,TAS	XLOC_126981(0.0),XLOC_097522(2.0)	NAC domain-containing protein;NAC domain-containing protein 42	Cstv	siRNA	F_PSUSStaeMirSir
S-019	AUAGAUGCCGGUGUCAACUAC	noHit	1B,1D	mRNA;	XLOC_009315(1.0),XLOC_003064(1.0),XLOC_014001(1.0)	Golgi SNAP receptor complex member 1;Subtilisin-like protease SDD1	Cstv	siRNA	F_PSUSStaeMirSir
S-020	CCCGGCAUCCCUCUUGAGCCC	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;			Cstv	siRNA	F_PSUSStaeMirSir
S-021	UGCUCCAUAUCAUGUACUGUA	noHit	6D		XLOC_128777(0.0),XLOC_129114(0.0),XLOC_128038(1.0),XLOC_086689(1.5),XLOC_138720(1.5)	Protein kinase domain-containing protein;Protein kinase superfamily protein;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	F_PSUSStaeMirSir
S-023	GAAGAGCGCGUGGCGAGGGAA	noHit	6B		XLOC_120603(0.0),XLOC_080897(1.5),XLOC_144265(2.0),XLOC_145760(2.0),XLOC_099981(2.0)	Chromo domain-containing protein;Chromdomain-containing protein CRD101 (Chromo domain-containing protein LHP1);ChSh domain-containing protein;BHLH domain-containing protein;Transcription factor bHLH93;Chromo domain-containing protein LHP1;GRAS domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-024	AUGUUUUCGUCUAGCAAUAUA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_145818(1.5),XLOC_139385(1.5),XLOC_130242(1.5),XLOC_130378(1.5),XLOC_105814(1.5),XLOC_105140(1.5),XLOC_096247(1.5),XLOC_100789(1.5),XLOC_092844(1.5),XLOC_006347(2.0),XLOC_139134(2.0),XLOC_146456(2.0),XLOC_144697(2.0),XLOC_122197(2.0),XLOC_034067(2.0),XLOC_043365(2.0),XLOC_034390(2.0),XLOC_071750(2.0)	Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;GRAS domain-containing protein;Cysteine-rich receptor-like protein kinase 41;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;zinc_ribbon_12 domain-containing protein	Cstv	siRNA	H_novel
S-026	UUAGACCUAGACAUGCAAGUA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0)	Protein kinase domain-containing protein	Cstv	siRNA	E_LiteratureTaeMiRNA 2020
S-029	CACCAUGUUUUCGUCUAGCAA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_043365(1.0),XLOC_034390(1.0)	Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26	CRup	siRNA	H_novel
S-033	GCUAUGGAAUGAUCAAGGAAU	noHit	7B	mRNA	XLOC_087543(1.5),XLOC_146342(1.5),XLOC_074081(1.5),XLOC_081366(2.0),XLOC_088588(2.0),XLOC_139134(2.0),XLOC_122717(2.0),XLOC_071855(2.0)	Cysteine-rich receptor-like protein kinase 15;DDT domain-containing protein;Protein kinase domain-containing protein;2-succinylbenzoate--CoA ligase;AMP-binding protein (Acetate/butyrate--CoA ligase AAE7 peroxisomal);AMP-binding enzyme family protein, expressed (Os03g0305100 protein)	Cstv	siRNA	H_novel
S-040	UCAAAGCAUCUUGACAGACCA	noHit	7B,7D	mRNA;Protein kinase domain-containing protein	XLOC_146342(0.0),XLOC_133900(1.5),XLOC_142630(1.5),XLOC_150455(1.5),XLOC_021130(2.0),XLOC_105582(2.0),XLOC_091135(2.0),XLOC_099608(2.0),XLOC_039035(2.0)	Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;RRM domain-containing protein;NB-ARC domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-041	UGCUAUGGAAUGAUCAAGGAA	noHit	7B	mRNA	XLOC_087543(0.5),XLOC_081366(1.5),XLOC_088588(1.5),XLOC_139134(1.5),XLOC_074081(1.5),XLOC_071855(1.5),XLOC_138261(2.0),XLOC_146342(2.0),XLOC_122717(2.0),XLOC_130378(2.0)	Cysteine-rich receptor-like protein kinase 15;DDT domain-containing protein;Protein kinase domain-containing protein;2-succinylbenzoate--CoA ligase;AMP-binding protein (Acetate/butyrate--CoA ligase AAE7 peroxisomal);AMP-binding enzyme family protein, expressed (Os03g0305100 protein)	Cstv	siRNA	H_novel
S-049	UCAAAGCAUCUUGGCAGACCA	noHit	7A	mRNAProtein kinase domain-containing protein;	XLOC_146342(0.5),XLOC_055924(2.0),XLOC_133900(2.0),XLOC_142630(2.0),XLOC_150455(2.0),XLOC_121066(2.0),XLOC_094389(2.0),XLOC_070069(2.0)	H_novelomogentisate geranylgeranyltransferase (HGGT);Protein kinase domain-containing protein;Homogentisate geranylgeranyltransferase;Translation factor GUF1 homolog, mitochondrial;Homogentisate geranylgeranyl transferase1;Digeranylgeranylglyceryl phosphate synthase;RRM domain-containing protein;Tr-type G domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-053	CACCCGCCGACGCCGGCUGCU	noHit	1A				Cstv	siRNA	F_PSUSStaeMirSir
S-055	AUAUUUGCAGGUUUUAGGUCU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_134508(1.5),XLOC_106062(1.5),XLOC_110271(1.5),XLOC_072262(1.5),XLOC_079925(2.0),XLOC_134491(2.0),XLOC_133919(2.0),XLOC_150286(2.0),XLOC_128866(2.0),XLOC_128299(2.0),XLOC_127888(2.0),XLOC_122197(2.0),XLOC_117801(2.0),XLOC_121992(2.0),XLOC_115725(2.0),XLOC_121085(2.0),XLOC_118923(2.0),XLOC_057401(2.0),XLOC_033068(2.0),XLOC_072674(2.0),XLOC_070884(2.0)	Cysteine-rich receptor-like protein kinase 15;SET domain-containing protein;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 19;Os02g0804500 protein;Cysteine-rich receptor-like protein kinase 26;E3 ubiquitin ligase (RING-type E3 ubiquitin transferase) (Serine/threonine-protein kinase);1,3-beta-glucan synthase;Putative WD40-like beta propeller repeat family protein;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Aldehyde oxygenase (deformylating)	Cstv	siRNA	D_LiteratureTaeMiRNA2015
S-056	CCAUGUUUUCGUCUAGCAAUA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_145818(1.5),XLOC_130242(1.5),XLOC_105140(1.5),XLOC_105814(1.5),XLOC_096247(1.5),XLOC_100789(1.5),XLOC_092844(1.5),XLOC_006347(2.0),XLOC_146456(2.0),XLOC_144697(2.0),XLOC_150350(2.0),XLOC_034067(2.0),XLOC_043365(2.0),XLOC_034390(2.0)	Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;GRAS domain-containing protein;Cysteine-rich receptor-like protein kinase 41;Disease resistance protein RPM1;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Disease resistance RPP13-like protein 4;zinc_ribbon_12 domain-containing protein	Cstv	siRNA	H_novel

S-060	CCAACGAUCGGUGACAGAGAA	noHit	2D	siRNA,TAS	XLOC_140387(2.0)		Cstv	siRNA	F_PSUSStaeMirSir
S-063	AUGGCACAAGUUUGCGAGGAC	noHit	6D		XLOC_129568(0.0),XLOC_129421(0.0),XLOC_030899(2.0),XLOC_054754(2.0),XLOC_055030(2.0),XLOC_084141(2.0),XLOC_105177(2.0),XLOC_099198(2.0),XLOC_091151(2.0),XLOC_044545(2.0),XLOC_078459(2.0)	Putative disease resistance protein RGA3;NB-ARC domain-containing protein;Disease resistance protein RPM1;SH3 domain-containing protein;DUF4220 domain-containing protein	Cstv	siRNA	H_novel
S-080	UGUAGACGAGCUCAGGGAGGA	noHit	6A,6B,6D	TAS,siRNA;	XLOC_120603(0.0),XLOC_011728(2.0),XLOC_055473(2.0),XLOC_057931(2.0),XLOC_075104(2.0)	Structural maintenance of chromosomes protein 2-2;Protein kinase domain-containing protein;SMC hinge domain-containing protein;Receptor-like protein kinase HSL1;3-hydroxyacyl-CoA dehydrogenase;Structural maintenance of chromosomes protein	Cstv	siRNA	G_BlastnPreMirSir
S-081	UUUGCGAAGGUGUUUUGAGAG	noHit	6A	siRNA,TAS	XLOC_088588(2.0),XLOC_085917(2.0)	Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Reverse transcriptase domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-082	UCCCCGAGCUCAAAUGACCCU	noHit	6D		XLOC_040446(2.0)	Protein kinase domain-containing protein;Formin-like protein 6	Cstv	siRNA	F_PSUSStaeMirSir
S-084	CCACACUCAAGAAUUUCUUGG	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0),XLOC_002751(1.5)	Reticulon-like protein;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-086	UCUUUCUUCGCCGUCCACCUC	noHit	6D	siRNA,TAS	XLOC_126981(0.0),XLOC_085620(1.5),XLOC_138740(1.5),XLOC_145226(1.5),XLOC_132021(1.5),XLOC_114162(1.5),XLOC_060160(1.5),XLOC_074566(1.5),XLOC_055690(2.0),XLOC_054833(2.0),XLOC_127022(2.0),XLOC_111028(2.0)	Myb-related protein P;SKP1-like protein 11;MYB transcription factor 78;MYB transcription factor SM155-1;MYB73;Expressed protein (Os12g0244200 protein) (cDNA clone:J023002G14, full insert sequence);Myb-related protein Myb4	Cstv	siRNA	G_BlastnPreMirSir
S-087	CCCUCGCACCGCGCUCUUCUU	noHit	6D	siRNA,TAS	XLOC_126981(0.0),XLOC_114162(1.5),XLOC_009127(2.0)	Transcription factor GTE1	Cstv	siRNA	G_BlastnPreMirSir
S-113	ACAUCAAGUGAUCGGACAUGA	noHit	6B	mRNA	XLOC_124504(0.0),XLOC_124282(0.0),XLOC_119081(0.0),XLOC_113448(0.0),XLOC_120281(0.0),XLOC_113331(1.0),XLOC_118389(1.5),XLOC_107834(1.5),XLOC_125876(2.0),XLOC_118267(2.0)	Putative late blight resistance protein-like protein R1B-16;AAA domain-containing protein;NB-ARC domain-containing protein;Disease resistance RPP13-like protein 4	Cstv	siRNA	F_PSUSStaeMirSir
S-114	ACCGAUGAUGACUCCCUCGCC	noHit	1B,1D	mRNA;			Cstv	siRNA	F_PSUSStaeMirSir
S-129	CACCGUCAUCCUCCCGUCGCC	noHit	6B	siRNA,TAS	XLOC_028047(0.5),XLOC_054624(2.0),XLOC_086663(2.0),XLOC_087696(2.0),XLOC_149065(2.0),XLOC_128482(2.0),XLOC_069811(2.0)	6(G)-fructosyltransferase;MADS-box transcription factor 29;Protein kinase domain-containing protein;MADS-box domain-containing protein;Os07g0209100 protein;Putative galactinol--sucrose galactosyltransferase 2	Cstv	siRNA	G_BlastnPreMirSir
S-130	GUCUUUCUUCGCCGCCCACCU	noHit	6B	siRNA,TAS	XLOC_126981(1.0),XLOC_097385(1.5),XLOC_036963(1.5),XLOC_138740(2.0),XLOC_133589(2.0),XLOC_145226(2.0),XLOC_132021(2.0),XLOC_127022(2.0),XLOC_101067(2.0)	Transcription factor RAX3;UDP-glycosyltransferase 73D1;Os09g0460300 protein (cDNA clone:J023102G04, full insert sequence);Glutaredoxin domain-containing protein;Abhydrolase_3 domain-containing protein;Glycosyltransferase;MYB73;Expressed protein (Os12g0244200 protein) (cDNA clone:J023002G14, full insert sequence);Myb-related protein Myb4;MYB transcription factor SM155-1;Os07g0159900 protein	Cstv	siRNA	G_BlastnPreMirSir
S-131	GUAGCCGGGAUAGACGAAGCU	noHit	1A		XLOC_055365(2.0)	Enhancer of mRNA-decapping protein 4;WD_REPEATS_REGION domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-145	CCAUUGUUGCACUAACCGCUG	noHit	2A	siRNA,TAS	XLOC_143130(1.5),XLOC_135320(1.5),XLOC_150667(1.5)	ADP-ribosylation factor GTPase-activating protein AGD1;Discolored-paralog2	Cstv	siRNA	F_PSUSStaeMirSir
S-146	CACAACAUGUGUACACUUCGU	noHit	1A	mRNA;	XLOC_007878(2.0)		Cstv	siRNA	H_novel
S-154	UUCAGCAUAAGAUCCGUAGGC	noHit	3D	mRNA;Os01g0721900 protein;Ankyrin repeat domain-containing protein 2A;WRKY domain-containing protein;Os03g0321700 protein (WRKY DNA binding domain containing protein, expressed) (cDNA clone:J033057D07, full insert sequence);GIY-YIG domain-containing protein;Mitochondrial outer membrane protein porin 1 (Voltage-dependent anion-selective channel protein 1) (OsVDAC1);Ankyrin repeat domain-containing protein 2;Disease resistance protein RPM1;Chorismate_bind domain-containing protein;Putative WRKY transcription factor 46;HABP4_PAI-RBP1 domain-containing protein;Mitochondrial	XLOC_053848(2.0)	malic domain-containing protein;Malic enzyme	Cstv	siRNA	F_PSUSStaeMirSir
S-157	GGAUGGUUUUGAAGUGAUAGC	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_047121(1.5),XLOC_068730(2.0),XLOC_106190(2.0),XLOC_098776(2.0),XLOC_093041(2.0),XLOC_047499(2.0)	Protein-serine/threonine phosphatase;Protein kinase domain-containing protein;Cellulose synthase;Os01g0708900 protein (cDNA clone:J013050L21, full insert sequence);Graves disease carrier protein	Cstv	siRNA	H_novel

S-158	UUCAAAGCAUCUUGACAGACC	noHit	7B,7D	mRNA;Protein kinase domain-containing protein	XLOC_146342(0.0),XLOC_063823(2.0),XLOC_055924(2.0),XLOC_121066(2.0)	H_novelomogentisate geranylgeranyltransferase (HGGT);Transcription factor GTE3, chloroplastic;Bromo domain-containing protein;Protein kinase domain-containing protein;Homogentisate geranylgeranyltransferase;Homogentisate geranylgeranyl transferase1;Digeranylgeranylglyceryl phosphate synthase	CRup	siRNA	F_PSUSStaeMirSir
S-159	UCUUGACAGACCAAAAUCCGU	noHit	7B,7D	mRNA;Protein kinase domain-containing protein	XLOC_144146(0.0),XLOC_146342(0.0),XLOC_138664(1.0),XLOC_121109(1.0),XLOC_103587(1.0),XLOC_075407(1.0),XLOC_006347(1.5),XLOC_008378(1.5),XLOC_016805(1.5),XLOC_023975(1.5),XLOC_137941(1.5),XLOC_134491(1.5),XLOC_133622(1.5),XLOC_148646(1.5),XLOC_144697(1.5),XLOC_146456(1.5),XLOC_128299(1.5),XLOC_117801(1.5),XLOC_119331(1.5),XLOC_121255(1.5),XLOC_122362(1.5),XLOC_122163(1.5),XLOC_105582(1.5),XLOC_099608(1.5),XLOC_092993(1.5),XLOC_070858(1.5),XLOC_072674(1.5),XLOC_072568(1.5),XLOC_012089(2.0),XLOC_011031(2.0),XLOC_058747(2.0),XLOC_138139(2.0),XLOC_139385(2.0),XLOC_140294(2.0),XLOC_134268(2.0),XLOC_145285(2.0),XLOC_114627(2.0),XLOC_110928(2.0),XLOC_120925(2.0),XLOC_114970(2.0),XLOC_106062(2.0),XLOC_105893(2.0),XLOC_041841(2.0),XLOC_033068(2.0),XLOC_070021(2.0)	Putative cysteine-rich receptor-like protein kinase 12;Histidine kinase;Protein kinase domain-containing protein (Fragment);Os11g0274100 protein (Fragment);Protein-serine/threonine phosphatase;Os04g0369066 protein;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 29;RING-type E3 ubiquitin transferase;Cysteine-rich receptor-like protein kinase 41;NB-ARC domain-containing protein;Transcription initiation factor TFIID subunit 4b;Putative WD40-like beta propeller repeat family protein;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Cysteine-rich receptor-like protein kinase 19;Os11g0608300 protein (cDNA, clone: J100045H04, full insert sequence)	Cstv	siRNA	H_novel
S-160	CGUAAUUUUUGGCACCAUGUU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_094765(1.5),XLOC_003942(2.0),XLOC_016828(2.0),XLOC_012116(2.0),XLOC_141488(2.0),XLOC_128556(2.0),XLOC_115567(2.0),XLOC_095239(2.0),XLOC_033068(2.0),XLOC_078255(2.0)	Myb_DNA-bind_3 domain-containing protein;DDE Tnp4 domain-containing protein;Os09g0122200 protein (Fragment);Os07g0115800 protein (Fragment);Protein kinase domain-containing protein;Os10g0468250 protein;Os09g0122100 protein;Glyoxalase I;Os03g0643050 protein (Fragment);Os03g0643100 protein (Fragment);WAC domain-containing protein;SWIM-type domain-containing protein (Fragment);Os07g0116050 protein;Nuclear pore complex protein NUP205;Lactoylglutathione lyase;Retrovirus-related Pol polyprotein LINE-1	Cstv	siRNA	H_novel
S-161	ACCAUGUUUUCGUCUAGCAAU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_043365(1.0),XLOC_034390(1.0),XLOC_019978(1.5),XLOC_145818(1.5),XLOC_130242(1.5),XLOC_132224(1.5),XLOC_105140(1.5),XLOC_105814(1.5),XLOC_096247(1.5),XLOC_100789(1.5),XLOC_092844(1.5),XLOC_006347(2.0),XLOC_140024(2.0),XLOC_146456(2.0),XLOC_144697(2.0),XLOC_114970(2.0),XLOC_034067(2.0)	ULP_PROTEASE domain-containing protein;WD_REPEATS_REGION domain-containing protein;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;GRAS domain-containing protein;Cysteine-rich receptor-like protein kinase 41;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Histidine kinase;zinc_ribbon_12 domain-containing protein	Cstv	siRNA	H_novel
S-162	GUUUUCGUCUAGCAAUAUAUU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_006347(0.5),XLOC_128874(1.0),XLOC_130242(1.5),XLOC_011424(2.0),XLOC_012789(2.0),XLOC_032270(2.0),XLOC_025207(2.0),XLOC_054768(2.0),XLOC_145818(2.0),XLOC_137377(2.0),XLOC_144697(2.0),XLOC_146456(2.0),XLOC_130378(2.0),XLOC_122751(2.0),XLOC_110077(2.0),XLOC_114493(2.0),XLOC_115363(2.0),XLOC_122549(2.0),XLOC_115422(2.0),XLOC_096247(2.0),XLOC_037198(2.0),XLOC_044266(2.0),XLOC_071750(2.0)	Pto-interacting protein 1;Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;Wall-associated receptor kinase 5;Os08g0278900 protein;Cysteine-rich receptor-like protein kinase 41;G-type lectin S-receptor-like serine/threonine-protein kinase SRK	Cstv	siRNA	H_novel
S-163	UAUAUUUGCAGGUUUUAGGUC	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_106062(1.5),XLOC_110271(1.5),XLOC_134508(2.0),XLOC_128299(2.0),XLOC_122197(2.0),XLOC_033068(2.0)	Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 26;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Aldehyde oxygenase (deformylating)	Cstv	siRNA	E_LiteratureTaeMiRNA 2020
S-164	CUAUGGAAUGAUCAAGGAAUU	noHit	7B	mRNA	XLOC_087543(1.5),XLOC_146342(1.5),XLOC_088512(2.0),XLOC_126080(2.0),XLOC_116792(2.0),XLOC_118327(2.0)	Cytosolic invertase 1 (OsCYT-INV1);Protein kinase domain-containing protein;Alkaline/neutral invertase;C2H2-type domain-containing protein	Cstv	siRNA	H_novel
S-165	UUCACUUAGACCUAGACAUGC	noHit	7A,7B	mRNA;Protein kinase domain-containing protein;	XLOC_146342(1.5),XLOC_012108(2.0)	DBD_Tnp_Mut domain-containing protein;Protein kinase domain-containing protein	Cstv	siRNA	H_novel
S-166	ACUUAGACCUAGACAUGCAAG	noHit	7A,7B	mRNA;Protein kinase domain-containing protein;	XLOC_146342(1.5)	Protein kinase domain-containing protein	Cstv	siRNA	H_novel
S-167	UCUAGGCGAAAAUAUAAUGUU	noHit	7B	mRNA	XLOC_146342(1.0),XLOC_017845(2.0),XLOC_062688(2.0)	Protein kinase domain-containing protein;Chromatin structure-remodeling complex protein BSH	Cstv	siRNA	H_novel

S-168	CCUUCUUUUCCUUGCCAUGUU	noHit	4A	siRNA,TAS,TASiRNA,SiRNA, mRNA	XLOC_081194(1.0),XLOC_087935(1.0),XLOC_072629(1.0),XLOC_083679(1.5),XLOC_125627(1.5),XLOC_075142(1.5),XLOC_015420(2.0),XLOC_021763(2.0),XLOC_030520(2.0),XLOC_062535(2.0),XLOC_058605(2.0),XLOC_086016(2.0),XLOC_085722(2.0),XLOC_080563(2.0),XLOC_113637(2.0),XLOC_105642(2.0),XLOC_090036(2.0),XLOC_099551(2.0),XLOC_048009(2.0),XLOC_048003(2.0),XLOC_039189(2.0),XLOC_034922(2.0),XLOC_078297(2.0),XLOC_070461(2.0)	Pre-mRNA-processing factor 17;WD_REPEATS_REGION domain-containing protein;Rho-GAP domain-containing protein;eRF1_1 domain-containing protein;Eukaryotic peptide chain release factor subunit 1-3;Putative mediator of RNA polymerase II transcription subunit 12;Med12 domain-containing protein;Serine/threonine-protein phosphatase BSL3;Phospholipid-transporting ATPase;BAG family molecular chaperone regulator 4 (BCL-2 binding anthanogene-1);EIP1 (Os01g0831200 protein) (cDNA clone:J023047C12, full insert sequence);Os03g0216800 protein (Polygalacturonase, putative, expressed) (cDNA clone:001-127-F09, full insert sequence);Serine/threonine-protein phosphatase;BAG family molecular chaperone regulator 4;Transducin/WD40 repeat-like superfamily protein;Os07g0648300 protein (Fragment);Os11g0153400 protein (Rac GTPase activating protein 1, putative, expressed);Alkaline/neutral invertase;Os03g0397500 protein (Pre-mRNA splicing factor PRP17, putative, expressed) (cDNA clone:J033044C19, full insert sequence);Rho GTPase-activating protein gacA;Putative eukaryotic peptide chain release factor subunit 1-3;RTR domain-containing protein;Eukaryotic	Cstv	siRNA	F_PSUSStaeMirSir
S-170	AGACCUAGACAUGCAAGUAUA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_023881(2.0)	Protein kinase domain-containing protein	Cstv	siRNA	H_novel
S-173	UUCGUCUAGCAAUAUAUUUGC	noHit	7B,7D	mRNA;Protein kinase domain-containing protein	XLOC_146342(0.0),XLOC_006347(1.0),XLOC_130242(1.0),XLOC_008828(1.5),XLOC_010495(1.5),XLOC_030502(1.5),XLOC_128874(1.5),XLOC_130580(1.5),XLOC_037198(1.5),XLOC_042923(1.5),XLOC_011424(2.0),XLOC_139167(2.0),XLOC_144697(2.0),XLOC_146456(2.0),XLOC_145565(2.0),XLOC_104894(2.0),XLOC_099259(2.0),XLOC_071750(2.0)	Pto-interacting protein 1;Cysteine-rich receptor-like protein kinase 15;Serine/threonine protein kinase;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 7;Os02g0819600 protein;DUF295 domain-containing protein;Cysteine-rich receptor-like protein kinase 26;DUF668 domain-containing protein;Cysteine-rich receptor-like protein kinase 41;Receptor-like serine/threonine-protein kinase SD1-6;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Cysteine-rich receptor-like protein kinase 6	Cstv	siRNA	H_novel
S-174	UCACGUGGACAUCAGUCCAUC	noHit	1A				Cstv	siRNA	F_PSUSStaeMirSir
S-175	AAGAAAUUCUUGAGUGUGGGU	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_119971(1.0),XLOC_002923(2.0),XLOC_055732(2.0),XLOC_062920(2.0),XLOC_057728(2.0),XLOC_048574(2.0),XLOC_035161(2.0)	ATP-dependent DNA helicase;Protein SRG1;Fe2OG dioxygenase domain-containing protein;Putative ABC transporter B family member 8;Os03g0856000 protein (Oxidoreductase, 2OG-Fe oxygenase family protein, expressed) (cDNA clone:J023105N19, full insert sequence);CBFD_NFYB_HMF domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-176	AGAGCAGUUGGGACUGCCCGC	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0),XLOC_043684(2.0)	Defective in cullin neddylation protein;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-178	ACCAAUGAUAACUCCCUCGCC	noHit	1B	mRNA			Cstv	siRNA	F_PSUSStaeMirSir
S-181	UUUCGUCUAGCAAUAUAUUUG	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_006347(0.5),XLOC_128874(0.5),XLOC_130242(1.0),XLOC_012789(1.5),XLOC_110077(1.5),XLOC_011424(2.0),XLOC_010495(2.0),XLOC_008828(2.0),XLOC_030502(2.0),XLOC_032270(2.0),XLOC_025207(2.0),XLOC_026378(2.0),XLOC_054768(2.0),XLOC_139167(2.0),XLOC_145818(2.0),XLOC_144697(2.0),XLOC_146456(2.0),XLOC_148306(2.0),XLOC_130580(2.0),XLOC_122751(2.0),XLOC_114493(2.0),XLOC_115363(2.0),XLOC_122549(2.0),XLOC_115422(2.0),XLOC_096247(2.0),XLOC_037198(2.0),XLOC_044266(2.0),XLOC_042939(2.0),XLOC_071750(2.0)	Pto-interacting protein 1;Cysteine-rich receptor-like protein kinase 15;Serine/threonine protein kinase;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 7;Os02g0819600 protein;DUF295 domain-containing protein;Cysteine-rich receptor-like protein kinase 26;Wall-associated receptor kinase 5;Cysteine-rich receptor-like protein kinase 41;G-type lectin S-receptor-like serine/threonine-protein kinase SRK;Cysteine-rich receptor-like protein kinase 6	Cstv	siRNA	H_novel
S-182	UGUUAAAAAUUGCGGAUUUUU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_081366(1.5),XLOC_088588(1.5),XLOC_139134(1.5),XLOC_130378(1.5),XLOC_071855(1.5),XLOC_087255(2.0)	Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Reverse transcriptase domain-containing protein	Cstv	siRNA	H_novel
S-184	AGAGAGGGUUGAUGCGUCGAA	noHit	6D	siRNA,TAS	XLOC_103592(1.5),XLOC_126432(2.0)	SHSP domain-containing protein;18.6 kDa class III heat shock protein (18.6 kDa heat shock protein) (OsHsp18.6);17.4 kDa class III heat shock protein	Cstv	siRNA	F_PSUSStaeMirSir
S-190	AGAGGAUCUUCGCCUGAUGGA	noHit	4A		XLOC_073324(0.0),XLOC_130357(1.5),XLOC_026946(2.0)	TAXi_C domain-containing protein;Aspartic proteinase nepenthesin-1;Aspartic proteinase nepenthesin-2;Peptidase A1 domain-containing protein	CRdw	siRNA	F_PSUSStaeMirSir
S-191	GGCACCGGCAGUAGUGGCACC	noHit	4A		XLOC_087806(1.5),XLOC_106564(1.5),XLOC_089841(1.5),XLOC_006724(2.0),XLOC_023026(2.0),XLOC_056281(2.0),XLOC_058764(2.0),XLOC_143795(2.0),XLOC_099082(2.0),XLOC_093774(2.0),XLOC_048375(2.0),XLOC_078898(2.0)	Amb_all domain-containing protein;PHD-type domain-containing protein;Os09g0298400 protein;WD_REPEATS_REGION domain-containing protein;N-acetyltransferase domain-containing protein;Membrane protein;Os04g0594500 protein;DUF674 family protein; (Fragment);RRM domain-containing protein;CCHC-type domain-containing protein;Auxin-responsive protein;Os01g0962200 protein (cDNA clone:J013114K13, full insert sequence);Auxin-responsive protein IAA13 (Indoleacetic acid-induced protein 13)	Cstv	siRNA	H_novel

S-194	CGAGAAAGGAGUGGAACACUU	noHit	6D		XLOC_129114(0.0),XLOC_033560(0.5),XLOC_128038(1.0),XLOC_004712(1.5),XLOC_004711(1.5),XLOC_085224(1.5),XLOC_035290(1.5),XLOC_028464(2.0),XLOC_025119(2.0),XLOC_120505(2.0),XLOC_042572(2.0)	Cytochrome P450 family 96 subfamily A polypeptide 1;EF-hand domain-containing protein;Reverse transcriptase domain-containing protein;Cytochrome P450 86B1;Os05g0181200 protein (Fragment);NB-ARC domain-containing protein;Disease resistance protein RPM1;CCHC-type domain-containing protein	Cstv	siRNA	H_novel
S-195	GACGAGAAAGGAGUGGAACAC	noHit	6D		XLOC_129114(0.0),XLOC_128038(0.0),XLOC_085224(2.0),XLOC_129697(2.0),XLOC_093827(2.0),XLOC_051999(2.0),XLOC_033560(2.0),XLOC_035290(2.0),XLOC_073876(2.0)	EF-hand domain-containing protein;Protein DETOXIFICATION (Multidrug and toxic compound extrusion protein);ThiF domain-containing protein;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	F_PSUSStaeMirSir
S-217	AGCUGCUCGUUCAUGGUUCCC	miR159	3B	miRNA,microRNA,MIR,mRNA	XLOC_061039(0.0),XLOC_068598(1.5),XLOC_146060(1.5),XLOC_130373(1.5),XLOC_128391(1.5),XLOC_052503(1.5),XLOC_145885(2.0),XLOC_137251(2.0)	Os06g0716800 protein (cDNA clone:001-016-A09, full insert sequence) (cDNA clone:J023099K10, full insert sequence);1-phosphatidylinositol-3-phosphate 5-kinase;F-box domain-containing protein;BZIP domain-containing protein	Cstv	siRNA	B_MiRBaseMonocot
S-218	UGAACCCUUCCACUUUGAUGG	noHit	1B,1D	mRNA;	XLOC_009544(0.0),XLOC_009315(0.0),XLOC_003064(0.0),XLOC_014001(0.0),XLOC_002046(1.5),XLOC_002434(1.5),XLOC_001632(1.5),XLOC_016582(1.5),XLOC_015132(1.5),XLOC_050539(1.5),XLOC_004490(2.0),XLOC_017079(2.0)	JmjC domain-containing protein;Sulfoquinovosyl transferase SQD2;Os01g0142300 protein;Golgi SNAP receptor complex member 1;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-219	AGGGUUCAUGCCAUGGCACUG	noHit	1B,1D	mRNA;	XLOC_016652(0.0),XLOC_009893(1.5),XLOC_016636(1.5),XLOC_144740(1.5),XLOC_066871(2.0),XLOC_063452(2.0),XLOC_066998(2.0),XLOC_068032(2.0),XLOC_059663(2.0),XLOC_054489(2.0),XLOC_033030(2.0)	Plastocyanin-like domain-containing protein;Cytochrome P450 71C2;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-220	CUACACCGAUGAUGACUCCCU	noHit	1B,1D	mRNA;	XLOC_009315(1.5),XLOC_003064(1.5),XLOC_014001(1.5)	Golgi SNAP receptor complex member 1;Subtilisin-like protease SDD1	Cstv	siRNA	F_PSUSStaeMirSir
S-221	GAUAGAUGGAAGAAAGAUGGA	noHit	4A,4D	SiRNA,TAS,TASiRNA,mRNA,siRNA;Alkaline/neutral invertase;	XLOC_084749(0.0),XLOC_076710(0.0),XLOC_074878(0.0),XLOC_000203(2.0),XLOC_080647(2.0),XLOC_074388(2.0)	Protein kinase domain-containing protein	Cstv	siRNA	G_BlastnPreMirSir
S-226	UUCUUCAAAGCAUCUUGACAG	noHit	7B,7D	mRNA;Protein kinase domain-containing protein	XLOC_146342(0.0),XLOC_055924(1.0),XLOC_121066(1.0),XLOC_011821(1.5),XLOC_139104(1.5),XLOC_146803(1.5),XLOC_130227(1.5),XLOC_041418(2.0)	H_novelomogentisate geranylgeranyltransferase (HGGT);Protein kinase domain-containing protein;Homogentisate geranylgeranyltransferase;Homogentisate geranylgeranyl transferase1;Digeranylgeranylglyceryl phosphate synthase	Cstv	siRNA	H_novel
S-227	AGAGAGGGUUGGUGUGUCGAA	noHit	6B		XLOC_120603(0.0),XLOC_009276(2.0)		Cstv	siRNA	F_PSUSStaeMirSir
S-230	GCUCAAGAGGGAUGCCGGGUU	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0)	Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-231	GAAGAGCGCGGUGCGAGGGAA	noHit	6D	siRNA,TAS	XLOC_144514(1.5),XLOC_003029(2.0),XLOC_023489(2.0),XLOC_124642(2.0),XLOC_118258(2.0),XLOC_104365(2.0),XLOC_096658(2.0)	SPX domain-containing protein;Phospholipase D	Cstv	siRNA	F_PSUSStaeMirSir
S-236	AAUCCGUAAUUUUUGGCACCA	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_115567(1.0),XLOC_081366(2.0),XLOC_088588(2.0),XLOC_114627(2.0),XLOC_072022(2.0),XLOC_071855(2.0),XLOC_072874(2.0)	Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;Putative late blight resistance protein-like protein R1C-3;Nuclear pore complex protein NUP205;Disease resistance protein RPM1;Disease resistance RPP13-like protein 4;Retrovirus-related Pol polyprotein LINE-1	Cstv	siRNA	F_PSUSStaeMirSir
S-250	ACCAAAAUCCGUAAUUUUUGG	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_025542(1.0),XLOC_114627(1.0),XLOC_033728(1.0),XLOC_043780(1.0),XLOC_011005(1.5),XLOC_011091(1.5),XLOC_066038(1.5),XLOC_085974(1.5),XLOC_087428(1.5),XLOC_148646(1.5),XLOC_117801(1.5),XLOC_106698(1.5),XLOC_102969(1.5),XLOC_091173(1.5),XLOC_099629(1.5),XLOC_057147(1.5),XLOC_070858(1.5),XLOC_077347(1.5),XLOC_069177(1.5),XLOC_005520(2.0),XLOC_017256(2.0),XLOC_067501(2.0),XLOC_059996(2.0),XLOC_059552(2.0),XLOC_079825(2.0),XLOC_081366(2.0),XLOC_088588(2.0),XLOC_138261(2.0),XLOC_137941(2.0),XLOC_145285(2.0),XLOC_148691(2.0),XLOC_131573(2.0),XLOC_130378(2.0),XLOC_119331(2.0),XLOC_109830(2.0),XLOC_105582(2.0),XLOC_053695(2.0),XLOC_057108(2.0),XLOC_057109(2.0),XLOC_072560(2.0),XLOC_072571(2.0),XLOC_072022(2.0),XLOC_071855(2.0),XLOC_070021(2.0),XLOC_069885(2.0)	Protein-serine/threonine phosphatase;Putative late blight resistance protein-like protein R1C-3;Proline-rich receptor-like protein kinase PERK1;Mitogen-activated protein kinase kinase 3;Disease resistance RPP13-like protein 4;Os11g0274100 protein (Fragment);Coatomer subunit beta'-2;Cysteine-rich receptor-like protein kinase 26;MKK3-2;FYVE-type domain-containing protein;Cysteine-rich receptor-like protein kinase 19;MKK3-1;Putative CRINKLY4-like receptor protein kinase family protein;Cysteine-rich receptor-like protein kinase 10;Callose synthase 2;RING-type E3 ubiquitin transferase;Os11g0608300 protein (cDNA, clone: J100045H04, full insert sequence);Cysteine-rich receptor-like protein kinase 15;Protein kinase domain-containing protein;GUB_WAK_bind domain-containing protein;Mitogen-activated protein kinase kinase 6;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel
S-252	UUCUUCAAAGCAUCUUGGCAG	noHit	7A	mRNAProtein kinase domain-containing protein;	XLOC_055924(0.0),XLOC_121066(0.0),XLOC_146342(0.5),XLOC_011821(2.0),XLOC_063823(2.0),XLOC_139104(2.0),XLOC_146803(2.0),XLOC_130227(2.0),XLOC_041418(2.0)	H_novelomogentisate geranylgeranyltransferase (HGGT);Transcription factor GTE3, chloroplastic;Bromo domain-containing protein;Protein kinase domain-containing protein;Homogentisate geranylgeranyltransferase;Homogentisate geranylgeranyl transferase1;Digeranylgeranylglyceryl phosphate synthase	CRdw	siRNA	H_novel

S-253	UUCAAAGCAUCUUGGCAGACC	noHit	7A	mRNAProtein kinase domain-containing protein;	XLOC_146342(0.5),XLOC_055924(1.0),XLOC_121066(1.0),XLOC_126981(2.0),XLOC_078123(2.0)	H_novelomogentisate geranylgeranyltransferase (HGGT);Protein kinase domain-containing protein;Homogentisate geranylgeranyltransferase;Cold acclimation protein WCOR615;Group3 late embryogenesis abundant protein;Homogentisate geranylgeranyl transferase1;Digeranylgeranylglyceryl phosphate synthase;Cold-responsive LEA/RAB-related COR protein	Cstv	siRNA	H_novel
S-254	UAGACCUAGACAUGCAAGUAU	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0)	Protein kinase domain-containing protein	Cstv	siRNA	E_LiteratureTaeMiRNA2020
S-258	GAAUAAGAAGAUGAAGAACGU	noHit	4D	siRNA,TAS,TASiRNA,SiRNA,mRNAAlkaline/neutral invertase;	XLOC_084749(0.0),XLOC_011785(2.0),XLOC_031057(2.0),XLOC_136037(2.0),XLOC_104894(2.0),XLOC_099259(2.0)	Protein kinase domain-containing protein;DUF668 domain-containing protein;NB-ARC domain-containing protein;Receptor-like serine/threonine-protein kinase SD1-6	Cstv	siRNA	G_BlastnPreMirSir
S-265	CUCCCUGAGCUCGUCUACACC	noHit	6A,6B,6D	,TAS,siRNA;	XLOC_126981(0.0),XLOC_114162(0.0),XLOC_087569(2.0),XLOC_079735(2.0)	SANT domain-containing protein	Cstv	siRNA	G_BlastnPreMirSir
S-266	AUAGCUGCAGCUCUUCAAGGC	noHit	6D		XLOC_129114(0.0),XLOC_128038(0.0),XLOC_134249(2.0),XLOC_141743(2.0),XLOC_151058(2.0),XLOC_033560(2.0)	RNA helicase;NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel
S-274	CAUCUUUCUUCCAUCUAUCCC	noHit	4A,4D	SiRNA,TAS,TASiRNA,mRNA,siRNA;Alkaline/neutral invertase;	XLOC_082991(0.0),XLOC_078297(0.0),XLOC_015017(2.0),XLOC_045491(2.0),XLOC_049509(2.0),XLOC_035208(2.0),XLOC_076950(2.0)	Cytochrome P450 CYP714B3;Calcium/calmodulin-dependent serine/threonine-protein kinase 1;Cytochrome P450 714B2;Protein kinase domain-containing protein;Secologanin synthase;Pept_C1 domain-containing protein;Alkaline/neutral invertase	Cstv	siRNA	D_LiteratureTaeMiRNA2015
S-275	CCUCCUUUCCCUGGCCAUGUU	noHit	4D	siRNA,TAS,TASiRNA,SiRNA,mRNAAlkaline/neutral invertase;	XLOC_082991(0.0),XLOC_045491(0.0),XLOC_083136(2.0),XLOC_124804(2.0),XLOC_111678(2.0),XLOC_117640(2.0),XLOC_112474(2.0),XLOC_037551(2.0),XLOC_033641(2.0),XLOC_079113(2.0),XLOC_074827(2.0)	Os02g0618400 protein (cDNA clone:J013104N22, full insert sequence);VWFA domain-containing protein;TFIIS N-terminal domain-containing protein;Transcription factor MYB48;Alkaline/neutral invertase;26S proteasome non-ATPase regulatory subunit 4	Cstv	siRNA	G_BlastnPreMirSir
S-279	AGAUAGGGAAACUGACAUGCC	noHit	6D		XLOC_129568(0.0),XLOC_129421(0.0),XLOC_020341(2.0),XLOC_087439(2.0),XLOC_046033(2.0)	NB-ARC domain-containing protein;Disease resistance protein RPM1;DUF4371 domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-288	CUUAGACCUAGACAUGCAAGU	noHit	7A,7B	mRNA;Protein kinase domain-containing protein;	XLOC_146342(1.0),XLOC_130096(2.0)	Protein kinase domain-containing protein;Intron_maturas2 domain-containing protein;Reverse transcriptase domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-291	GCAAUGGAAGGAUGACGACGG	noHit	6D	siRNA,TAS	XLOC_027230(1.5),XLOC_118105(2.0),XLOC_036838(2.0)	Os05g0376400 protein (Fragment);Os02g0586400 protein (cDNA clone:006-207-A09, full insert sequence) (cDNA clone:J023004J08, full insert sequence);Ras-related protein RABB1a;OSJNBa0029H02.8 protein (Os04g0499100 protein)	Cstv	siRNA	F_PSUSStaeMirSir
S-292	GCCUGAGAGGUGGUGCUCGUA	noHit	6D	siRNA,TAS			Cstv	siRNA	G_BlastnPreMirSir
S-293	UUGCAAGGUCAAGAACAGCCU	noHit	6A,6D	TAS,siRNA;	XLOC_031900(0.0),XLOC_120603(0.5),XLOC_012899(2.0),XLOC_084749(2.0),XLOC_076710(2.0),XLOC_074878(2.0)	Branched-chain-amino-acid aminotransferase;	Cstv	siRNA	G_BlastnPreMirSir
S-297	GAGGAGGAAACGAAGGGGGAG	noHit	4D	siRNA,TAS,TASiRNA,SiRNA,mRNAAlkaline/neutral invertase;	XLOC_084749(0.0),XLOC_044089(1.5),XLOC_078496(1.5),XLOC_009857(2.0),XLOC_004986(2.0),XLOC_026323(2.0),XLOC_024585(2.0),XLOC_030166(2.0),XLOC_065575(2.0),XLOC_083512(2.0),XLOC_077514(2.0),XLOC_126289(2.0),XLOC_105610(2.0),XLOC_105015(2.0),XLOC_042250(2.0),XLOC_074878(2.0),XLOC_074137(2.0)	PseudoU_synth_2 domain-containing protein;AP2-EREBP-type transcription factor;NAB domain-containing protein;Os02g0234500 protein (cDNA clone:J013128H21, full insert sequence);Cytochrome b-c1 complex subunit Rieske, mitochondrial;Lipid transfer protein;Putative complex I intermediate-associated protein 30;RNA pseudourine synthase 6, chloroplastic;Os07g0557100 protein (cDNA clone:001-016-H11, full insert sequence) (cDNA clone:J033081M13, full insert sequence);Os01g0973300 protein (Fragment);AP-3 complex subunit beta;CCHC-type domain-containing protein;AAI domain-containing protein	Cstv	siRNA	G_BlastnPreMirSir
S-298	CACCCGCACAAACUCUGCCCU	noHit	1B	mRNA	XLOC_032290(2.0),XLOC_024422(2.0)		Cstv	siRNA	H_novel
S-299	CCUCGUGUCCAAGAGGGCAGA	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0),XLOC_132396(2.0)	Phospholipase D;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-300	ACCCGGCAUCCCUCUUGAGCC	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;			Cstv	siRNA	F_PSUSStaeMirSir
S-302	ACAAAAGACAAGUACUGCGCC	noHit	5A	mRNADisease resistance protein RPM1;	XLOC_058128(2.0)		Cstv	siRNA	F_PSUSStaeMirSir
S-305	UAGUAGGGUUUCCAUGUCUGU	noHit	1A		XLOC_016652(1.5)	Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-306	ACAAAAGCAUGAUCCGGUGGA	noHit	6D		XLOC_129568(0.0),XLOC_129421(0.0),XLOC_056369(2.0)	NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel
S-311	CCAGAUUACGUCCAGAUUGUU	noHit	1A,1B	mRNA;Transport inhibitor response 1;Transport inhibitor response 1-like protein Os05g0150500 (TIR1-like protein);Transp_inhibit domain-containing protein			Cstv	siRNA	F_PSUSStaeMirSir
S-314	AUGCAAGUAUAUUUCUAGGCG	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_022228(1.0),XLOC_030945(1.0),XLOC_040188(1.0),XLOC_057870(2.0)	Protein kinase domain-containing protein;Peroxisome biogenesis protein 6;Os01g0603100 protein (cDNA clone:J033046J06, full insert sequence)	Cstv	siRNA	H_novel

S-315	CGGCACCGGCAGUAGUGGCAC	noHit	4A		XLOC_106564(1.5),XLOC_023026(2.0),XLOC_024371(2.0),XLOC_081166(2.0),XLOC_087278(2.0),XLOC_114908(2.0),XLOC_099082(2.0),XLOC_093774(2.0),XLOC_048375(2.0),XLOC_078898(2.0)	GDSL esterase/lipase;Os09g0298400 protein;WD_REPEATS_REGION domain-containing protein;Membrane protein;GDSL-like Lipase/Acylhydrolase family protein, expressed (Os03g0683800 protein) (cDNA clone:J023013F22, full insert sequence);Lipase_GDSL domain-containing protein;Os01g0962200 protein (cDNA clone:J013114K13, full insert sequence);CCHC-type domain-containing protein	Cstv	siRNA	H_novel
S-316	AUGGAGAUCGCAGGCUGCUGC	noHit	4A		XLOC_073324(0.0)		CRdw	siRNA	H_novel
S-318	CCGAGCUCAAAUGGUCUACUU	noHit	1D	mRNA; (Fragment);NB-ARC domain-containing protein;Putative disease resistance RPP13-like protein 3	XLOC_018048(0.0),XLOC_146067(1.5),XLOC_004769(2.0),XLOC_011485(2.0),XLOC_005733(2.0),XLOC_017858(2.0),XLOC_012782(2.0),XLOC_139841(2.0),XLOC_110980(2.0)	Putative disease resistance RPP13-like protein 3;Putative disease resistance RPP13-like protein 2;Disease resistance protein RPP13; (Fragment);NB-ARC domain-containing protein;Glyco_transf_64 domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-324	UACUAUAUCCGGAUUCUGUAU	noHit	6B	mRNA	XLOC_118662(0.0),XLOC_117365(0.0),XLOC_113448(0.5),XLOC_028963(1.5),XLOC_118389(1.5),XLOC_113331(1.5),XLOC_124504(2.0),XLOC_124282(2.0)	Putative late blight resistance protein-like protein R1B-16;AAA domain-containing protein;NB-ARC domain-containing protein;Disease resistance RPP13-like protein 4	Cstv	siRNA	F_PSUSStaeMirSir
S-328	ACAGAGGAAAUGACUUGGGCU	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_009544(0.0),XLOC_015377(1.0),XLOC_094970(1.0),XLOC_001722(2.0),XLOC_092433(2.0)	tRNA (Cytosine(34)-C(5))-methyltransferase;SAM_MT_RSMB_NOP domain-containing protein;Os09g0471900 protein	Cstv	siRNA	H_novel
S-329	AUUAAGCUUCGAGAUUGUUGU	noHit	1B	mRNA	XLOC_016652(1.0),XLOC_104229(1.5),XLOC_127940(2.0)	PNPLA domain-containing protein;Peptidylprolyl isomerase;Patatin;Triacylglycerol lipase SDP1;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-330	UUGCAACAAGAUGAUUAAGCU	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0),XLOC_028303(2.0),XLOC_127281(2.0),XLOC_114934(2.0)	Chaperone protein ClpB 2;S-acyltransferase;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-331	CAUGCCAUGGCACUGCUGCUG	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_016652(0.0),XLOC_126007(2.0),XLOC_120265(2.0),XLOC_071789(2.0)	GATA transcription factor;GATA transcription factor 5;F-box domain-containing protein;GATA-type domain-containing protein;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-332	CAGUGCCAUGGCAUGAACCCU	noHit	1B,1D	mRNA;Subtilisin-like protease SDD1;	XLOC_009544(0.0),XLOC_009315(0.5),XLOC_003064(0.5),XLOC_014001(0.5),XLOC_002046(1.5),XLOC_002434(1.5),XLOC_001632(1.5),XLOC_016582(1.5),XLOC_015132(1.5),XLOC_127810(2.0),XLOC_121024(2.0),XLOC_115193(2.0)	Chlorophyll a-b binding protein, chloroplastic;Golgi SNAP receptor complex member 1;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-333	CAGCCGGCGUCGGCGGGUGGC	noHit	1A		XLOC_001578(0.0),XLOC_000695(1.5),XLOC_027322(1.5),XLOC_071533(1.5),XLOC_009315(2.0),XLOC_007388(2.0),XLOC_088835(2.0),XLOC_115984(2.0),XLOC_106230(2.0),XLOC_044559(2.0)	Os09g0260000 protein;LIM zinc-binding domain-containing protein;Protein kinase domain-containing protein;Cysteine-rich receptor-like protein kinase 8;Transcription factor GAMYB;Lectin receptor kinase (Lectin receptor kinase 1);F-box domain-containing protein;Lectin receptor kinase;Subtilisin-like protease SDD1	Cstv	siRNA	H_novel
S-344	AAAAAUUGCGGAUUUUUGGCU	noHit	7B	mRNA	XLOC_146342(1.0),XLOC_139391(2.0),XLOC_145885(2.0),XLOC_145208(2.0),XLOC_130100(2.0),XLOC_048529(2.0)	1-phosphatidylinositol-3-phosphate 5-kinase;Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase	Cstv	siRNA	H_novel
S-347	GUGGGCGGCGAAGAAAGACGG	noHit	6B	siRNA,TAS	XLOC_120603(0.0),XLOC_039307(1.5),XLOC_088372(2.0),XLOC_134608(2.0),XLOC_125224(2.0),XLOC_113393(2.0),XLOC_115778(2.0),XLOC_033145(2.0),XLOC_033668(2.0)	Protein kinase domain-containing protein;Glutaredoxin domain-containing protein;BUB1 N-terminal domain-containing protein;S5A_REDUCTASE domain-containing protein;Os07g0159900 protein	Cstv	siRNA	F_PSUSStaeMirSir
S-348	GUCUAGCAAUAUAUUUGCAGG	noHit	7A,7B,7D	mRNA;Protein kinase domain-containing protein;	XLOC_146342(0.0),XLOC_144188(1.0),XLOC_146456(1.0),XLOC_144697(1.0),XLOC_006347(1.5),XLOC_138664(1.5),XLOC_128874(1.5),XLOC_121255(1.5),XLOC_106969(1.5),XLOC_008378(2.0),XLOC_028802(2.0),XLOC_139634(2.0),XLOC_147970(2.0),XLOC_130242(2.0),XLOC_128194(2.0),XLOC_122677(2.0),XLOC_120925(2.0),XLOC_112063(2.0),XLOC_122301(2.0),XLOC_108363(2.0),XLOC_101380(2.0),XLOC_043133(2.0),XLOC_069475(2.0)	Pto-interacting protein 1;Putative disease resistance protein RGA4;Protein kinase domain-containing protein;Receptor-like serine/threonine-protein kinase;Cysteine-rich receptor-like protein kinase 29;Disease resistance protein RGA2;Wall-associated receptor kinase 5;RNA helicase;WAT1-related protein;NB-ARC domain-containing protein;Cysteine-rich receptor-like protein kinase 41;G-type lectin S-receptor-like serine/threonine-protein kinase SRK	CRup	siRNA	H_novel
S-358	CGACGGGAGGAUGACGGUGGU	noHit	6B	siRNA,TAS	XLOC_120603(0.0),XLOC_032875(1.5),XLOC_003762(2.0),XLOC_030610(2.0),XLOC_061233(2.0),XLOC_087005(2.0),XLOC_081210(2.0),XLOC_143213(2.0),XLOC_150345(2.0),XLOC_074924(2.0)	BHLH domain-containing protein;Os03g0604600 protein (Fragment);Os02g0710300 protein (cDNA clone:002-143-H06, full insert sequence);RING-type E3 ubiquitin transferase;Transcription factor bHLH83;Probable E3 ubiquitin-protein ligase BAH1-like 1;Probable E3 ubiquitin-protein ligase BAH1-like 2;Protein HIRA	Cstv	siRNA	G_BlastnPreMirSir
S-360	UCAAAACACCUUGGCAAAGUA	noHit	6D	siRNA,TAS	XLOC_126981(0.0),XLOC_114162(1.5),XLOC_009150(2.0)	Oligopeptide transporter 7	CRdw	siRNA	G_BlastnPreMirSir

S-361	CCUCCCCCUUCGCUUCCUCUU	noHit	4A	siRNA,TAS,TASiRNA,SiRNA, mRNA	XLOC_098195(1.0),XLOC_003248(1.5),XLOC_082991(1.5),XLOC_107918(1.5),XLOC_001486(2.0),XLOC_017653(2.0),XLOC_058245(2.0),XLOC_085089(2.0),XLOC_084922(2.0),XLOC_089093(2.0),XLOC_145430(2.0),XLOC_128488(2.0),XLOC_134814(2.0),XLOC_114768(2.0),XLOC_103961(2.0),XLOC_096679(2.0),XLOC_096292(2.0),XLOC_051896(2.0),XLOC_038756(2.0),XLOC_040201(2.0),XLOC_078297(2.0),XLOC_073718(2.0),XLOC_070548(2.0),XLOC_075193(2.0),XLOC_075186(2.0),XLOC_070010(2.0)	Leucoanthocyanidin dioxygenase;17.4 kDa class I heat shock protein 3;17.4 kDa class I heat shock protein;Putative serine/threonine-protein kinase;Serine/threonine-protein kinase TOUSLED;FHA domain-containing protein;Sugar transporter ERD6-like 6;Os01g0954500 protein (cDNA clone:J033093N15, full insert sequence);Bidirectional sugar transporter SWEET2a;Zinc finger protein 8;Anthocyanidin synthase;Protein NRT1/ PTR FAMILY 3.1;Tousled-like protein kinase;PPR_long domain-containing protein;C2H2-type domain-containing protein;Bidirectional sugar transporter SWEET;dCTP pyrophosphatase 1;Macrophage migration inhibitory factor (Tautomerase/MIF superfamily protein);Leucoanthocyanidin dioxygenase 1 (LDOX1) (Leucocyanidin oxygenase 1);Os03g0216000 protein (Zinc finger family protein, putative, expressed) (cDNA clone:002-148-B06, full insert sequence);Glycosyltransferase;Alkaline/neutral invertase;Os03g0749800 protein (Tousled-like kinase 2, putative, expressed) (cDNA clone:J013098L24, full insert sequence);Tetratricopeptide repeat protein 38;Expressed protein (Os03g0710700 protein);MFS domain-containing protein;Os09g0416900 protein;Macrophage migration inhibitory factor family protein	Cstv	siRNA	F_PSUSStaeMirSir
S-366	ACAGAAUCCGGAUAUAGUACU	noHit	6B	mRNA	XLOC_028774(2.0)	Protein kinase domain-containing protein;Os07g0152200 protein (cDNA, clone: J090052O12, full insert sequence)	Cstv	siRNA	F_PSUSStaeMirSir
S-369	GAGAAAGGAGUGGAACACUUA	noHit	6D		XLOC_129114(0.0),XLOC_028464(1.0),XLOC_025119(1.0),XLOC_128038(1.0),XLOC_042572(1.0),XLOC_004712(1.5),XLOC_004711(1.5),XLOC_055091(1.5),XLOC_062995(1.5),XLOC_116520(1.5),XLOC_033560(1.5),XLOC_003102(2.0),XLOC_021114(2.0),XLOC_081115(2.0),XLOC_034633(2.0)	Cytochrome P450 family 96 subfamily A polypeptide 1;Cytochrome P450 86B1;Os05g0181200 protein (Fragment);Putative 26S proteasome non-ATPase regulatory subunit 7;Os04g0466600 protein (cDNA, clone: J100056P20, full insert sequence);HVA22-like protein;NB-ARC domain-containing protein;Reverse transcriptase domain-containing protein;CCHC-type domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel
S-370	CACAGAGGUGCGGAAGCUGAC	noHit	6D		XLOC_129421(0.0),XLOC_131205(1.5),XLOC_002729(2.0)	O-methyltransferase ZRP4;5-pentadecatrienyl resorcinol O-methyltransferase;Methyltransf_2 domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel
S-378	GAGGAAGCGAAGGGGGAGGAU	noHit	4A	siRNA,TAS,TASiRNA,SiRNA, mRNA	XLOC_074878(0.0),XLOC_084749(0.5),XLOC_077943(1.0),XLOC_014056(1.5),XLOC_017957(1.5),XLOC_012221(1.5),XLOC_085368(1.5),XLOC_133826(1.5),XLOC_119854(1.5),XLOC_112295(1.5),XLOC_106459(1.5),XLOC_048232(1.5),XLOC_044481(1.5),XLOC_034777(1.5),XLOC_077320(1.5),XLOC_076710(1.5),XLOC_018097(2.0),XLOC_023164(2.0),XLOC_022873(2.0),XLOC_011270(2.0),XLOC_018453(2.0),XLOC_062349(2.0),XLOC_089141(2.0),XLOC_151460(2.0),XLOC_129865(2.0),XLOC_105691(2.0),XLOC_096649(2.0),XLOC_053926(2.0),XLOC_053339(2.0),XLOC_057141(2.0),XLOC_057481(2.0),XLOC_044202(2.0),XLOC_035766(2.0),XLOC_034479(2.0),XLOC_078917(2.0)	ABI3VP1 transcription factor (B3 domain-containing transcription factor NGA2) (Fragment);Os02g0697300 protein (TOM1 protein) (cDNA clone:J013047D24, full insert sequence);TOM1-like protein 2;B3 domain-containing protein Os03g0120900;TF-B3 domain-containing protein;Nitrate transporter 1.4;Os01g0747800 protein (cDNA clone:J033023G10, full insert sequence);VQ domain-containing protein;Double-stranded RNA-binding protein 3;Beta-galactosidase;H15 domain-containing protein;Retrotrans_gag domain-containing protein;FYVE-type domain-containing protein;Glucan endo-1,3-beta-D-glucosidase;Rab3 GTPase-activating protein catalytic subunit;Lipase_GDSL domain-containing protein;GRAS domain-containing protein;Os05g0150400 protein (Fragment);Retrovirus-related Pol polyprotein from transposon TNT 1-94;Double-stranded RNA-binding protein 2;Os03g0638800 protein (Protein kinase, putative, expressed) (cDNA, clone: J100090O16, full insert sequence);Protein kinase domain-containing protein;Auxin response factor;Histone H1;Histone H1 WH1A.2;1,4-alpha-D-glucan glucanohydrolase;KIX_2 domain-containing protein	Cstv	siRNA	G_BlastnPreMirSir
S-379	UCUUGACCUUGUAAGACCCAA	noHit	4A,4D	SiRNA,TAS,TASiRNA,mRNA ,siRNA;Alkaline/neutral invertase;	XLOC_082991(0.0),XLOC_045491(0.0),XLOC_078297(0.0),XLOC_007620(0.5),XLOC_001556(0.5),XLOC_014542(0.5),XLOC_064742(0.5),XLOC_047957(0.5),XLOC_057702(0.5),XLOC_056846(0.5),XLOC_006150(1.0),XLOC_010092(1.5),XLOC_007597(1.5),XLOC_001935(1.5),XLOC_015057(1.5),XLOC_062065(1.5),XLOC_062262(1.5),XLOC_054662(1.5),XLOC_048256(1.5),XLOC_048685(1.5),XLOC_105996(2.0),XLOC_101183(2.0),XLOC_089940(2.0)	AB hydrolase-1 domain-containing protein;Alpha/beta-Hydrolases superfamily protein;Hydrolase_4 domain-containing protein;Auxin response factor;Os12g0152600 protein (Rf1 protein, mitochondrial, putative, expressed);Auxin response factor 2 (ETTIN-like protein 2) (OsETTIN2);Alkaline/neutral invertase;Auxin response factor 3	Cstv	siRNA	F_PSUSStaeMirSir
S-381	UAUGGAAUGAUCAAGGAAUUU	noHit	7B	mRNA	XLOC_087543(1.5),XLOC_146342(1.5),XLOC_145356(2.0)	Protein kinase domain-containing protein;Acetyl-CoA carboxylase;Acetyl-CoA carboxylase 1	Cstv	siRNA	H_novel
S-384	AUGGAAUGAUCAAGGAAUUUG	noHit	7B	mRNA	XLOC_146342(1.5),XLOC_058266(2.0)	Protein kinase domain-containing protein	Cstv	siRNA	H_novel
S-385	CUUUGUCAAGGUGUUUGAGAG	noHit	6B	siRNA,TAS	XLOC_120603(0.0),XLOC_108359(1.5),XLOC_101447(1.5),XLOC_091108(1.5),XLOC_123929(2.0)	Protein NRT1/ PTR FAMILY 7.1;Os02g0716800 protein;Calcium-dependent protein kinase 32	Cstv	siRNA	G_BlastnPreMirSir

S-420	ACCUCGACAAGAACAACAGCU	noHit	2B,2D	mRNA;Disease resistance protein RPM1;NB-ARC domain-containing protein;	XLOC_083776(2.0),XLOC_140336(2.0),XLOC_144590(2.0),XLOC_145395(2.0),XLOC_131736(2.0)	E3 ubiquitin-protein ligase RGLG2;Copine-1;Os06g0608800 protein (cDNA clone:J013115P09, full insert sequence) (cDNA clone:J023106M02, full insert sequence);E3 ubiquitin-protein ligase RGLG1;Copine domain-containing protein;31 kDa ribonucleoprotein, chloroplastic;RING-type domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-440	GGUGGAUAGCCGAAGGCCUUG	noHit	1D	mRNA; (Fragment);NB-ARC domain-containing protein;Putative disease resistance RPP13-like protein 3	XLOC_114304(2.0)	Disease resistance RPP8-like protein 3;NB-ARC domain-containing protein	Cstv	siRNA	H_novel
S-442	ACCUCGACAAGAACAGCAGCU	noHit	2A	mRNACysteine-rich receptor-like protein kinase 10;Disease resistance protein RPM1;NB-ARC domain-containing protein;Protein kinase domain-containing protein	XLOC_010418(2.0)	Os05g0120800 protein (cDNA clone:J013087A14, full insert sequence)	Cstv	siRNA	F_PSUSStaeMirSir
S-443	UUCCACGCGUAUCGGUUUGCC	noHit	6B	mRNAPutative late blight resistance protein-like protein R1B-16;NB-ARC domain-containing protein;	XLOC_117365(0.0)	NB-ARC domain-containing protein	Cstv	siRNA	F_PSUSStaeMirSir
S-447	GUGGACUGCGAAGAAAGACUG	noHit	6A	siRNA,TAS	XLOC_079597(1.5)		Cstv	siRNA	F_PSUSStaeMirSir
S-449	GAACAAGAAGAUGAAGAGUGU	noHit	4A	siRNA,TAS,TASiRNA,SiRNA,mRNA	XLOC_076710(0.0),XLOC_074878(0.0),XLOC_104894(1.0),XLOC_099259(1.0),XLOC_054541(1.5),XLOC_101050(1.5),XLOC_089129(2.0),XLOC_086478(2.0),XLOC_142714(2.0),XLOC_142810(2.0),XLOC_151337(2.0)	Protein kinase domain-containing protein;ABC transporter C family member 10;AAA domain-containing protein;ABC transporter C family member 8;Retrotrans_gag domain-containing protein;DUF668 domain-containing protein;Glycosyltransferase;Receptor-like serine/threonine-protein kinase SD1-6	Cstv	siRNA	G_BlastnPreMirSir
S-454	AAAACACAUACACUUCUUCUG	noHit	2A		XLOC_137656(1.5),XLOC_095151(1.5)		CRdw	siRNA	H_novel
S-462	CACCUGAGGUCGUUCAAUGUU	noHit	6D		XLOC_129568(0.0),XLOC_129421(0.0)	NB-ARC domain-containing protein;Disease resistance protein RPM1	Cstv	siRNA	H_novel

APPENDIX C

SIWALK

Journal Title Here, 2025, pp. 1–22

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Improving siRNA Prediction in Arabidopsis with Machine Learning

Chao-Jung Wu[1] and Abdoulaye Baniré Diallo[1,*]

[1]Département d'informatique, Université du Québec à Montréal, P.O. Box 8888 Downtown Station, H3C 3P8, Québec, Canada
[*]Corresponding author, email: diallo.abdoulaye@uqam.ca
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Plant small interfering RNAs (siRNAs) regulate gene expression but are difficult to identify due to confounding factors with other RNA, limited resources, and lack of prediction criteria. Existing methods rely on phase scoring, which measures alignment periodicity, but often fails to distinguish siRNAs from other small RNAs and leads to high false-positive rates. We propose to improve *Arabidopsis* siRNA prediction by incorporating features derived from phased siRNA precursors beyond periodicity into a machine learning classifier. Leveraging a Random Forest classifier, siWalk combines biogenesis knowledge and mathematical modeling of siRNA precursors. Its algorithms detect phased positions, define and tag phasing boundaries, compute key attributes, unify related segments, and sort siRNAs by calculating weights for their effector properties. SiWalk predicts siRNA precursors from sRNA-seq alignments with greater precision and accuracy than existing phase scoring methods. It validates highly conserved siRNA loci by highlighting the influential expressional and structural features for siRNA prediction. By focusing on structure-only features, siWalk can also identify localizations of effector siRNAs in a sequence without requiring sRNA-seq data. SiWalk, with its robust learning and sorting kernel, enriches the toolkit for annotating genome-wide phased siRNA precursors in diverse tissues and treatments and for selecting functional siRNAs in *Arabidopsis*.

Key words: Machine learning in siRNA prediction, siRNA localization, siRNA features, Functional genomics, Agritech

Introduction

Plant small interfering RNAs (siRNAs) play key roles in post-transcriptional gene silencing [Agrawal et al., 2003, Li et al., 2017], regulating critical processes such as development [Zhang et al., 2019, Liu et al., 2020], stress responses [Khraiwesh et al., 2012] and pathogen defense [Llave, 2010, Zhang and Zhu, 2011]. The generation of siRNAs begins with the formation of double-stranded RNA (dsRNA) from sources such as viral infections, transposable elements, RNA-dependent RNA polymerase (RDR) activity on single-stranded RNA (ssRNA) transcripts [Baulcombe, 2004, Vaucheret and Voinnet, 2024, Axtell, 2013] or the self-folding of ssRNA [Kakrana et al., 2018, Piriyapongsa and Jordan, 2008]. Dicer-like enzymes cleave double-stranded RNA (dsRNA) into 21- to 24-nucleotide siRNAs [Kim et al., 2005, Fukudome and Fukuhara, 2017] through sequential cycles. This process produces siRNA **phasing** clusters from dsRNA substrates, which are regularly spaced and periodically aligned on the precursor genome [Howell et al., 2007, Chen et al., 2007] (illustrated in Figure S1). In some pathways, microRNA (miRNA) triggers siRNA production by cleaving ssRNAs, which are then converted into dsRNA by RDR [Allen et al., 2005, Axtell et al., 2006, Chen et al., 2010]. Finally, selected siRNAs are loaded into the RNA-induced silencing complex (RISC), where they function as active **effectors** and guide the complex to silence target genes [John et al., 2004, Dai et al., 2018].

SiRNAs are promising biomarkers for monitoring plant health and improving agricultural outcomes. Many existing siRNA prediction methods essentially predict siRNA precursors from small RNA sequencing (sRNA-seq) data, and assign the most abundant siRNA species as the effector if its position corresponds to Dicer processing cycles. They rely on one of three main phase scoring schemes to evaluate siRNA production from expressed genomic regions: Howell [Howell et al., 2007], Chen [Chen et al., 2007] and Guo [Guo et al., 2014]. These scoring approaches represent the earliest efforts to disentangle confounding factors associated with other RNA molecules and identify phasing patterns among small RNAs aligned to a precursor. Howell et al. [2007] introduced the phase transformation score, a metric that quantifies phasing strength by linking the periodic alignment of RNAs to their abundance. Chen et al. [2007] achieved siRNA detection by using the hypergeometric distribution to evaluate the occupancy of in-phase positions by expressed siRNAs among all in-phase positions in a search window. It ignores sequencing read accumulation and instead emphasizes rigorous statistical assessment of phasing significance. Guo et al. [2014] advanced siRNA prediction with the software

© The Author 2025. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

PhaseTank, which connects miRNA triggers to siRNA loci and their downstream targets—an area beyond the scope of this study. Guo's phase scoring approach integrates siRNA count, abundance and phased ratio to quantify phasing strength and quality.

While phase scoring methods have laid the foundation for subsequent advancements in siRNA prediction, they are not without notable limitations. First, their ability to differentiate phased RNAs from others has not been evaluated. Although originally developed in *Arabidopsis*, these scoring schemes have been widely adopted for plant siRNA discovery in various crops [Zhang et al., 2012, De Paoli et al., 2009, Zheng et al., 2014, Dotto et al., 2014, Kakrana et al., 2018]. They typically rely on fixed thresholds without quantitative justification. Assessing phase scores across various thresholds would deepen our understanding of their optimal use in predicting siRNA precursors.

Second, interpreting phase scores remains challenging because they alone do not fully distinguish siRNAs from other RNAs that align to adjacent genomic regions. Since scoring schemes focus on alignment patterns, they often overlook the structural and biochemical properties that are crucial for siRNA biogenesis. Members of the siRNA superfamily are traditionally classified based on various characteristics into distinct types, such as miRNAs, piRNAs (piwi-interacting), ta-siRNAs (trans-acting), nat-siRNAs (natural antisense), hc-siRNAs (heterochromatic) and mirtrons (intronic), among others [Piriyapongsa and Jordan, 2008, Farazi et al., 2008, Rorbach et al., 2018, Axtell and Meyers, 2018, Kakrana et al., 2018, Agrawal et al., 2003]. The characteristics for classification include the mechanism of dsRNA formation, their genetic origin (e.g., coding, heterochromatic, terminal-repeat, intronic, viral), the post-transcriptional modifications they regulate (e.g., cleavage, methylation), precursor structure (e.g., hairpin, length) and tissue specificity (e.g., reproductive tissues). Incorporating these characteristics into a classifier could improve siRNA prediction beyond relying only on alignment. Building on the study by Leclercq et al. [2013], which improved miRNA prediction by training a Random Forest classifier on 100 diverse features, siRNA prediction can similarly benefit from leveraging machine learning's capacity to learn a wide range of features and to distinguish between classes.

Third, although singly phased miRNAs have well-defined prediction criteria [Meyers et al., 2008, Axtell and Meyers, 2018, Wu et al., 2019], the criteria for multiply phased siRNAs remain poorly established. This gap contributes to higher false-positive rates and less reliable siRNA detection.

Fourth, siRNA research is constrained by the availability of specialized software and databases. Unlike miRNA, which has over 160 analysis tools [Lukasik et al., 2016], only a few are available for siRNA: Notable software includes ShortStack [Johnson et al., 2016] and PHASIS [Kakrana et al., 2017], and databases such as sRNAanno [Chen et al., 2021] and MPSS [Nakano et al., 2020]. A new prediction tool with enhanced performance would enrich the bioinformatics toolkit and advance siRNA research.

Fifth, existing resources assign different identifiers and coordinates to the same phasing events (Table S1, creating barriers to data standardization and comparative analyzes. The assignment of a standardized genomic tag to each locus would resolve discrepancies in siRNA precursor identification and facilitate better integration between studies.

Sixth, for a given precursor, one might be curious about the potential siRNA production sites. However, sRNA-seq data, typically used by existing methods to predict precursors and assign effectors based on expression rather than directly predicting them, are not available in this scenario. Thus, localizing effector sites directly from the precursor sequence requires a new classifier independent of sRNA-seq.

In this study, we present siWalk, a novel bioinformatics method that uses machine learning to predict siRNAs with improved precision, accuracy, and biological insight. SiWalk overcomes the limitations of conventional methods by integrating a comprehensive knowledge of siRNA biogenesis, including structural and biochemical properties that influence interactions and stability, and the periodicity and abundance of sequence alignments. These elements are incorporated into a hybrid machine learning and mathematical scoring method for precise precursor identification and effector localization. The primary contributions of this study include:

- SiWalk is the first tool to apply machine learning for siRNA prediction, integrating current biogenesis knowledge with mathematical modeling. Predict siRNA precursors from sRNA-seq data and improve precision and accuracy compared to existing methods. Furthermore, it predicts the most likely localizations of effector siRNAs in a sequence, without the requirement of sRNA-seq data.
- Rationally designed algorithms identify phased positions, define locus boundaries, compute key attributes, and merge related phasing segments.
- Feature analysis reveals key structural and nucleotide motifs that could be crucial to distinguish true siRNAs, providing insight into siRNA biogenesis and biological relevance.
- A standardized genomic tag is proposed to reduce discrepancies in siRNA-generating loci identification across datasets.

By improving precursor prediction, advancing effector siRNA localization, refining phase boundary and tag, and demonstrating model generalization, siWalk enriches the toolkit for annotating genome-wide phased siRNA precursors in diverse tissues and treatments and reduces the barriers associated with selecting functional siRNAs in *Arabidopsis*.

Materials and methods

Datasets

Reference contigs of Arabidopsis siRNA-generating loci

Since no central repository for plant endogenous siRNA precursors was available, we obtained the coordinates of *Arabidopsis* phased siRNA precursors from four studies: 21 precursors from Chen et al. [2007], 18 from Howell et al. [2007], 37 from Nakano et al. [2020], and 35 from Chen et al. [2021]. All identified loci corresponded to 21-nt (nucleotide) siRNAs. Many loci from these sources overlapped but did not align exactly, likely due to variations in expression levels between libraries, precursor boundary definition and annotation approaches.

To streamline the analysis, we indexed the entire genome into **contigs**, each spanning 5250 nucleotides with a 250-nt shared region between consecutive contigs, ensuring that events at the junctions were captured. To reduce ambiguity, we unified overlapping precursors using the standardized contig tags (Table S1). This process yielded 51 reference contigs of siRNA-generating loci, each 5250-nt long (Table S2).

Table 1. Training dataset.

Class	Segment count	%Total	Ratio	Corresponding to contig count
Negative	174856	94.24	16	6034
Total	185539	100		∼27000 (by genome size)

Arabidopsis sRNA-seq libraries

Arabidopsis sRNA-seq libraries were obtained from NCBI Gene Expression Omnibus repository: two from GSM518432 and GSM738727 [Lei and Sun, 2014], and 27 from GSE44622 [Jeong et al., 2013]. Using the least stringent criterion (at least three nearby phase positions occupied by 21-nt RNAs), 50 of 51 reference contigs were expressed in these 29 libraries.

Curation of Arabidopsis positive and negative sets generation

As shown in Table S1, phasing events are dynamic with variable start and end positions at the same locus. To capture the variability across libraries and experiments and to provide examples for classifier training, we curated phasing segments with precise boundaries from sRNA-seq libraries. This approach ensured the retrieval of relevant precursor sequences and corresponding effector siRNAs for feature generation. sRNA-seq libraries were aligned to genome using Bowtie [Langmead et al., 2009] allowing no mismatches, and partitioned into contigs using SAMtools [Danecek et al., 2021]. We computed six phase scores (Supplementary C) for each search window by analyzing the RNA alignments within that region. Within a contig, nearby in-phase segments and those sharing the same putative effector siRNA sequences were merged into a single segment (see Section D). A general filter (requiring the segment to be expressed, phase positions to be occupied, and sufficient effector reads to be present) was applied to reduce instances for training efficiency. Significant phasing events were identified by modeling phase scores with gamma distributions and applying a local false discovery rate (FDR) threshold. A phasing segment was assigned to the positive set of siRNA-generating precursors:

- if its contig was among the reference contigs, and
- if at least two of its six phase scores were significant (FDR < 0.05), and
- if its **effseq** was present in at least two libraries, where effseq is the most abundant 21-nt siRNAs considering phase drift by 1 or 2 positions;
- otherwise, it was placed in the negative set.

The final training dataset, compiled from 29 *Arabidopsis* sRNA-seq libraries, includes 10683 positive and 174856 negative segment examples (Table 1).

Feature vectors and training

Each training example was represented by **245 raw features** (summarized in Table S6, see Supplementary I for details). These features, categorized into **expressional** and **structural**, were derived from the genomic alignment of small RNAs, encompassing primary sequence, abundance, alignment patterns, precursor structural layout, folding motifs, and molecular interactions. Features of interest included expression statistics from sRNA-seq alignments that capture phasing, such as phase scores, the count of phase positions expressed by 21-nt RNAs, their accumulation at phase positions, and strand-specific expression proportions from the precursor. Precursor length, G/C content, nucleotide composition, and mono-, di-, and tri-nucleotide frequencies within the primary effector siRNA sequence, as well as nucleotide types at key effector positions, were also considered [Lück et al., 2019]. These sequence-based features can influence the strand selection and RISC loading process, ultimately affecting siRNA functionality [Takeda et al., 2008]. Structural motifs (loops, bulges, hairpins, triplet folding modes) [Leclercq et al., 2013] and miRNA targeting [Axtell et al., 2006] were included to capture the molecular interaction behavior of siRNA precursors, which may impact their stability and processing by Dicer-like enzymes.

Features were obtained using computational tools, string pattern searches and statistics. Phasing potential embedded in alignment patterns was evaluated by phase scores (Supplementary C). RNAfold was used to determine the minimum free energy (MFE) and structural layout of candidate precursors [Lorenz et al., 2011]. The precursor hairpin structure was analyzed by MIRcheck [Jones-Rhoades and Bartel, 2004]. To identify miRNAs that trigger siRNA production, *Arabidopsis* miRNAs were retrieved from miRBase v22.1 [Kozomara et al., 2019], and cleavage on candidate precursors was predicted using miRanda [John et al., 2004].

We used scikit-learn [Pedregosa et al., 2011] for data preprocessing, feature selection, classifier training and evaluation. Features were discretized and encoded into **1502 attributes** (Supplementary I.2). The rank scores, representing the importance of individual features for precursor classification, were provided by the chi-squared test [McHugh, 2013], and the most relevant attributes were selected by SelectKBest.

SiWalk utilizes a **Random Forest** classifier, which aggregates multiple decision trees on subsamples to improve overall accuracy [Breiman, 2001]. The rationale for selecting it over alternative classifiers such as **Gradient Boosting** [Friedman, 2001], **Logistic Regression** [Yu et al., 2011], and **Support Vector Machine** [Chang and Lin, 2011] is detailed in the Results section. Pairing with **AdaBoost**, which focuses on misclassifications and combines the weighted votes of weak classifiers [Freund and Schapire, 1997], can further enhance accuracy. The optimal parameters maximizing precision are listed in Table S3.

The performance of each classifier, trained on all or the subset of best attributes, was evaluated using 5-fold cross-validation strategy. In standard split, data were split into training and testing sets based on class labels to ensure proportional representation in each fold. In stratified split, data were divided to preserve the distribution of both class and library labels, ensuring testing on unseen libraries and reducing the risk of overfitting to specific library characteristics. Metrics assessed included precision, sensitivity, specificity, accuracy, Matthew's correlation coefficient (MCC, addressing class imbalance) and F1 score, as described previously [Leclercq et al., 2013].

siRNA indication score

Drawing from Leclercq et al. [2013]'s approach to miRNA localization, which showed that scoring improved localization, we found that combining machine learning with scoring also enhances siRNA localization. Take a 300-nt precursor, each nucleotide is a potential starting position for the effector siRNA. Considering siRNAs of 19 to 23 nt provides 1500 possible position-length effector candidates. To identify the

best candidates, we adopted and refined their approach as follows.

The process first sorts candidates using machine learning classifications, followed by a secondary sorting based on indication scores. Presenting a short list (size $= t$) of top candidates increases the likelihood of including the true effector.

The classification for each candidate of length l at position p on the precursor (length $= n \gg t$) is determined by a machine learning model M, with the siRNA length l set to a range of $DicerCall \pm 2$, where $DicerCall = \{18, 19, .., 31\}$ represents the interval of Dicer cleavage. The range for l is $19 \leq l \leq 23$ if $DicerCall = 21$.

$$siRNA_{true} = \left\{ (p,\, l) \middle| \begin{matrix} M(p,l) = True, \\ 19 \leq l \leq 23,\ 0 \leq p \leq n - l + 1 \end{matrix} \right\} \quad (1)$$

The **indication score** is then used to sort $siRNA_{true}$ or to identify additional siRNAs until the list reaches size t. It is calculated in three steps:

- The base score for each candidate is calculated as the sum of the products of importance, correlation, and value for each **structural** feature i associated with the siRNA [Chao et al., 2022]. Feature importance is determined by their ranking score from the chi-squared test. Correlation between a variable and a binary classification is assessed using the Point-Biserial Correlation [Kornbrot, 2014].

$$score(p,l) = \sum_i importance_i \times correlation_i \times value_i \quad (2)$$

- The consensus scores for p being the starting position $S(p)$ or the ending position $E(p)$ are defined as in Leclercq et al. [2013].
- The most likely siRNA start and end positions are determined by the largest combinations of start and end position scores [Leclercq et al., 2013]. The $argsort$ function returns a sorted list of most likely siRNAs from a given precursor.

$$\{siRNA_1, siRNA_2, \dots\} = \left\{ (p,\, l) \middle| \underset{\substack{19 \leq l \leq 23, \\ 0 \leq p \leq n-l+1}}{argsort} \quad [S(p) + E(p+l-1)] \right\}$$

Results

We developed siWalk, a hybrid classifier that leverages machine learning and integrates mathematical scoring to enhance siRNA prediction. SiWalk operates in two modes:

- **Prediction mode**, where it predicts genome-wide siRNA precursors from sRNA-seq data and assign the effector siRNAs.
- **Localization mode**, where it evaluates a given long RNA sequence and predicts the most likely effector siRNA localizations.

We curated phased siRNA precursors originating from *Arabidopsis* reference contigs. A set of expressional and structural features were extracted from the examples and ranked by their importance for precursor classification. Using these features, machine learning models from various algorithms were trained and evaluated, and a final model was selected for its predictive performance. In localization mode, a sorting process was incorporated alongside machine learning to compensate for the absence of sRNA-seq data, as expression-derived phasing boundaries were unavailable for feature generation. Here, we presented the features for phased siRNA prediction, the training and evaluation of siRNA classifiers, and the delivery of siWalk.

Evaluation of individual predictive features

Feature construction was shaped by previous studies of miRNA duplex and hairpin characterization [Leclercq et al., 2013], siRNA precursor prediction [Chen et al., 2007, Howell et al., 2007, Guo et al., 2014] and properties relevant to siRNA production [Axtell et al., 2006, Lück et al., 2019, Takeda et al., 2008]. The predictive power of **1502 attributes**, generated through the encoding of **245 raw features**, was evaluated.

We observed that the attributes ranked as highly important by the chi-squared test generally showed strong positive or negative correlations with precursor classes (Figure S3). As shown in Table 2, important attributes for precursor classes included scoring (e.g., k, Guo, effector abundance, pval, Howell), miRNA targeting, siRNA nucleotide compositions, precursor length and MFE. Key nucleotides included siRNA mono-, di-, and tri-nucleotides at the 5' end positions 1 and 2 (attribute names involving 5p1, 5p2), combinations of the first nucleotides at both ends (5p1, 3p3), and triplets at the center (md); They involved positions known to affect thermodynamics in determining precursor cleavage and strand selection [Takeda et al., 2008]. The nucleotide at position 3 of the 3' end (3p3) marked the cleavage site of the complementary strand. Triplet folding modes and nucleotide expression frequencies may also influence accurate precursor prediction. Since some attributes were structural and did not require expression data, using them was expected to help determining effector **localization** within the precursor. Among the key structural attributes (Table 2), miRNA targeting, precursor MFE, effector nucleotide composition, triplet folding mode, and the start position within the precursor could be significant factors driving effector siRNA production.

To explore whether these important features can support siRNA precursor classification, we examined how classifier performance changes with increasing features. Selecting top-ranked features from a set of 245 showed that predictive performance plateaued between 45 and 115 features with reduced computation time (Figure S4), suggesting that this range captures the essential features for classification. Considering the performance stabilization within this range and the practical constraints of evaluating multiple subsets, we selected 100 attributes and extended the analysis to various classifiers. As shown in Table S5A, this subset often reduced computation time across classifiers compared to using the full set. Further evaluations are presented in the following sections.

Training and evaluation of siRNA classifiers

We assessed the predictive performance of several machine learning models, including Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVC), each paired with complete or partial feature subsets, or structural-only attributes (Table 3). These models

Table 2. Top predictive features for precursor **pred**iction and effector **loca**lization.

Attribute	Description	Property	**Usage**	Rank score
k	number of phase positions expressed by 21-nt RNAs within the 9-cycle window	phase score	pred	3822.9
ext_k	updated number of phasing positions after increasing the boundary beyond the 9-cycle window	phase score	pred	2447.4
n	number of all positions expressed by 21-nt RNAs within the 9-cycle window	phase score	pred	1143.5
Guo_b_ccdf	Guob, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	1045.7
twohitBool_True	miRNA targets the precursor twice or more	miRNA targeting	pred, loca	999.5
eff_frq	frequency of the most abundant effector siRNA allowing drift	phase score	pred	850.3
length	length of the precursor	length	pred	807.5
md3C	expression-weighted nucleotide type percentage at **md3** on effector	nucleotide frequency	pred	803.0
eff5p1di_TA	di-nucleotide type at **5p1** on effector	nucleotide type	pred, loca	783.1
BestScore	the highest targeting score among miRBase microRNAs	miRNA targeting	pred, loca	715.0
md2C	expression-weighted nucleotide type percentage at **md2** on effector	nucleotide frequency	pred	705.0
3p3G	expression-weighted nucleotide type percentage at **3p3** on effector	nucleotide frequency	pred	696.8
eff5p1tri_TAA	tri-nucleotide type at **5p1** on effector	nucleotide type	pred, loca	669.7
pval_b	pvalb, the minimum score in a precursor's scanning windows	phase score	pred	639.8
max_mfe	the maximum MFE of folding sequences in a precursor's scanning windows	MFE	pred	616.4
premfe	MFE of the precursor in [Rbound, Lbound], calculated by RNAfold	MFE	pred, loca	614.6
NBtripletC	occurrences of triplet C around eff_seq	triplet	pred, loca	554.6
5p1T	expression-weighted nucleotide type percentage at **5p1** on effector	nucleotide frequency	pred	531.9
5p1C	expression-weighted nucleotide type percentage at **5p1** on effector	nucleotide frequency	pred	527.1
eff5p2tri_ATG	tri-nucleotide type at **5p2** on effector	nucleotide type	pred, loca	525.3
eff5p1_T	mono-nucleotide type at **5p1** on effector	nucleotide type	pred, loca	514.0
eff5p1_3p3_TG	the first and the last nucleotides (**5p1, 3p3**) on effector	nucleotide type	pred, loca	513.5
effmd4tri_AAG	tri-nucleotide type at **md4** on effector	nucleotide type	pred, loca	499.7
GC%	expression-weighted percentage of G and C combined in a precursor	nucleotide frequency	pred	487.3
effmd1tri_ATG	tri-nucleotide type at **md1** on effector	nucleotide type	pred, loca	485.6
Howell_ccdf	Howell, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	481.1
5p4C	expression-weighted nucleotide type percentage at **5p4** on effector	nucleotide frequency	pred	479.1
Howellb_ccdf	Howellb, significant if < 0.05, Complementary Cumulative Distribution Function	phase score	pred	479.0
effmd2tri_GTA	tri-nucleotide type at **md2** on effector	nucleotide type	pred, loca	470.9
3p4C	expression-weighted nucleotide type percentage at **3p4** on effector	nucleotide frequency	pred	454.4
effmd2tri_AGA	tri-nucleotide type at **md2** on effector	nucleotide type	pred, loca	450.2
eff_pos	coordinate of the most abundant effector siRNA allowing drift	abundance	pred	449.1
dist_5p	eff_seq relative position on precursor	length	pred, loca	447.9
5p5C	expression-weighted nucleotide type percentage at **5p5** on effector	nucleotide frequency	pred	443.8
eff3p3_G	mono-nucleotide type at **3p3** on effector	nucleotide type	pred, loca	427.4

were compared to phase scoring methods with FDR and established siRNA prediction software, ShortStack v3.8 and PHASIS v3.3. Since segment-level information was not available for external software, we mapped their predictions onto contigs for comparing with the 50 expressed reference contigs. To maintain a 1:16 class ratio similar to our training dataset (Table 1), we arbitrarily assigned 818 negative contigs.

Although the actual class distribution in real-world scenarios is unknown, the 1:16 ratio reasonably approximates the assumption that most genomic segments express or regulate other RNA types, with only an unspecified fraction producing siRNAs. We believe that modeling siRNA production on an imbalanced dataset reflects the biological reality, and ensures better applicability of the classifier.

Boosting predictive power

As shown in Table 3A-B, phase scoring methods provided low F1 and MCC, highlighting their limitations. Although Guo identified most positives (TP approaching 5.76%), it suffered from low specificity due to a bias towards predicting positives. The accuracies of Howell, Guo, pval, ShortStack, and PHASIS remain below 94.24%, making them worse than random guessing (Figure S5). In contrast, machine learning approaches delivered higher precision, sensitivity, and accuracy, with fewer false positives (Table 3C). Take GBA100, it achieved an *accuracy gain* of 0.0346 over random guessing (60% of maximum 0.0576); other models showed similar gains. Recognizing that preselected fixed thresholds may not be optimal, we assessed ROC curves of scoring methods across various score thresholds when predicting precursor coordinates from 29 sRNA-seq libraries (Figure S6). We compared these methods to GBA100 by varying probability

Table 3. Performance evaluation of various classifiers with imbalanced examples (5.76% positive, 94.24% negative) **(A)** Phase scoring with FDR < 0.05, predicting a segment dataset of 185539 examples. **(B)** Existing phase scoring software such as ShortStack (Guo >= 30) and PHASIS (pval <= 0.005), predicting 50 expressed reference contigs and 818 negatives contigs across 29 sequencing libraries. **(C)** Machine learning using either the top 100 attributes, the full feature set, or the top 100 structure-derived attributes, predicting the same segment dataset with 5-fold cross-validation. **(D)** Machine learning with representative models (GBA100 and RFAs100), training and predicting the library-stratified segment datasets with 5-fold cross-validation.

(A) Phase score	TP%	TN%	FP%	FN%	Precision	Sensitivity	Specificity	Accuracy	MCC	F1
Howell	1.022	90.841	3.401	4.736	0.231	0.177	0.964	0.919	0.160	0.201
Howellb	1.739	79.770	14.472	4.019	0.107	0.302	0.846	0.815	0.094	0.158
Guo	5.725	7.082	87.161	0.033	0.062	0.994	0.075	0.128	0.063	0.116
Guob	5.740	18.608	75.634	0.018	0.071	0.997	0.197	0.243	0.116	0.132
pval	0.579	89.895	4.347	5.179	0.118	0.101	0.954	0.905	0.059	0.108
pvalb	1.075	88.093	6.150	4.683	0.149	0.187	0.935	0.892	0.109	0.166
(B) Existing software										
ShortStack (Guo)	0.727	87.216	7.024	5.033	0.385	0.126	0.926	0.879	0.152	0.167
PHASIS (pval)	2.275	91.694	2.546	3.485	0.599	0.395	0.973	0.940	0.427	0.422
(C) ML model										
GB100	3.446	93.248	0.994	2.312	0.805	0.598	0.989	0.967	0.671	0.673
GBall	2.720	93.456	0.786	3.038	0.798	0.472	0.992	0.962	0.591	0.583
GBA100	4.460	93.225	1.017	1.298	0.857	0.775	0.989	0.977	0.794	0.795
GBAs100	3.270	93.401	0.841	2.488	0.802	0.568	0.991	0.967	0.657	0.662
RF100	2.673	93.835	0.407	3.085	0.894	0.464	0.996	0.965	0.622	0.597
RFA100	2.915	93.813	0.430	2.843	0.904	0.506	0.995	0.967	0.653	0.631
RFAs100	3.840	93.806	0.436	1.918	0.911	0.667	0.995	0.976	0.766	0.765
LR100	1.970	92.919	1.323	3.787	0.680	0.342	0.986	0.949	0.445	0.431
SVC100	2.083	93.870	0.372	3.675	0.880	0.362	0.996	0.960	0.542	0.501
(D) Stratified										
GBA100-GS	2.677	93.392	0.453	3.478	0.857	0.430	0.995	0.961	0.589	0.569
RFAs100-GS	3.958	93.495	0.349	2.198	0.923	0.641	0.996	0.974	0.757	0.756

Abbreviations: Gradient Boosting with 100 features (GB100), all features (GBall) or AdaBoost (GBA); GBA with 100 structural features (GBAs100); Random Forest (RF); Logistic Regression (LR); Support Vector Machine (SVC); GBA100 on library-stratified datasets (GBA100-GS).

thresholds. GBA100 achieved a higher area under the curve, consistently outperforming all phase scoring methods except PHASIS, highlighting the enhanced ability of machine learning to differentiate between positives and negatives.

Minimizing false positives

Given the predominance of the negative class in the training set, a naive classifier is inclined to predict examples as negatives over positives. Therefore, we prioritized **precision** to minimize false positives, at the same time considering **MCC** to address class imbalance and enhance overall performance, along with runtime efficiency when selecting the final classifier. The trade-off between precision and MCC is illustrated in Figure S7. Phase scoring methods clustered in the lower-left corner of the plot, showing poor precision; They detected true positives but struggle with false positives, likely due to reliance on limited expression features. ShortStack had an MCC of 0.152, indicating marginal improvement over random guessing (MCC = 0), and a precision of 0.385, indicating many incorrect positive predictions. PHASIS, locating at the center of the plot, performed reasonably well with an MCC of 0.43 and a precision of 0.6 but may still be unsatisfactory. In contrast, all machine learning methods outperformed the scoring methods, with the GBA and RFA families leading in precision and MCC, as shown in the upper-right corner of the plot, suggesting superior predictive capabilities.

Versatile classifier for precursor and effector prediction

The **GBA100** model, using the 100 best overall attributes, achieved a precision of 0.857 and an MCC of 0.794. The **RFAs100** model, using 100 best structural attributes, offered a precision of 0.911 and an MCC of 0.766. These two models emerged from the GBA and RFA families as the leading candidates for final consideration. While both models performed strongly, RFAs100 excelled at detecting true positives, and GBA100 provided better overall performance. To facilitate the prediction of siRNA production sites for a given precursor, **RFAs100 is the optimal choice, offering a versatile model** that relies solely on static structural features without reliance on expression features.

Generalization across experiments

The *Arabidopsis* libraries spanning various tissues and treatments were sourced from distinct studies [Lei and Sun, 2014, Jeong et al., 2013]. To test model generalization across diverse experimental conditions, preparations, sequencing depths, and biological contexts, performance was evaluated on unseen libraries via stratified library sampling (Table S4, Figures S5 and S7). Compared to counterpart models trained on standard split datasets, GBA100-GS and RFAs100-GS models maintained strong precision and accuracy on library-stratified datasets, with a slight decrease in MCC. Similar performance between standard and stratified cross-validation highlights model robustness and ability to generalize.

Reliable prediction of conserved precursors

The **RFAs100** and GBA100 models showed strong reliability in identifying conserved precursors, as demonstrated by their substantial overlap with highly conserved reference contigs (Figure 1, left panel) and agreement with predictions from established methods (Figure 1, middle and right panels). This suggests that both models are reliable tools for genomic precursor predictions, producing results aligned with widely accepted benchmarks.

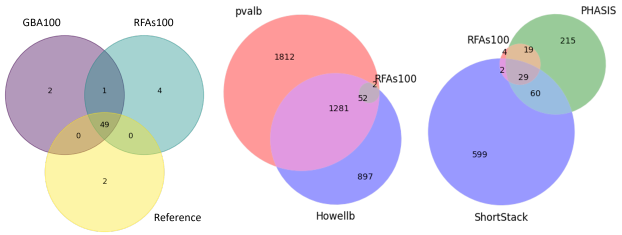


Fig. 1. Comparison of contigs predicted by siWalk RFAs100 classifier with reference contigs, GBA100 classifier, scoring methods Howellb and pvalb, and existing tools ShortStack and PHASIS. Background consists of 6083 contigs based on the training dataset; or ∼27000 contigs based on genome size (Table 1).

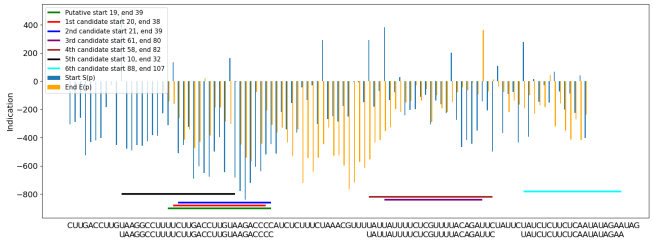


Fig. 2. Effector siRNA localization in an example precursor. Six candidates are identified by siWalk as the most likely effectors for a precursor (Chr3, positions 5862187-5862334), extracted and trimmed around the known ta-siRNA sequence UUCUUGACCUUGUAAGACCCC (positions 19-39). This ta-siRNA is derived from the TAS3 precursor PHAS21-21 [Chen et al., 2021].

Effector siRNA localization in precursor

We selected **RFAs100**, a machine learning model that can predict precursors without relying on expression-derived features. As most existing siRNA predictors rely on expression data and focus on precursors, we tested whether this model could locate effector production sites without such data. We developed a siRNA sorting process (see Section 2.3), powered by the **RFAs100** classifier and the indication score, to predict potential effectors directly from sequence information. We demonstrated it using the ta-siRNA sequence UUCUUGACCUUGUAAGACCCC [Chen et al., 2021, Fahlgren et al., 2006]. The genomic sequence surrounding the ta-siRNA was extracted, and siRNA candidates of various lengths at each position were evaluated using RFAs100, along with indication scores computed from $S(p)$ and $E(p)$ (Figure 2). Among the six identified effector candidates, three clustered near the known site, and two were positioned ∼20 nucleotides downstream. This process successfully pinpoints the approximate location of effectors, showcasing its ability to predict biologically relevant siRNA localization.

We assessed the offsets of the best (1st) candidates for each available positive precursor and an equal number of randomly selected negatives from the training dataset (Figure 3). To facilitate the analysis, each precursor was standardized to 121 nucleotides, with the start position of putative siRNAs centered at position 50 ± 5 for consistent comparison. SiRNAs located far from the center were excluded. This yielded 6386 positive examples from an initial 10683. The negative set exhibited a uniform distribution of offsets ranging from -50 to 50 nucleotides, and the positive set demonstrated a clear peak near zero between -10 to 10. The distributions between the

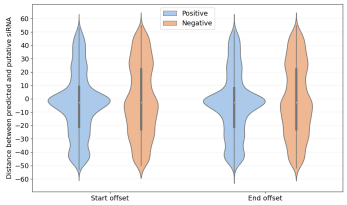


Fig. 3. Offset distribution of the 1st effector candidates for 6386 positive and negative precursors, each standardized to 121 nucleotides with putative siRNAs centered. Significant differences in the distributions of absolute offsets between positive and negative populations were confirmed by the Kruskal-Wallis test [Kruskal and Wallis, 1952]: Start offset $p = 2.97^{-67}$; End offset $p = 5.25^{-73}$.

two populations were significantly different, suggesting that the sorting process reliably identifies potential effector siRNAs.

The siWalk program

SiWalk is an open-source python program available on GitHub (https://github.com/bioinfoUQAM/siWalk). It offers modules for genome-wide precursor prediction, effector localization without expression data, and customized training using user-labeled data, as shown in Figure S11.

SiWalk can either use pre-trained models provided with the program or be trained on a user-labeled precursor dataset. Due to its 3.6 GB size, RFAs100 is not distributed with siWalk; instead, the instructions for model generation are provided. The GBA100 and GBAs100 models are available for reference. Optionally, users can label segment classes to create a training dataset, and a script is provided to train a custom model. Training on a dataset derived from 29 *Arabidopsis* libraries takes approximately 30 minutes on a single 3.2 GHz CPU.

In **prediction** mode, siWalk accepts SAM files as input, which are generated by aligning sRNA-seq libraries to a reference genome. The feature annotation stage takes 86 ± 101 minutes per library using 64 CPUs at 2.40 GHz with Spark parallel processing [Zaharia et al., 2016], based on processing 29 *Arabidopsis* SAM files. The precursor prediction stage takes around 1 minute per annotated *Arabidopsis* library. Although siWalk requires extra time for feature generation, it can be sped up by adding CPUs. SiWalk's robustness and generalization enable users to apply it to investigate genome-wide siRNA precursors expressed in their own sequencing datasets.

In **localization** mode, siWalk accepts a long nucleotide sequence as input and returns six siRNA candidates. Predicting siRNA from a 121-nt precursor takes approximately 8 minutes. Users can provide a precursor of any length to predict the localization of effectors, but note that processing time increases with precursor length since the scanning is conducted at each nucleotide.

Discussion

Predicting siRNAs from precursors is complex due to the difficulty of pinpointing precise Dicer cleavage sites and understanding how the active effector is selected by RISC. While the formation of double-stranded siRNA precursors is increasingly understood, the sequence and structural signals guiding these processes remain unclear. As portrayed in this study, siWalk is an accurate, reliable and generalized tool for the prediction of conserved siRNA precursors and for localizing

effectors within a sequence. SiWalk identifies potential siRNA precursors across the entire genome, enabling researchers to investigate siRNA expression patterns and explore regulatory mechanisms influenced by RNA interference. This facilitates the study of siRNAs' involvement in stress responses and other regulatory processes. By revealing influential features, siWalk facilitates biologists to hypothesize the relationship between biogenesis with feature importance and to uncover their evolutionary origins. These insights could inform the development of refined criteria for siRNA prediction, rooted in the biological significance of key features. The siRNA production sites identified by siWalk could also guide RNA interference vector design. This could enhance the amplification of desired siRNA thereby reducing off-target toxicity. On the other hand, its framework is tailored and benchmarked for regularly phased siRNAs, restricting its capacity to predict irregularly phased siRNAs, which arise from variations in precursor origins. It is possible to apply siWalk in other plant species, but its scope is constrained by the absence of comprehensive ground truth datasets similar to miRBase for miRNAs. Therefore, the training in this study has been limited to *Arabidopsis*, leaving its applicability to other plant species untested. Despite some limitations, accurate siRNA precursor and localization prediction by siWalk opens opportunities to drive advancements in identifying biomarkers of plant health and agricultural innovations, such as enhancing crop resilience and productivity.

Conflicts of interest

None declared.

Funding

Natural Sciences and Engineering Research Council of Canada, CGSD research grant to CJW and RGPIN Discovery Grants to ABD. Fonds de recherche du Québec - Nature et technologies, B2X research grant to CJW.

Data availability

Supplementary data are available at *Journal Name* online.

Acknowledgments

The authors would like to thank Armand Bandiang Massoua, Dr. Jean Danyluk, Dr. Amine M. Remita and Dr. Hayda Almeida for helpful discussions. This research was supported by the computational clusters and dedicated staff of The Digital Research Alliance of Canada.

References

N. Agrawal, P. Dasaradhi, A. Mohmmed, P. Malhotra, R. K. Bhatnagar, and S. K. Mukherjee. Rna interference: biology, mechanism, and applications. *Microbiology and molecular biology reviews*, 67(4):657–685, 2003.

E. Allen, Z. Xie, A. M. Gustafson, and J. C. Carrington. microrna-directed phasing during trans-acting sirna biogenesis in plants. *Cell*, 121(2):207–221, 2005.

M. J. Axtell. Classification and comparison of small rnas from plants. *Annual review of plant biology*, 64(1):137–159, 2013.

M. J. Axtell and B. C. Meyers. Revisiting criteria for plant microrna annotation in the era of big data. *The Plant Cell*, 30(2):272–284, 2018.

M. J. Axtell, C. Jan, R. Rajagopalan, and D. P. Bartel. A two-hit trigger for sirna biogenesis in plants. *Cell*, 127(3): 565–577, 2006.

D. Baulcombe. Rna silencing in plants. *Nature*, 431(7006): 356–363, 2004.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Y.-S. Chao, C.-J. Wu, Y.-C. Lai, H.-T. Hsu, Y.-P. Cheng, H.-C. Wu, S.-Y. Huang, and W.-C. Chen. Diagnostic accuracy of symptoms for an underlying disease: a simulation study. *Scientific Reports*, 12(1):13810, 2022.

C. Chen, J. Li, J. Feng, B. Liu, L. Feng, X. Yu, G. Li, J. Zhai, B. C. Meyers, and R. Xia. srnaanno—a database repository of uniformly annotated small rnas in plants. *Horticulture Research*, 8, 2021. URL `http://www.plantsrnas.org`. Accessed on February 20, 2024.

H.-M. Chen, Y.-H. Li, and S.-H. Wu. Bioinformatic prediction and experimental validation of a microrna-directed tandem trans-acting sirna cascade in arabidopsis. *Proceedings of the National Academy of Sciences*, 104(9):3318–3323, 2007.

H.-M. Chen, L.-T. Chen, K. Patel, Y.-H. Li, D. C. Baulcombe, and S.-H. Wu. 22-nucleotide rnas trigger secondary sirna biogenesis in plants. *Proceedings of the National Academy of Sciences*, 107(34):15269–15274, 2010.

X. Dai, Z. Zhuang, and P. X. Zhao. psrnatarget: a plant small rna target analysis server (2017 release). *Nucleic acids research*, 46(W1):W49–W54, 2018.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.

E. De Paoli, A. Dorantes-Acosta, J. Zhai, M. Accerbi, D.-H. Jeong, S. Park, B. C. Meyers, R. A. Jorgensen, and P. J. Green. Distinct extremely abundant sirnas associated with cosuppression in petunia. *Rna*, 15(11):1965–1970, 2009.

M. C. Dotto, K. A. Petsch, M. J. Aukerman, M. Beatty, M. Hammell, and M. C. Timmermans. Genome-wide analysis of leafbladeless1-regulated and phased small rnas underscores the importance of the tas3 ta-sirna pathway to maize development. *PLoS genetics*, 10(12):e1004826, 2014.

N. Fahlgren, T. A. Montgomery, M. D. Howell, E. Allen, S. K. Dvorak, A. L. Alexander, and J. C. Carrington. Regulation of auxin response factor3 by tas3 ta-sirna affects developmental timing and patterning in arabidopsis. *Current biology*, 16 (9):939–944, 2006.

T. A. Farazi, S. A. Juranek, and T. Tuschl. The growing catalog of small rnas and their association with distinct argonaute/piwi family members. *Development*, 135(7): 1201–1214, 2008.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

A. Fukudome and T. Fukuhara. Plant dicer-like proteins: double-stranded rna-cleaving enzymes for small rna biogenesis. *Journal of Plant Research*, 130:33–44, 2017.

Q. Guo, X. Qu, and W. Jin. Phasetank: genome-wide computational identification of phasirnas and their regulatory cascades. *Bioinformatics*, 31(2):284–286, 2014.

M. D. Howell, N. Fahlgren, E. J. Chapman, J. S. Cumbie, C. M. Sullivan, S. A. Givan, K. D. Kasschau, and J. C. Carrington. Genome-wide analysis of the rna-dependent rna polymerase6/dicer-like4 pathway in arabidopsis reveals dependency on mirna-and tasirna-directed targeting. *The Plant Cell*, 19(3):926–942, 2007.

D.-H. Jeong, S. R. Thatcher, R. S. Brown, J. Zhai, S. Park, L. A. Rymarquis, B. C. Meyers, and P. J. Green. Comprehensive investigation of micrornas enhanced by analysis of sequence variants, expression patterns, argonaute loading, and target cleavage. *Plant physiology*, 162(3): 1225–1245, 2013.

B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. Human microrna targets. *PLoS biology*, 2(11): e363, 2004.

N. R. Johnson, J. M. Yeoh, C. Coruh, and M. J. Axtell. Improved placement of multi-mapping small rnas. *G3: Genes, Genomes, Genetics*, 6(7):2103–2111, 2016.

M. W. Jones-Rhoades and D. P. Bartel. Computational identification of plant micrornas and their targets, including a stress-induced mirna. *Molecular cell*, 14(6):787–799, 2004.

A. Kakrana, P. Li, P. Patel, R. Hammond, D. Anand, S. M. Mathioni, and B. C. Meyers. Phasis: a computational suite for de novo discovery and characterization of phased, sirna-generating loci and their mirna triggers. *BioRxiv*, page 158832, 2017.

A. Kakrana, S. M. Mathioni, K. Huang, R. Hammond, L. Vandivier, P. Patel, S. Arikit, O. Shevchenko, A. E. Harkess, B. Kingham, et al. Plant 24-nt reproductive phasirnas from intramolecular duplex mrnas in diverse monocots. *Genome research*, 28(9):1333–1344, 2018.

B. Khraiwesh, J.-K. Zhu, and J. Zhu. Role of mirnas and sirnas in biotic and abiotic stress responses of plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(2):137–148, 2012.

D.-H. Kim, M. A. Behlke, S. D. Rose, M.-S. Chang, S. Choi, and J. J. Rossi. Synthetic dsrna dicer substrates enhance rnai potency and efficacy. *Nature biotechnology*, 23(2):222, 2005.

D. Kornbrot. Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*, 2014.

A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones. mirbase: from microrna sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019. URL `https://mirbase.org`. Accessed on May 7, 2024.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3): R25, 2009.

M. Leclercq, A. B. Diallo, and M. Blanchette. Computational prediction of the localization of microrna within their pre-mirna. *Nucleic acids research*, 41(15):7200–7211, 2013.

J. Lei and Y. Sun. mir-prefer: an accurate, fast and easy-to-use plant mirna prediction tool using small rna-seq data. *Bioinformatics*, 30(19):2837–2839, 2014.

S. Li, C. Castillo-González, B. Yu, and X. Zhang. The functions of plant small rna s in development and in stress responses. *The Plant Journal*, 90(4):654–670, 2017.

Y. Liu, C. Teng, R. Xia, and B. C. Meyers. Phasirnas in plants: their biogenesis, genic sources, and roles in stress responses, development, and reproduction. *Plant Cell*, 32(10):3059–3080, 2020.

C. Llave. Virus-derived small interfering rnas at the core of plant–virus interactions. *Trends in plant science*, 15(12): 701–707, 2010.

R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, Nov 2011.

S. Lück, T. Kreszies, M. Strickert, P. Schweizer, M. Kuhlmann, and D. Douchkov. sirna-finder (si-fi) software for rnai-target design and off-target prediction. *Frontiers in Plant Science*, 10:1023, 2019.

A. Lukasik, M. Wójcikowski, and P. Zielenkiewicz. Tools4mirs–one place to gather all the tools for mirna analysis. *Bioinformatics*, 32(17):2722–2724, 2016.

M. L. McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

B. C. Meyers, M. J. Axtell, B. Bartel, D. P. Bartel, D. Baulcombe, J. L. Bowman, X. Cao, J. C. Carrington, X. Chen, P. J. Green, et al. Criteria for annotation of plant micrornas. *The Plant Cell*, 20(12):3186–3190, 2008.

M. Nakano, K. McCormick, C. Demirci, F. Demirci, S. G. R. Gurazada, D. Ramachandruni, A. Dusia, J. A. Rothhaupt, and B. C. Meyers. Next-generation sequence databases: Rna and genomic informatics resources for plants. *Plant physiology*, 182(1):136–146, 2020. URL `https://mpss.danforthcenter.org`. Accessed on February 27, 2024.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Piriyapongsa and I. K. Jordan. Dual coding of sirnas and mirnas by plant transposable elements. *Rna*, 14(5):814–821, 2008.

G. Rorbach, O. Unold, and B. M. Konopka. Distinguishing mirtrons from canonical mirnas with data exploration and machine learning methods. *Scientific reports*, 8(1):7560, 2018.

A. Takeda, S. Iwasaki, T. Watanabe, M. Utsumi, and Y. Watanabe. The mechanism selecting the guide strand from small rna duplexes is different among argonaute proteins. *Plant and cell physiology*, 49(4):493–500, 2008.

H. Vaucheret and O. Voinnet. The plant sirna landscape. *The Plant Cell*, 36(2):246–275, 2024.

C.-J. Wu, A. M. Remita, and A. B. Diallo. Mirlibspark: a scalable ngs plant microrna prediction pipeline for multi-library functional annotation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 669–674, 2019.

H.-F. Yu, F.-L. Huang, and C.-J. Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85:41–75, 2011.

M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

C. Zhang, G. Li, J. Wang, and J. Fang. Identification of trans-acting sirnas and their regulatory cascades in grapevine. *Bioinformatics*, 28(20):2561–2568, 2012.

H. Zhang and J.-K. Zhu. Rna-directed dna methylation. *Current opinion in plant biology*, 14(2):142–147, 2011.

Z. Zhang, S. Teotia, J. Tang, and G. Tang. Perspectives on micrornas and phased small interfering rnas in maize (zea mays l.): functions and big impact on agronomic traits enhancement. *Plants*, 8(6):170, 2019.

Y. Zheng, S. Wang, and R. Sunkar. Genome-wide discovery and analysis of phased small interfering rnas in chinese sacred lotus. *PloS one*, 9(12):e113790, 2014.

Wu and Diallo, Improving siRNA Prediction in Arabidopsis with Machine Learning

Appendix: supplementary data

Alignment of phasing siRNAs on precursors

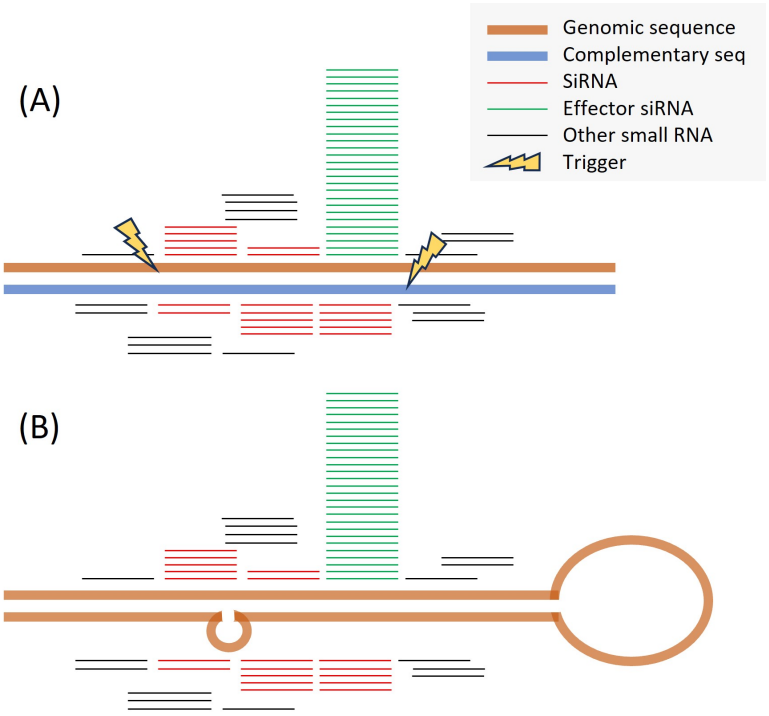


Fig. S1. Phasing patterns of small RNAs aligned to genomic sequences that produce dsRNA precursors. Duplexes align periodically at regular intervals and arise from two contexts: (A) complementary genomic strands, and (B) a folded single-stranded RNA derived from one genomic strand. Other RNAs may align randomly. Duplex production may be triggered by signals such as miRNA-mediated cleavage. Phase scores quantify phasing patterns by measuring the periodic alignment of small RNAs.

Phase scores

Phase scores quantify phasing potential by measuring RNA alignment patterns from Dicer's cyclical cleavage, which produces siRNA clusters with regularly spaced duplexes along the precursor (Figure S1). The default search window for a phasing segment spans 169 nucleotides, corresponding to 9 cleavage **cycles** each representing a 21-nucleotide (nt) interval known as **DicerCall** $= \{18, 19, \ldots, 31\}$. Reads from sense and antisense strands are combined by consolidating their start coordinates, accounting for the 2-nt offset of the antisense strand in the duplex. To compute a phase score, the following parameters are considered:

- **k** the number of phase positions expressed by 21-nt RNAs,
- **n** the number of all positions expressed by 21-nt RNAs,
- **N** the number of all positions expressed by small RNAs of any length,
- **p** the number of reads for 21-nt RNAs at phase positions,
- **u** the number of reads for 21-nt RNAs at all positions,
- **U** the number of reads for small RNAs of any length, and
- **maxf** the number of reads for the most abundant 21-nt RNAs at phase positions, including those shifted by 1 or 2 positions to account for phase drift.
- **effseq** (**eff_seq**) or putative effector siRNA, is the RNA sequence at the maxf position.

Existing scoring methods either evaluate phasing strength by assessing the accumulation of phased reads (e.g., Howell [Howell et al., 2007], Guo [Guo et al., 2014]) or determine phasing significance by comparing occupied phase positions to all positions (e.g., Chen p-value [Chen et al., 2007]). The scores and parameters we investigated include $Howell(\mathbf{p}, \mathbf{u}, \mathbf{k})$ [Chen et al., 2021]

and variant $Howellb(\mathbf{p}, \mathbf{k})$ [Howell et al., 2007], $Guo(\mathbf{u}, \mathbf{k}, \mathbf{p}, \mathbf{maxf})$ [Guo et al., 2014] and variant $Guob(\mathbf{U}, \mathbf{k}, \mathbf{p}, \mathbf{maxf})$, as well as $pval(\mathbf{k}, \mathbf{n})$ [Chen et al., 2007, 2021] and variant $pvalb(\mathbf{k}, \mathbf{N})$.

$$\text{Howell}(p, u, k) = \begin{cases} 0, & \text{if } k < 3, \\ \log\left[\left(1 + 10 \cdot \frac{p}{1+u}\right)^{k-2}\right], & \text{if } k \geq 3. \end{cases} \tag{3}$$

$$\text{Howellb}(p, k) = \begin{cases} 0, & \text{if } k < 3, \\ \log\left[(1+p)^{k-2}\right], & \text{if } k \geq 3. \end{cases} \tag{4}$$

$$\text{pval}(k, n) = \begin{cases} 1, & \text{if } k < 3, \\ \sum_{i=k}^{m} \frac{\binom{20m}{n-i} \cdot \binom{m}{i}}{\binom{21m}{n}}, & \text{if } k \geq 3 \text{ and } n \geq i, \end{cases} \tag{5}$$

where $m = (2c - 1)$ is a constant if cycle c is given.

$$\text{pvalb}(k, N) = \text{pval}(k, N) \tag{6}$$

$$\text{Guo}(u, k, p, \text{maxf}) = \begin{cases} 0, & \text{if } k < 3, \\ \left(\frac{1}{u}\right) \cdot k \cdot \ln(p) \cdot \text{maxf}, & \text{if } k \geq 3, \end{cases} \tag{7}$$

$$\text{Guob}(U, k, p, \text{maxf}) = \text{Guo}(U, k, p, \text{maxf}) \tag{8}$$

Algorithms for determining phasing boundaries

Figure S2 illustrates how phasing boundaries are determined and the core steps of the implementation. An in-phase position is at a multiple of the *DicerCall* interval from a starting position *pos*. A phased position is an in-phase position occupied by RNAs of *DicerCall* length. The search window for a phasing segment spans from *pos* to $pos + DicerCall \times (cycle - 1)$. The booleans *pos.watson* and $(pos+2).crick$ denote the RNA duplex counterparts of *DicerCall* length expressed on the Watson or Crick strand. List functions MIN, MAX and UNIQUE return the minimum, maximum and unique values in a list. LENGTH returns the length of the list. The default search window spans 169 nucleotides, corresponding to 9 Dicer cleavage *cycles*, each spacing 21-nt known as $DicerCall = \{18, 19, \ldots, 31\}$.

After identifying phased positions (Algorithm 1) and establishing locus boundaries (Algorithm 2), a function computes attributes (e.g., nucleotide frequency, folding mode), which are stored in the segment object. Segments with shared effector siRNAs are aggregated (Algorithm 3).

Algorithm 1 Get Phased Positions within a Window

```
1: function GetPhasedPositions(pos, DicerCall, cycle): list of integers
2:     Initialize an empty list L and a positive integer posNext
3:     for i in {0, 1, 2, .., cycle − 1} do
4:         posNext ← pos + DicerCall × i
5:         if posNext.watson or (posNext + 2).crick then
6:             Append posNext to L
7:         end if
8:     end for
9:     return L
10: end function
```

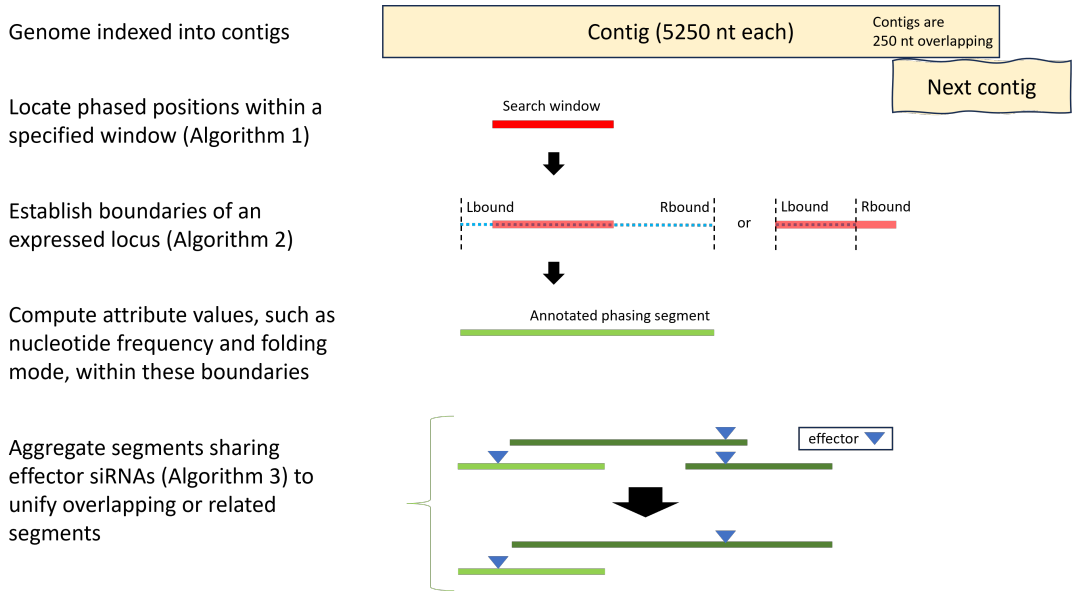


Fig. S2. Illustration of algorithms for identifying phasing event boundaries. A phasing segment is identified through: Labeling the source contig, Identifying expressed phasing positions within an initial search window, Refining boundaries by searching nearby positions, Computing features to annotate phasing segments, and Aggregating segments that share effector siRNAs.

Algorithm 2 Get Boundaries of An Expressed Locus

function GetBoundaries(pos, $DicerCall$, $cycle$): 2 integers
  Initialize $Lbound = 0$, $Rbound = 0$ and $posTmp = 0$
  Initialize empty lists L and L_0
  ▷ Step 1: Collect expressed phase positions from pos
  $L \leftarrow$ GetPhasedPositions(pos, $DicerCall$, $cycle$)
  ▷ Step 2: Scan nearby in-phase positions
  if Length(L) > 1 **then**
    for $direction$ in {1, -1} **do** ▷ 1 for downstream, -1 for upstream
      $posTmp \leftarrow pos + DicerCall \times cycle \times direction$
      $L_0 \leftarrow$ GetPhasedPositions($posTmp$, $DicerCall$, $cycle$)
      while Length(L_0) > 1 **do**
        Append L_0 to L
        $posTmp \leftarrow posTmp + DicerCall \times cycle \times direction$
        $L_0 \leftarrow$ GetPhasedPositions($posTmp$, $DicerCall$, $cycle$)
      end while
    end for
    ▷ Step 3: Determine boundaries
    $Lbound \leftarrow$ Min(L) $-DicerCall$
    $Rbound \leftarrow$ Max(L) $+DicerCall$
  end if
  return Max(0, $Lbound$), $Rbound$ ▷ Return adjusted Lbound and Rbound
end function

Algorithm 3 Aggregate Segments

function SOFT_AGGREGATE(G: list of segment objects): list of segment objects
  Initialize empty lists L, T and M
  $L \leftarrow$ Unique($segment.effpos$ in G)
  for $effpos$ in L **do**
    Create a new segment object S with $S.effpos \leftarrow effpos$
    $T \leftarrow$ segments in G where $segment.effpos = effpos$
    $S.Lbound \leftarrow$ Min($Lbound$ of segments in T)
    $S.Rbound \leftarrow$ Max($Rbound$ of segments in T)
    ▷ Aggregate remaining attributes with specific rules
    for all remaining attributes in segment **do**
      $S.attribute \leftarrow$ Min/Max(attribute values in T)
    end for
    Append S to M
  end for
  return M
end function

Reference contigs

Table S1. Example of a reference contig. A standardized genomic contig represents this Chromosome (Chr) 1 phasing locus, which has been assigned varying IDs and genomic coordinates in four different sources.

ID	Chr	Start	End	Length	Contig	Source
At1g62910	1	23299057	23300958	1901	1-23295000-23300250	[Chen et al., 2007, Howell et al., 2007]
PHAS21-7	1	23299479	23300850	1371	1-23295000-23300250	[Chen et al., 2021]
AT_TAIR10.1.23299482.23300923	1	23299482	23300923	1441	1-23295000-23300250	[Nakano et al., 2020]

Table S2. *Arabidopsis* literature-supported siRNA-generating contigs [Chen et al., 2007, Howell et al., 2007, Nakano et al., 2020, Chen et al., 2021]. (A) Complete list of reference contigs and (B) contigs of TAS (trans-acting siRNA) genes.

(A) Reference contig			
1–4180000-4185250	1–23410000-23415250	3–1955000-1960250	5–5460000-5465250
1–4295000-4300250	1–23415000-23420250	3–1965000-1970250	5–9785000-9790250
1–4350000-4355250	1–23450000-23455250	3–1970000-1975250	5–15555000-15560250
1–4365000-4370250	1–23485000-23490250	3–5860000-5865250	5–15550000-15555250
1–5295000-5300250	1–23490000-23495250	3–6390000-6395250	5–15755000-15760250
1–17885000-17890250	1–23505000-23510250	3–6915000-6920250	5–16635000-16640250
1–17890000-17895250	1–23585000-23590250	3–8525000-8530250	5–17560000-17565250
1–18545000-18550250	1–25275000-25280250	3–8530000-8535250	5–17565000-17570250
1–23175000-23180250		3–9415000-9420250	5–20130000-20135250
1–23205000-23210250		3–9870000-9875250	5–23390000-23395250
1–23200000-23205250		3–9865000-9870250	5–24305000-24310250
1–23295000-23300250			
1–23300000-23305250	2–7190000-7195250	4–5760000-5765250	
1–23305000-23310250	2–11720000-11725250	4–8145000-8150250	
1–23385000-23390250	2–16535000-16540250	4–8380000-8385250	
(B) Contig of TAS			
1–18545000-18550250	2–11720000-11725250	3–5860000-5865250	5–20130000-20135250
	2–16535000-16540250	3–9415000-9420250	5–23390000-23395250

Notes. All identified loci correspond to 21-nt Dicer call. No phasing locus was found in mitochondrial or chloroplast genomes. Refer to file **ath_ref_contigs.xlsx** for data source mapping.

Classifier hyper-parameters, performance, and time efficiency

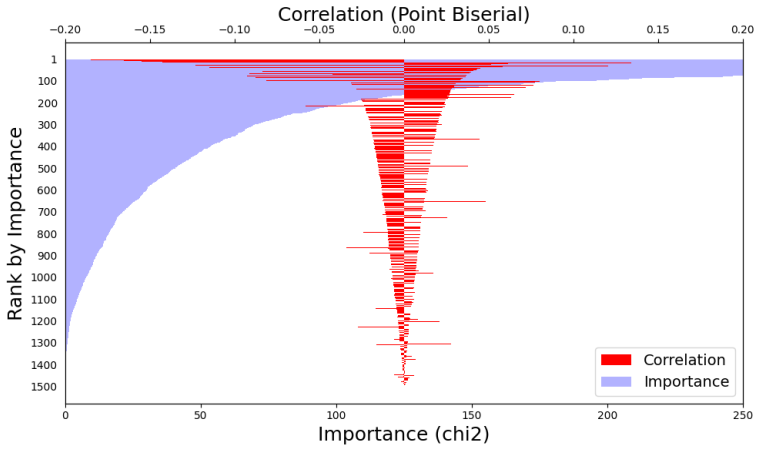


Fig. S3. Attributes ranked by importance and their correlation with precursor prediction.

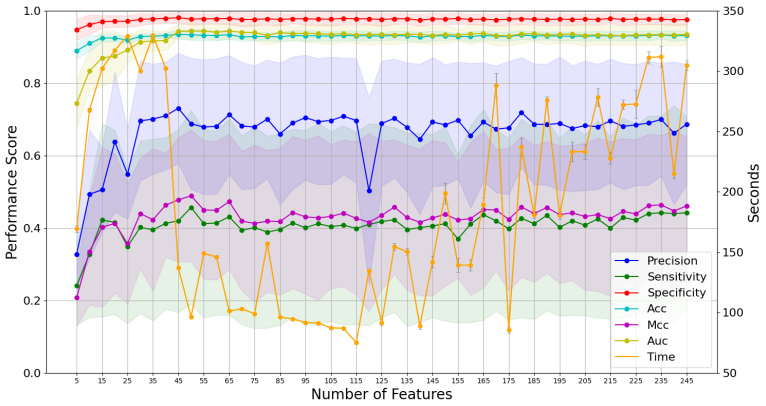


Fig. S4. Relationship between the number of features and classifier performance. The Gradient Boosting classifier with AdaBoost was evaluated on a reduced training dataset (10093 examples) using 5-fold cross-validation to assess performance of leading features at five intervals of increasing feature counts.

Table S3. Hyper-parameters of classifiers optimized for precision using Grid Search.

Classifier	Parameters
Gradient Boosting	learning_rate = 0.2, max_features = log2, min_samples_leaf = 4, and n_estimators = 200
Random Forest	max_depth = 30, max_features = log2, min_samples_leaf = 6, and n_estimators = 200
Logistic Regression	C = 0.1, max_iter = 10000, and solver = sag
Support Vector Machine	C = 0.01, degree = 4, gamma = 0.1, and kernel = poly

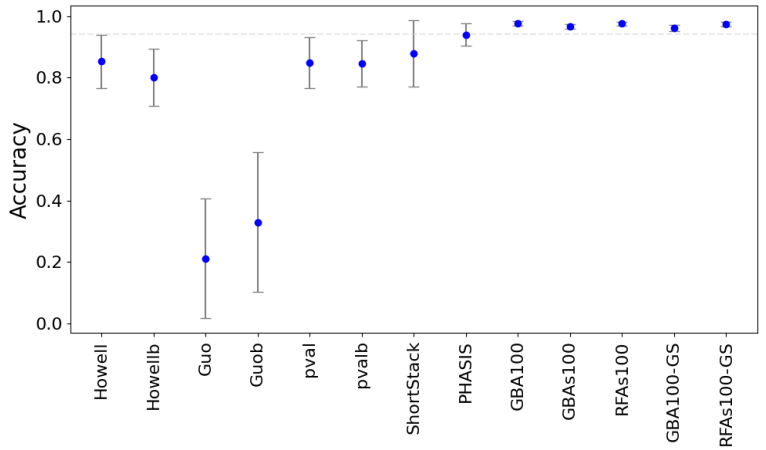


Fig. S5. Comparison of accuracies of selected classifiers with random guessing (0.9424). Since negative examples make up 94.24% in the dataset, a naive classifier that always predicts negative would achieve an accuracy of 0.9424.

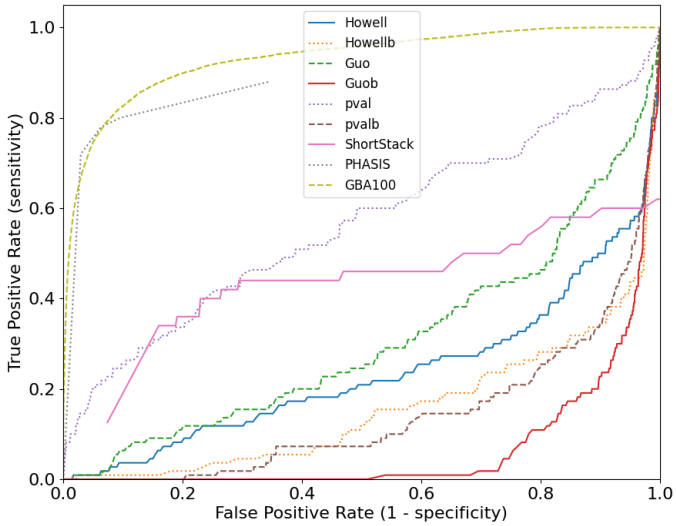


Fig. S6. ROC curves (Receiver-Operating Characteristic) of selected classifiers predicting from sRNA-seq libraries, showing their ability to distinguish true positives from false positives.

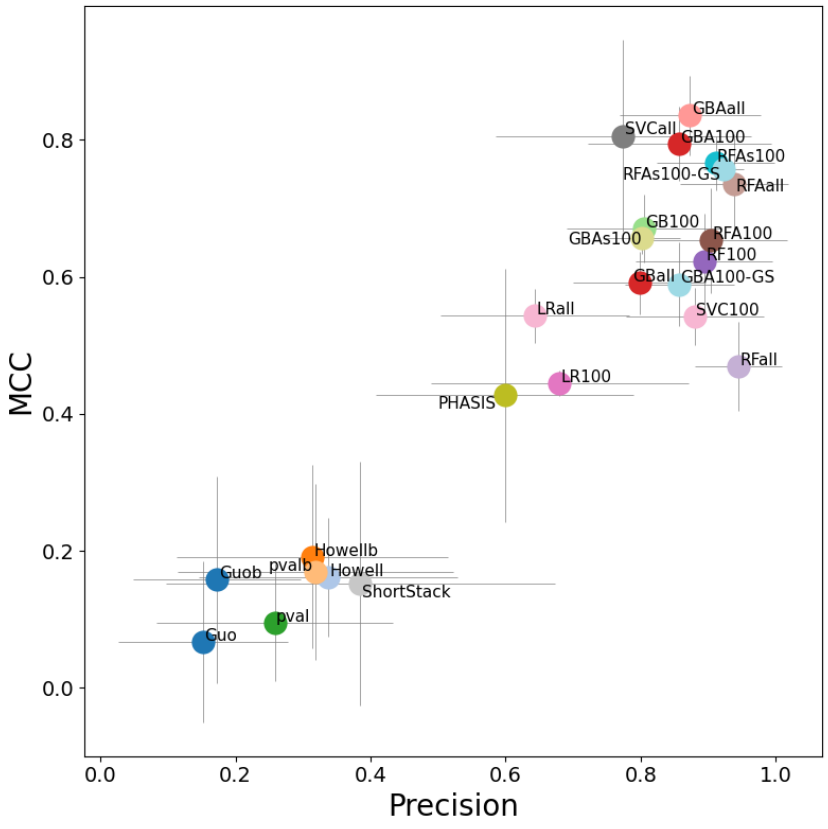


Fig. S7. Comparison of Precision and MCC among classifiers.

Table S4. Performance between standard and stratified split strategies. Extracted from Table 3. *GS*: training on library-stratified datasets.

	TP%	TN%	FP%	FN%	Precision	Sensitivity	Accuracy	MCC
GBA100	4.460	93.225	1.017	1.298	0.857	0.775	0.977	0.794
GBA100-GS	2.677	93.392	0.453	3.478	0.857	0.430	0.961	0.589
RFAs100	3.840	93.806	0.436	1.918	0.911	0.667	0.976	0.766
RFAs100-GS	3.958	93.495	0.349	2.198	0.923	0.641	0.974	0.757

Table S5. Time efficiency of various prediction methods. (A) Total time spent on Grid Search with 5-fold cross-validation for a dataset of 185539 examples. Using a subset of 100 attributes often reduces computation time compared to using the full set with the same classifier. (B) Average prediction time per library across 29 libraries. RFAs100, which uses 100 structural features, processes libraries more slowly than ShortStack and PHASIS due to its larger feature count and the additional time required for feature generation.

(A) Grid Search	Total (minutes)	Std Dev
GB100	2.2	na
GBall	7.7	na
GBAs100	34.7	na
GBAall	62.2	na
RF100	7.4	na
RFall	6.4	na
RFAs100	165.7	na
RFAall	303.7	na
LR100	0.5	na
LRall	16.3	na
SVC100	265.7	na
SVCall	1900.7	na
(B) Per Lib	Average (minutes)	
ShortStack	12.5	6.9
PHASIS	0.4	nd
RFAs100	86.-	101.-

Abbreviations *na*: not applicable (total time, run once); *nd*: not determined; Gradient Boosting with 100 features (GB100), all features (GBall) and AdaBoost (GBA); GBA with 100 structural features (GBAs100); Random Forest (RF); Logistic Regression (LR); Support Vector Machine (SVC).

Illustrations of more effector prediction examples

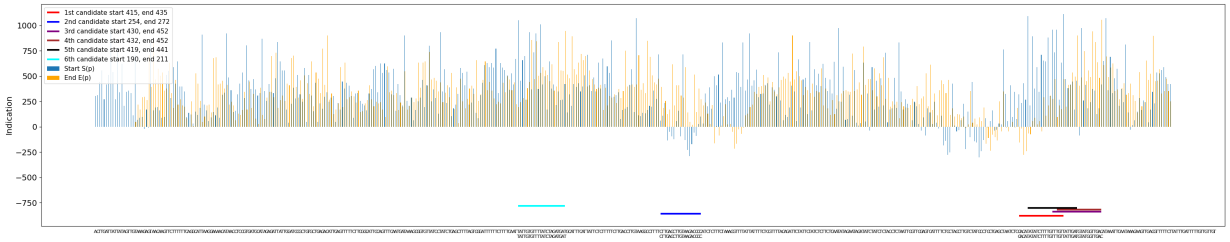


Fig. S8. Effector localization analysis of the full-length **PHAS21-21** precursor. This precursor is located on Chr3 (positions 5861956–5862442, length 487 [Chen et al., 2021]) and is a TAS3 precursor that produces siRNAs to regulate plant auxin signaling by targeting ARFs [Fahlgren et al., 2006]. The known effector tasiARF_1 (sequence: TTCTTGACCTTGTAAGGCCTT) is localized between positions 232 and 252, but was not detected by the sorting process. The known localization of tasiARF_2 (sequence: TTCTTGACCTTGTAAGACCCC), between positions 253 and 273, was identified near the second candidate in the analysis.

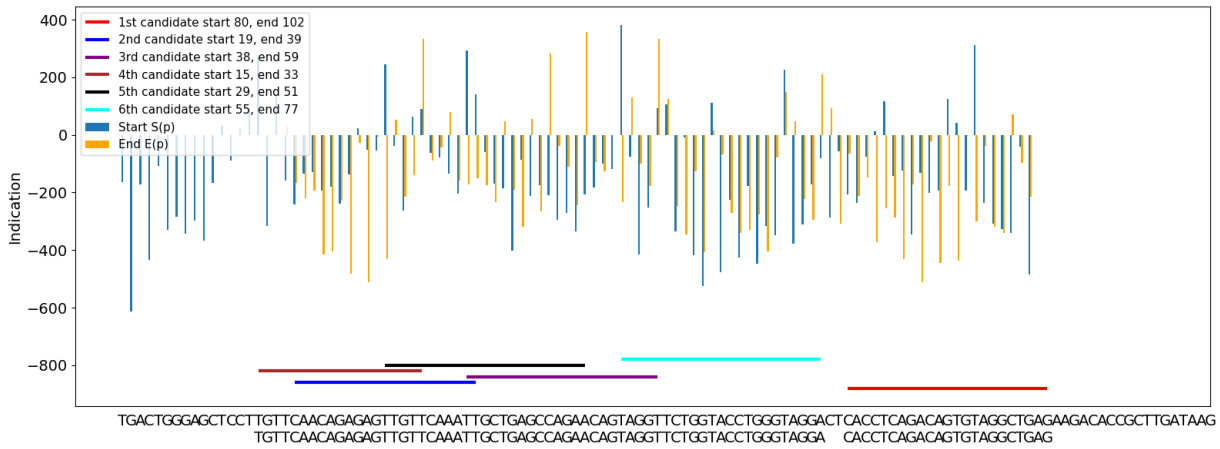


Fig. S9. An example randomly selected from the positive training set. The precursor sequence (Chr1, positions 17891352-17891519) was extracted and trimmed around the putative effector siRNA, located between positions 79 and 99.

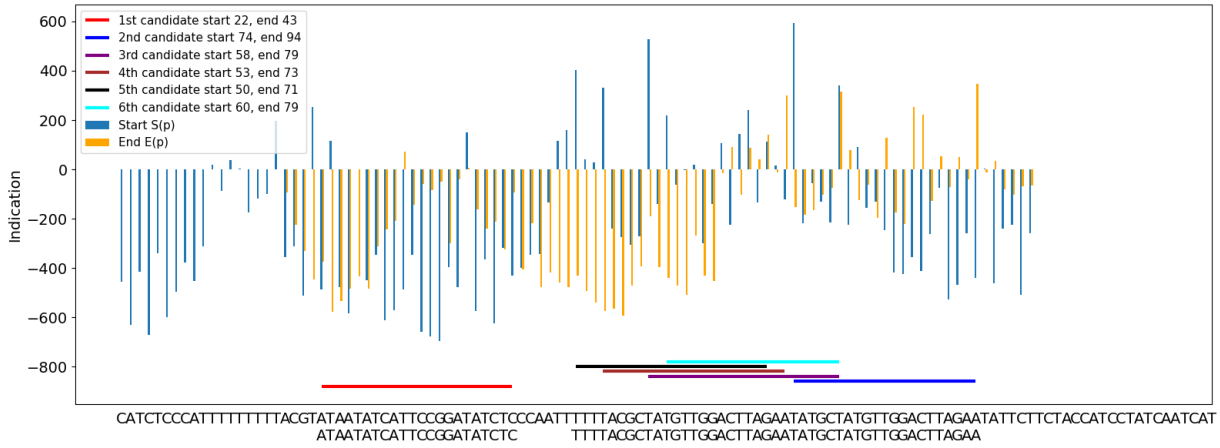


Fig. S10. Another example randomly selected from the positive training set. The precursor sequence (Chr1, positions 18549331-18549709) was extracted and trimmed around the putative effector siRNA, located between positions 50 and 70.

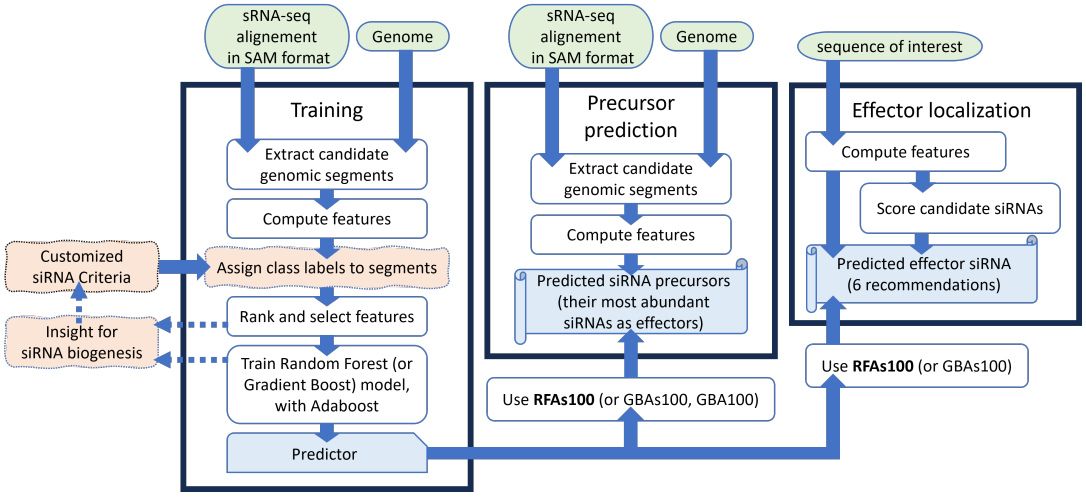


Fig. S11. Workflow of the siWalk algorithm. Three modules are provided. **Precursor prediction**: genome-wide prediction of precursors and assigning effectors. **Effector localization**: predicting effector localization in a sequence. **Training**: training a custom model with user-labeled siRNA examples. The miRBase as an input for computing features is not shown. Wiggled boxes indicate operations during training that could be performed by the user.

Feature definitions and complete list

Feature value space and distribution

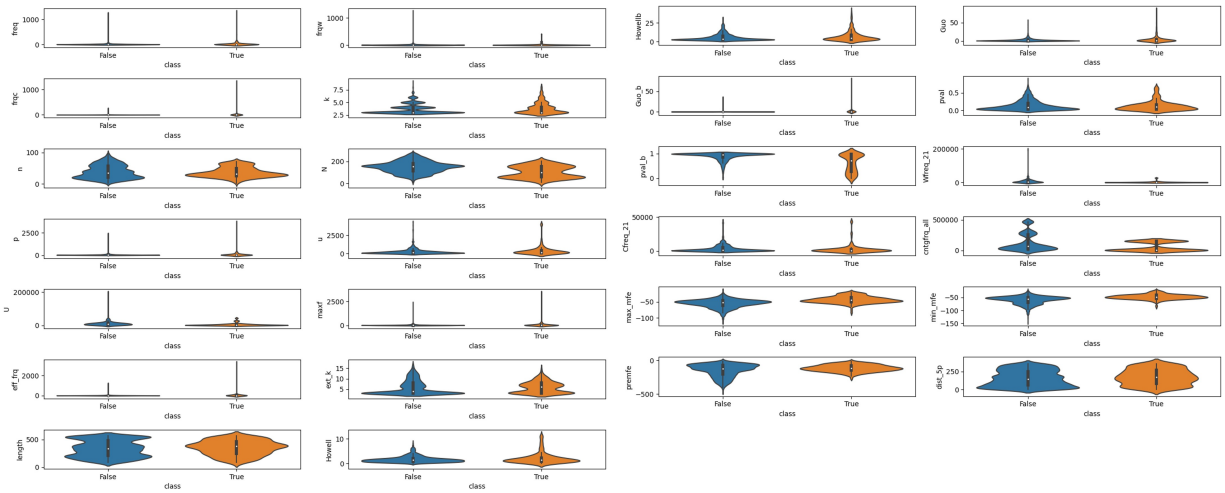


Fig. S12. Distribution of selected numerical features. Violin plots illustrate the value density of each feature within true and false classes of phasing segments in the training dataset.

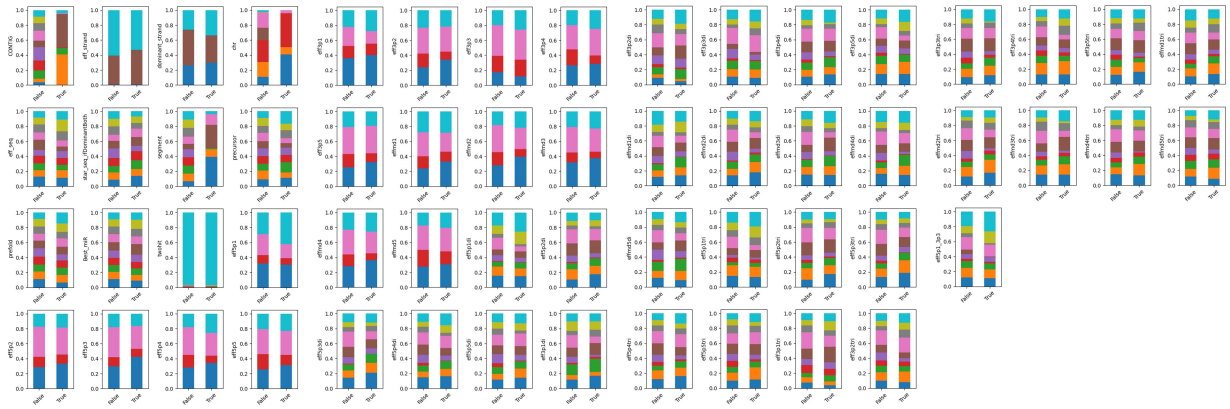


Fig. S13. Distribution of selected categorical features. Stacked bar charts display the value space and corresponding percentages for each feature within true and false classes of phasing segments in the training dataset.

Feature to attribute

To ensure comparability across libraries, numerical features were discretized with QuantileTransformer, and categorical features were encoded with OneHotEncoder (scikit-learn). QuantileTransformer scales numerical features to a uniform distribution, assigning a numerical value to each resulting attribute. OneHotEncoder transforms categorical features by creating binary (true/false) attributes for each unique category, which can expand a single categorical feature into multiple attributes. For example,

- the categorical feature **eff3p3**, with possible values **A**, **T**, **C**, and **G**, was one-hot encoded into four binary attributes: **eff3p3_A**, **eff3p3_T**, **eff3p3_C**, and **eff3p3_G**, each taking **True** or **False** values.
- the categorical feature **md3C**, with two values (**True** or **False**), was one-hot encoded into a single binary attribute **md3C**.

Feature naming

Figure S14 provides a guide for the naming of those features related to positional attributes. Attributes prefixed with **eff** denote structural features that specify the base type. Conversely, unprefixed attributes (e.g., **md3C**) represent expression features, capturing the expression level associated with the respective base type.

Example attribute names:

- **md3C**: The 12th position of the siRNA is Cytosine.
- **eff5p1di_TA**: The di-nucleotide starting at position 1 is Thymidine-Adenosine.
- **eff5p1_3p3_TG**: The first nucleotide is Thymidine, and the last is Guanine.
- **5p4C**: The 4th position is Cytosine.
- **3p4C**: The 18th position (21-mer), 19th (22-mer), 20th (23-mer), or 21st (24-mer) is Cytosine.
- **eff3p3_G**: The 19th position (21-mer), 20th (22-mer), 21st (23-mer), or 22nd (24-mer) is Guanine.

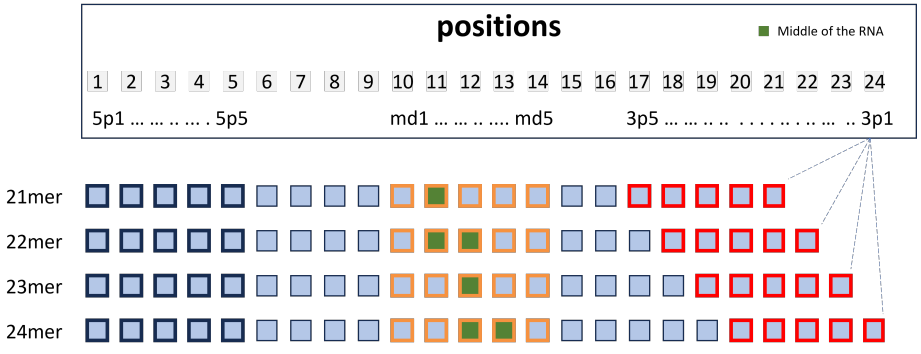


Fig. S14. Illustration of feature naming for positional attributes.

Feature definitions

Table S6. SiWalk feature summary. The 245 raw features are categorized by their expression-dependent and structural properties. The numbers of features considered for precursor prediction and effector localization are specified.

Feature property	Precursor prediction	Effector localization	Description
Expression:			
Phase score	14	0	Howell, Chen, and Guo scores, and phasing-related expression statistics
Ratio	1	0	Proportion of forward reads relative to the total reads
Nucleotide frequency	65	0	Expression-weighted nucleotide total frequency of the precursor or positional frequency on the effector
Abundance	2	0	Coordinate of the effector, the most abundant siRNA allowing for drift
Structural:			
Length	3	2	Precursor length and relative position of the effector siRNA on the precursor
Nucleotide type	46	46	Nucleotide composition at specific positions on the effector (as mono-di-, or tri-nucleotides)
Base pairing	5	5	Percentage of bases forming base pairs around the effector (-4 to +4 from start and end of the effector on the precursor)
MFE	5	1	MFE of precursors at varying lengths
Triplet	44	44	Frequency of effector triplet folding mode
Hairpin	9	3	Hairpin-forming potential of the precursor
Bulge	25	25	Bulge length around the effector
Loop	23	23	Loop length around the effector
MiRNA targeting	3	3	Number and score of miRNA targeting the precursor

Table S7 (see file complete_feature_list.tsv) shows the complete feature list. To outline their functions, these attributes were grouped by **Property1** into 111 **expression attributes** and 1391 **structural attributes**, and were further categorized by **Property2**.

Rank scores (feature importance) were determined by the chi-squared test. Correlation between a variable and a binary classification was assessed using the Point-Biserial Correlation. The Usage column records which features were incorporated into precursor **pred**ition or siRNA localization **vali**dation tasks.

Table S7. See file complete_feature_list.tsv

bioinfoUQAM / siWalk

Code Issues Pull requests Actions Projects Security Insights Settings

MIT license

0 stars 0 forks 4 watching Branches Activity Custom properties Tags

Private repository

main 1 Branch 0 Tags Go to file t Go to file Add file Code

JulieCJWu Delete src/barplot_indication.py 086038a · 2 months ago

dbs	Delete dbs/test	2 months ago
lib	Delete lib/test	2 months ago
model	Delete model/test	2 months ago
src	Delete src/barplot_indication.py	2 months ago
workdir	Delete workdir/test	2 months ago
LICENSE	Add files via upload	2 months ago
README.md	Update README.md	2 months ago

README MIT license

siWalk

SiWalk is a machine learning-based tool that predicts conserved siRNA production in Arabidopsis using a Random Forest classifier, complementing established methods. It refines predictions by incorporating phase boundaries and a standardized genomic contig index. SiWalk identifies siRNA precursors and active siRNAs based on structural and nucleotide patterns, while improving accuracy and minimizing false positives.

Copyright (c) 2024 Chao-Jung Wu, see License.

Citation:

Please cite the paper for your work using this tool.

Title: Leveraging Machine Learning for Improved siRNA Prediction in Plant

Cite: Chao-Jung Wu, Abdoulaye Baniré Diallo

BibTex:

```
@article{... ...
}
```

Link to retrieve the paper: http...

Dependencies:

Languages: `python/3.10.2` (scipy, statsmodels, sklearn, numpy)

Packages: `spark/3.3.0` , `bowtie/1.3.0` , `samtools/1.17` , `viennarna/2.5.1` , `perl/5.30.1`

Distributed with siWalk: `miRanda` , `miRCheck`

Environments: Users should run `siWalk` on servers that support parallel computing with `Spark` . This program was developed on Compute Canada clusters, primarily on Narval.

Manual

Follow this manual to download and execute the siWalk project on a server. You can use Compute Canada clusters or any other server infrastructure.

M1. Download siWalk from GitHub

The siWalk project is available on GitHub at github/JulieCJWu/siWalk.

To access the release version associated with the paper, visit: siWalk/releases/241106testlargefile.

- Download the assets and install the source code.
- Place the file **background.tsv** in `/path/siWalk/dbs/` folder.
- Place the file **TAIR10_genomic.fa** in `/path/siWalk/dbs/TAIR10` folder.

M2. Generate the RFAs100 Model

3.1. Generate the **RFAs100** model (described in the paper but too large to be distributed with siWalk) using the following command. Note that this step takes approximately 30 minutes on a personal computer.

```
cd /path/siWalk/src/
annotated_training_file=../dbs/background.tsv
python siWalk_pickle_localization.py $annotated_training_file
```

3.2. Rename the resulting model file to `RFAs100.pkl` and move it to `/path/siWalk/model/` .

M3. Description of supporting files

Supporting files, including data for **Arabidopsis**, are bundled with siWalk. Use these files to quickly start analyzing Arabidopsis sRNA-seq data.

- **background.tsv**: This file served as the Arabidopsis training set in the paper and contains a total of 185539 segment samples.
- **TAIR10_genomic.fa**: The Arabidopsis genome is provided for your convenience.
- dbs/**mature.fa**: miRNAs from miRBase v22.1
- dbs/**GSM1087998_C0RC1.txt**: an example expression library in readcount format, sourced from the GSE44622 series
- dbs/**feature_definition.txt**: definitions of variables that may appear in the analysis
- model/**GBA100.pkl**: machine learning model using Gradient Boosting with AdaBoost, based on the top 100 features as described in the paper
- model/**GBAs100.pkl**: machine learning model using Gradient Boosting with AdaBoost, based on the top 100 **structural** features as described in the paper
- model/**RFAs100.pkl**: machine learning model using Random Forest with AdaBoost, based on the top 100 **structural** features as described in the paper. This file is generated by users following the steps in **M2**.
- model/**Arabidopsis_feature_importance_n_correlation.tsv**: the feature evaluation datafile to pair with `GBA100.pkl`
- model/**Arabidopsis_strcture_feature_importance_n_correlation.tsv**: the feature evaluation datafile to pair with `GBAs100.pkl` and `RFAs100.pkl`

M4. Overview of core scripts

- **submit_siWalk_precursor_features.sh**: see details in **A2**.
- **siWalk_classify_precursors.py**: see details in **A3**.
- **siWalk_predict_siRNA_location.py**: see details in **B1**.
- **siWalk_pickle_precursor.py**: see details in **X2**.
- **siWalk_pickle_localization.py**: see details in **X2**.

To support the prediction of precursors from expression libraries aligned in SAM format (**A1**), two distinct scripts are provided: one for computing precursor features (**A2**) and another for classifying precursors (**A3**). These steps remain separate to allow for custom precursor labeling before classification, providing greater flexibility in the process.

For effector prediction from a precursor, a script scans the precursor sequence to identify and report the top six predicted siRNA candidates (**B1**).

To support the creation of a machine learning model using a custom training set (**X2**), two scripts are provided: one for generating a model using all feature types, and another for generating a model using only the structural features.

M5. Run siWalk on a server

siWalk can be executed on various server infrastructures, such as Compute Canada or Microsoft Azure. The setup involves modifying the job submission script according to the server environment.

5a. Use Compute Canada clusters (e.g., Narval)

- **Modify the Submission File**: siWalk provides a template submission file for Narval on Compute Canada. To use it:
  - Update your **credentials** in the template.
  - Customize job parameters (e.g., CPUs, memory) based on your computational needs.
- **Submit the Job**:
  The `submit_siWalk_precursor_features.sh` script requires **SLURM** and **Spark**. After customizing the submission file, submit it using the SLURM scheduling command. See details in **A2**.

5b. Use a Login Node for simple tasks on Compute Canada clusters

The scripts `siWalk_classify_precursors.py`, `siWalk_predict_siRNA_location.py`, and `siWalk_pickle_precursor.py`, along with the steps to create a SAM file, are relatively simple tasks that can be executed on a personal computer with a properly set up environment, should you prefer that approach.

For running these tasks on a Compute Canada cluster, set up the environment as follows: First, nevigate to a subfolder of siWalk,

```
cd /path/siWalk/workdir/
```

Then,

```
module load StdEnv/2020 viennarna/2.5.1 gcc/9.3.0
module load blast+/2.14.0 samtools/1.17 bowtie/1.3.0 scipy-stack
chmod +x ../lib/miranda
virtualenv --no-download env
source env/bin/activate
pip install --no-index --upgrade pip
pip install --no-index statsmodels sklearn
```

5c. Use other servers

For servers like **Microsoft Azure**, follow their specific instructions for job submission. Although this has not been explicitly tested by the authors, the general steps are as follows:

- Upload the siWalk files to the server.
- Set up the environment by installing the necessary dependencies and configuring any required settings.
- Submit the job using Azure's task scheduling tools.

Module A. precursor prediction

A1. Create SAM file

siWalk requires **SAM** files as input for prediction. To create these alignment files, follow the steps below to align sRNA-seq libraries to the appropriate genome reference. While siWalk does not directly support this step, you can easily achieve it by:

1. Define the path to the `genome_file`.

2. Set the prefix for the genome using `index_prefix` .
3. Build `bowtie` genome reference index.
4. Create genome `.fai` index.

Example set up:

```
genome_file=/path/TAIR10_genomic.fa
index_prefix=/path/bowtie_index/TAIR10_genomic
```

Then run:

```
bowtie-build -f --quiet $genome_file $index_prefix
samtools faidx $genome_file
```

5. Prepare the library in `readcount` or `raw` format:
   - Ensure the library is in `readcount` format (each line contains a sequence and its occurrence) or `raw` format (each line represents an occurrence of a sequence).
   - Use the provided helper script to convert `readcount` to `raw` format.
6. Align the library using `bowtie` :
   - Align your sRNA-seq library to the genome using `bowtie` and output the results in `.sam` format.
   - Depending on the size of the genome and library, processing time may vary (e.g., 20 minutes for **Arabidopsis**, 20 hours for **wheat** using 64 CPUs).
7. Obtain associated sorted `.bam` and index `.bai` / `.csi` files.

Example set up:

```
path_siWalk=/path/siWalk
filename=GSM1087998_C0RC1
infile=$path_siWalk/dbs/$filename.txt # in readcount format
```

Then run:

```
mkdir $path_siWalk/input
raw=$path_siWalk/input/$filename.raw
python $path_siWalk/src/rc2raw.py $infile > $raw
samfile=$path_siWalk/input/$filename.sam
bamfile=$path_siWalk/input/sorted_$filename.bam
bowtie -r $raw -x $index_prefix --threads 64 -v 0 --mm -a --sam --no-unal > $samfile
samtools sort $samfile > $bamfile
samtools index $bamfile #\\ or \\ samtools index -c $bamfile # for large genome
```

A2. Generate features

8. Enter your credentials in the `submit_siWalk_precursor_features.sh` submission file.
   Edit the following variables in the file to specify the input path for the `.sam` , `.bam` , and `.bai` files, as well as the file path to the reference genome.

```
INPUT_SAM_PATH=/path/siWalk/input
genome_file=/path/TAIR10_genomic.fa
```

9. Submit the job to the queue from the `workdir` folder, and take note of the **job ID**.

```
cd /path/siWalk/workdir/
sbatch submit_siWalk_precursor_features.sh
```

10. Locate output files:
    - Output files, including the `annotation_file` of expression and structural features (`$filename.contig_features.tsv`), will be available in the folder `/path/siWalk/output/$jobID/$filename/` .

Check that the output files have been generated:

```
ls /path/siWalk/output/$jobID/$filename/
annotation_file=/path/siWalk/output/$jobID/$filename/$filename.contig_features.tsv
head -n 3 $annotation_file
```

A3. Predict precursor segments

11. Run precursor classification:
    - Use the script `siWalk_classify_precursors.py` to predict precursor segments by passing the `annotation_file` from the previous step.

```
cd /path/siWalk/src/
python siWalk_classify_precursors.py $annotation_file
```

12. Retrieve prediction results:
    - The output, including the predicted siRNA-generating loci, will be available as `prefix.prediction.tsv` in the same folder as the `annotation_file` .
    - Key columns include:
      - **CONTIG**: 5250-nt genomic region defined by coordinates (e.g., `3__5860000_5865250` on chromosome 3).
      - **eff_strand**: coordinate of the most abundant effector siRNA allowing drift.
      - **eff_seq**: sequence of the most abundant effector siRNA.
      - **segment**: predicted start and end positions of the siRNA-generating locus.
      - **Predicted_Class**: classification of the locus as an siRNA precursor (`true` or `false`) predicted by the model.

Check the output file:

```
file=/path/siWalk/output/$jobID/$filename/*prediction.tsv
head -n 3 $file
```

Module B. effector prediction from a precursor

B1. Provide your precursor sequence, and generate structural features

1. **Run the siWalk script** to scan the precursor sequence and identify the most likely effector siRNA location:
    - The script `siWalk_predict_siRNA_location.py` requires an input sequence (**priseq**) and the **DicerCall** parameter to define the dicing interval (e.g., 21 nt).
    - For optimal performance, it is recommended to limit precursor sequences to a **maximum of 200 nucleotides** to ensure efficient identification of local siRNA candidates, although there is no limitation on sequence length.

```
cd /path/siWalk/src/
priseq=CTTGACCTTGTAAGGCCTTTTCTTGACCTTGTAAGACCCCATCTCTTTCTAAACGTTTTATTATTTTCTCGTTTTACAGATTCTATTCTATCTCTTCTCAATATAGAATAGATATCTATC
DicerCall=21
python siWalk_predict_siRNA_location.py $priseq $DicerCall
```

B2. Obtain the recommendations of the most likely start and end positions of effector siRNA on a precursor

2. Retrieve the result file from the `/path/siWalk/output/timestamp/` folder. You will find the following output files:
    - **yourPrecursor.effector_localization_indication.tsv**: Contains the predicted effector localization information.
    - **yourPrecursor.effector_localization_indication.png**: A graphical representation of the predicted localization.
    - **yourPrecursor.effector_localization_indication_top6_recommendation.tsv**: Contains the top 6 predicted siRNA candidates from the precursor.

The most likely effector positions will also be displayed in the **standard output**. Stdout example:

```
> yourPrecursor, predicted siRNA start: 38, end: 59, score: 749.860201
```

Module X. train a machine learning model

X1. Create the training dataset

1. **Construct the training dataset**
    - For each precursor candidate, compute the relevant features as outlined in **A2**.
2. **Annotate the dataset with class labels**
    - Add a `consistent` column, where each instance is labeled as **True** or **False**, indicating whether the candidate corresponds to a true precursor (**True**) or a non-precursor (**False**).
    - Ensure that the dataset contains a sufficient number of positive and negative examples.
    - This labeling process is user-defined and should be tailored according to the specific dataset and classification requirements.
3. The resulting file will be the `annotated_training_file` for the next step.

X2. Generate the model

4. Considering **top 100 overall attributes**, run the script `siWalk_pickle_precursor.py` to train the machine learning model:
    - This script will produce several output files, including:
        - output/prefix_**GradientBoostingWithAdaBoost_model.pkl**: trained model with a timestamp for version control.
        - output/prefix_**feature_importance_n_correlation.tsv**: feature evaluation file listing the ranked importance of features and their correlation with precursor classes.
    - The model generated by this script, trained on the provided `background.tsv` file, is the **GBA100** model described in the paper, which is also included with the siWalk distribution.
    - Models in this category are intended solely for precursor predictions.

```
cd /path/siWalk/src/
annotated_training_file=../dbs/background.tsv #\\ or \\ annotated_training_file=from_X1
python siWalk_pickle_precursor.py $annotated_training_file
```

5. Considering **top 100 structural attributes**, run the script `siWalk_pickle_localization.py` to train the machine learning model:
    - This script will produce several output files, including:
        - output/prefix_**RandomForestWithAdaBoost_model.pkl**: trained model with a timestamp for version control.
        - output/prefix_**strcture_feature_importance_n_correlation.tsv**: feature evaluation file listing the ranked importance of structural features and their correlation with precursor classes.
    - The model generated by this script, trained on the provided `background.tsv` file, is the **RFAs100** model (also detailed in **M2**).
    - Models in this category are suitable for both precursor and effector predictions.

```
cd /path/siWalk/src/
annotated_training_file=../dbs/background.tsv #\\ or \\ annotated_training_file=from_X1
python siWalk_pickle_localization.py $annotated_training_file
```


Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

APPENDIX D

ARCTICFOX

A quantitative profiling framework for rAAV genome population

Chao-Jung Wu *+1,2, Hui-wen Liu +#2, Dongdong Lin 3, Svetlana Bergelson 2,
Baohong Zhang 3, Abdoulaye Baniré Diallo #1, Wei Zhang 2

* Presenting author
\+ Equal contribution
\# Corresponding authors
1. Computer Science, UQAM, Montréal, Canada
2. Technical Development, Biogen, Cambridge, USA
3. Computational Biology & Genomics, Biogen, Cambridge, USA

Emails: wu.chaojung@gmail.com; newiuhuil@gmail.com; dongdong.lin@biogen.com; svetlana.bergelson@biogen.com; baohong.zhang@biogen.com; diallo.abdoulaye@uqam.ca; wei.zhang@biogen.com

Key words: gene therapy vector quality control, rAAV genotyping, NGS long-read sequencing data analysis, residual DNA

Abstract

Recombinant adeno-associated virus (rAAV)-mediated gene therapy has been applied for human diseases. However, undesirable DNA can be encapsulated in viral particle due to cargo truncation and recombination with helper plasmids and host cell during production. Such impurities can be functional. The emerging long-read sequencing has achieved AAV single-genome resolution. While vector engineering improves cargo integrity, a comprehensive profiling method is needed to compare rAAV genome populations manufactured from different batches and vector designs. Here we propose a Python-based profiling framework to classify and quantitate residual DNA species in rAAV products. We designed a universal rAAV reference that contains 9 genetic components including a gene of interest, genes on the helper plasmids, and human genome, which are commonly used in rAAV production, as well as lambda genome for calibrating size bias from sequencing or DNA extraction procedures. Each read is genotyped as an encapsulated rAAV genome carrying cargo or chimeric cargo-residual DNAs, based on the presence of universal components. The cargo integrity and the functional potential of impurity are assessed based on both alignment pattern and percentage. Then, we applied this standardized profiling framework on rAAV long-read datasets and demonstrated that we could quantitatively compare rAAV qualities among various designs. When read abundance was calibrated, we could also compare genotype distributions between batches or processes. This new analysis framework will facilitate process development to ensure therapeutic product safety and quality.

Assessing functional impurities in rAAV production platforms by long-read sequencing

Chao-Jung Wu [*+1,2], **Hui-wen Liu** [+2], **Lauren Tereshko** [2], **Dongdong Lin** [3], **Svetlana Bergelson** [2], **Baohong Zhang** [3], **Abdoulaye Baniré Diallo** [#1], **Cullen Mason** [#2]

1. Informatique, UQÀM, Montréal, Canada
2. Technical Development, Biogen, Cambridge, USA
3. Computational Biology & Genomics, Biogen, Cambridge, USA
* Presenting author; + Equal contribution; # Corresponding authors

Abstract: (Background) Recombinant adeno-associated virus (rAAV)-mediated gene therapy has been applied for human diseases. **(Problematics)** However, the rAAV capsids contain heterogeneous mixtures of full-length and truncated genomes and, depending on the manufacturing process, residual host cell and plasmid DNA. Therefore, a method is needed to characterize the encapsidated DNA of rAAV in order to support process development and batch release. **(Approach)** The emerging long-read sequencing (LRS) has achieved AAV single-genome resolution. Here we propose a Python-based LRS profiling framework to classify and quantitate residual DNA species in rAAV products. We designed a reference that contains universal genetic components that are commonly used in rAAV production, including AmpR, KanR, Rep and Cap genes along with HPV18, Ad5 and hg38 genomes. **(Datasets)** We accessed the impurities of rAAV production from public and in-house LRS datasets. **(Results)** Analyzing the lambda fragments supplemented in these datasets showed that sequencing introduced size biases, which couldn't be fully corrected by regression but is improvable within library preparation. Functional potential of impurities were assessed through indicators derived from long-read alignments, which enabled us to quantitatively compare impurities between manufacturing batches. We demonstrated that LRS provides informative metrics for rAAV production and can facilitate process development to ensure therapeutic product safety and quality.

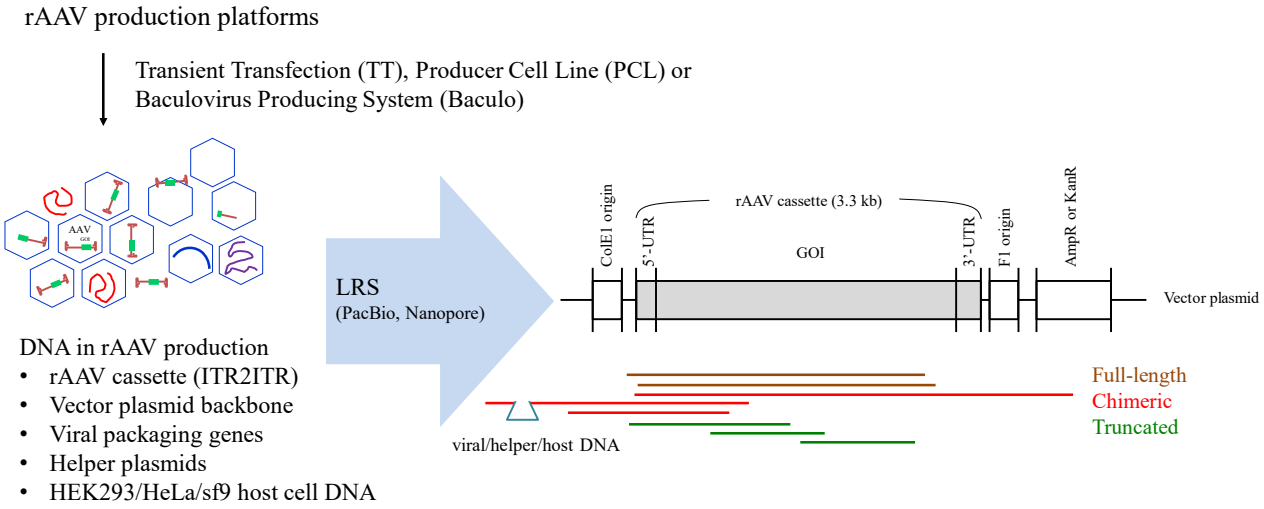
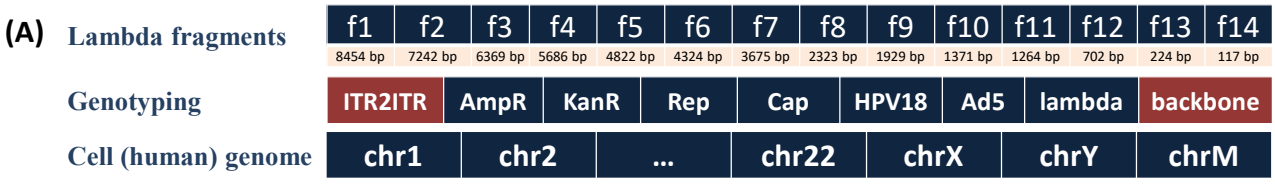


Figure 1. rAAV profiling by long-read sequencing. The rAAV capsids contain heterogeneous mixtures of full-length and truncated genomes and residual host cell and plasmid DNA, depending on the manufacturing process [1]. The emerging long-read sequencing (LRS) has AAV single-genome resolution [2]. We assessed in this study how LRS can help to classify and quantitate residual DNA species in rAAV products.



(B) Output metrics, ITR2ITR as example:

- **Positivity rate** = reads containing ITR2ITR / total reads in library
- **Intact rate** = reads are full length ITR2ITR / reads containing ITR2ITR; full-length threshold is 90%.
- **Mass percentage** = nucleotides aligned to ITR2ITR / total nucleotides in library

Figure 2. (A) the design of reference genomes and (B) the definition of metrics embedded in our LRS analytical framework. The transgene in ITR2ITR is default Cas9 derived from pX602 (Addgene) or specific. The backbone is default pX602 removing transgene or specific. Lambda DNA-BstEII digest (New England Biolabs) is often supplemented as sequencing control during library preparation.

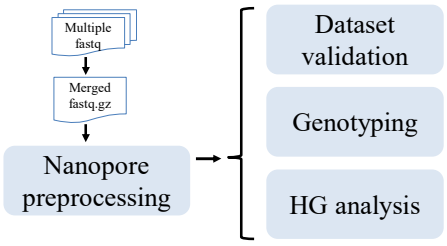


Figure 3. LRS pipeline for assessing functional impurities in rAAV, including three analytic modules embedded in the python-based software ArcticFox: **Dataset validation** to validate usable dataset, analyze lambda fragments and report **picking up** by template length and **reading through** by template length. **Genotyping** to analyze residual DNA based on several pre-defined chromosomes, reporting output metrics**. HG analysis** to analyze residual DNA coming from host genome hg38, reporting output metrics.

Table 1. LRS datasets of rAAV productions from public and in-house sources, including (a) in-house, (b) SRA PRJNA383145 [2], (c) SRA PRJNA608034 [3], and (d) SRA PRJNA794197 [4], representing various rAAV production systems; and sequencing data obtained from PacBio and Nanopore long-read sequencers.

Sample name	Transgene	Sequencer	Dataset source
Nanopore.Sample01	BIIB-X	MinION Mk1B	a
Nanopore.Sample05	BIIB-X	MinION Mk1B	a
PacBio.Sample01	BIIB-X	PACBIO Sequel II	a
PacBio.Sample05	BIIB-X	PACBIO Sequel II	a
pt2018_78	EGFP	PacBio RS II	b
pt2018_79	EGFP	PacBio RS II	b
pt2018_80	EGFP	PacBio RS II	b
pt2018_81	EGFP	PacBio RS II	b
pt2020_10 (Lib1)	Cas9	PACBIO Sequel	c
pt2020_09 (Lib2)	Cas9	PACBIO Sequel	c
pt2020_08 (Lib3)	Cas9	PACBIO Sequel	c
pt2020_07 (Lib4)	Cas9	PACBIO Sequel	c
pt2020_06 (Lib5)	mCherry	PACBIO Sequel	c
pt2020_05 (Lib6)	mCherry	PACBIO Sequel	c
pt2020_04 (Lib7)	EGFP	PACBIO Sequel	c
pt2020_03 (Lib8)	EGFP	PACBIO Sequel	c
pt2022_03	EGFP	PACBIO Sequel II	d
pt2022_04	EGFP	PACBIO Sequel II	d

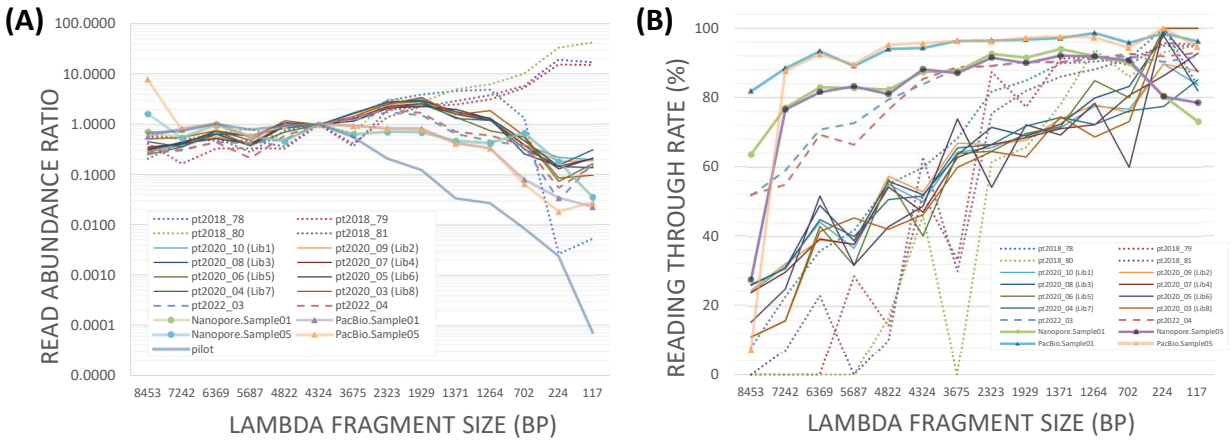


Figure 4. Size bias related to sequencing (A) regarding **picking up** template based on template length and (B) regarding complete **reading through** template based on template length. The *BstEII*-digestion of lambda genome gives rise to 14 fragments representing a range of template sizes that are theoretically of equal abundance ratio. After sequence identification, read abundance of each lambda fragment was measured and the abundance ratios were calculated against the fragment of 4324 bp. The theoretical reading through rate is 100%.

formula 1: $R_o = f(L_t) = b_o + b_1 L_t + b_2 L_t^2 + \ldots$

formula 2: $L_t = g(L_o) = c_o + c_o L_o$

formula 3: $A_c = h(L_o) = A_o / R_o = A_o / f(L_t) = A_o / f(g(L_o))$

Formulas for Modeling size biases of read abundance by regression. Given A_o: Observed read abundance, L_t: Theoretical template length, L_o: Observed read length and R_o: Read abundance ratio (against baseline f6) obtained by analyzing lambda fragments, the corrected read abundance A_c in formula 3 can be obtained by determining formula 1 and 2, where formula 1 models the **size bias** regarding **picking up** template based on template length trained on lambda; and formula 2 models the **size bias** regarding complete **reading through** template based on template length trained on lambda. The model of formula 1 is a polynomial regression, and that of formula 2 is a linear regression.

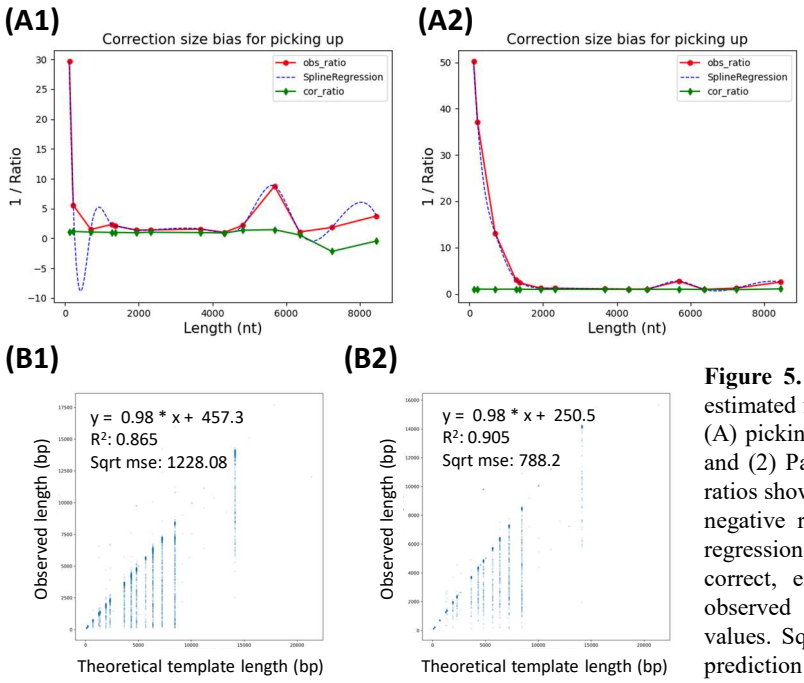


Figure 5. Correcting size biases of read abundance estimated from training on lambda in sample01 regarding (A) picking up and (B) reading through in (1) Nanopore and (2) PacBio. Picking up bias can be corrected if the ratios show a steady trend without fluctuations, otherwise negative ratio and overfitting occur in the polynomial regression model. Reading through is very difficult to correct, especially that the theoretical lengths of an observed small read have a wide range of possible values. Sqrt mse indicates on average how far (bp) the prediction of the linear regression model is from the theoretical length.

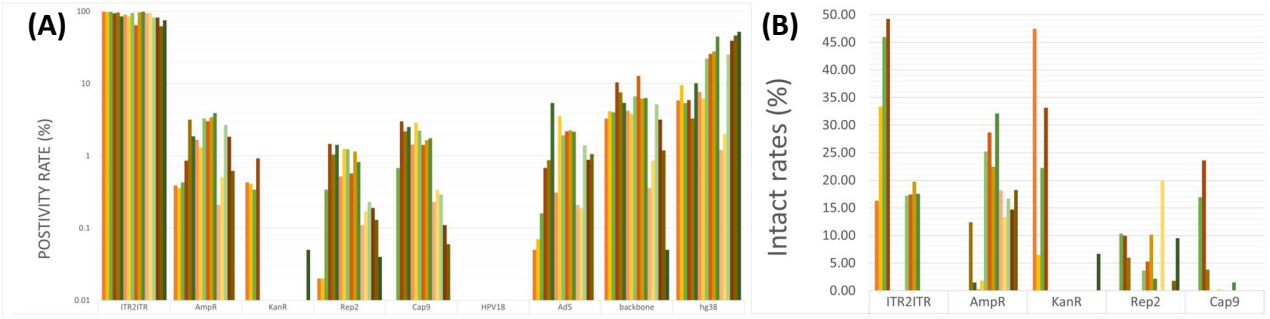
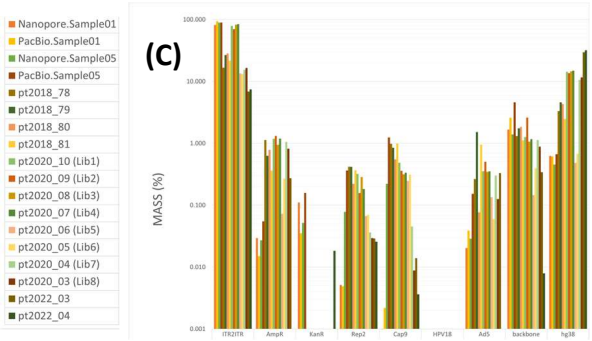


Figure 6. LRS sequencing provides several output metrics to evaluate functional potential of impurities, such as (A) Percentages of DNA populations based on reference genome positivity rates, (B) Intact rates of reference genomes, and (C) Mass percentages of reference genomes inferred from base count. These impurity metrics of the universal genetic components - AmpR, KanR, Rep and Cap genes and host genomes – allowed for quantitative comparison of impurities from different manufacturing batches; while the positivity rate can be compared if the ITR2ITR transgene between batches are the same.



Conclusion:

❖ LRS provides an opportunity to quick profile the drug substance before delivery or assess the outcome of a production procedure change or a vector design change.

- ✓ The universal reference genomes capture the genetic components that are commonly used in rAAV production.
- ✓ In addition to mass percentages of reference genomes, LRS provides the positivity rates of DNA populations and intact rates of reference genomes, which could not be obtained via short-read sequencing (SRS).

❖ LRS is prone to size bias. Supplement of Lambda could be used to evaluate the quality of the library preparation and validate a usable dataset.

- ✓ The biases are similar within an experiment but different between experiments, indicating that the bias comes from library preparation rather than the sequencers.
- ✓ The size biases could be modeled by regression. However, overfitting may happen due to insufficient datapoints (e.g., 14 datapoints from standard fragments); and it is difficult to predict the theoretical lengths of an observed small read for modeling reading through. The correction power of regression is therefore limited. Improving library preparation is more intuitive than correcting the biases mathematically.

References

References:

[1] Michael F Naso, Brian Tomkowicz, William L Perry III, and William R Strohl. "Adeno-associated virus (aav) as a vector for gene therapy". BioDrugs, 31(4):317–334, 2017

[2] Tai, Phillip WL, et al. "Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras." *Molecular Therapy-Methods & Clinical Development* 9 (2018): 130-141.

[3] Tran, Ngoc Tam, et al. "AAV-genome population sequencing of vectors packaging CRISPR components reveals design-influenced heterogeneity." *Molecular Therapy-Methods & Clinical Development* 18 (2020): 639-651.

[4] Tran, Ngoc Tam, et al. "Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity." *Human Gene Therapy* 33.7-8 (2022): 371-388.

BIBLIOGRAPHY

Adams, D., Gonzalez-Duarte, A., O'Riordan, W. D., Yang, C.-C., Ueda, M., Kristen, A. V., Tournev, I., Schmidt, H. H., Coelho, T., Berk, J. L. *et al.* (2018). Patisiran, an rnai therapeutic, for hereditary transthyretin amyloidosis. *New england journal of medicine*, *379*(1), 11–21.

Agharbaoui, Z., Leclercq, M., Remita, M. A., Badawi, M. A., Lord, E., Houde, M., Danyluk, J., Diallo, A. B. et Sarhan, F. (2015). An integrative approach to identify hexaploid wheat mirnaome associated with development and tolerance to abiotic stress. *BMC genomics*, *16*(1), 339.

Agrawal, N., Dasaradhi, P., Mohmmed, A., Malhotra, P., Bhatnagar, R. K. et Mukherjee, S. K. (2003). Rna interference: biology, mechanism, and applications. *Microbiology and molecular biology reviews*, *67*(4), 657–685.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91–93.

Akpinar, B. A., Kantar, M. et Budak, H. (2015). Root precursors of micrornas in wild emmer and modern wheats show major differences in response to drought stress. *Functional & integrative genomics*, *15*(5), 587–598.

Allen, E., Xie, Z., Gustafson, A. M. et Carrington, J. C. (2005). microrna-directed phasing during trans-acting sirna biogenesis in plants. *Cell*, *121*(2), 207–221.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–3402.

Andrews, S. *et al.* (2010). Fastqc: a quality control tool for high throughput sequence data. *Bioinformatics*, *Babraham Institute (Cambridge, UK)*.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J. *et al.* (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), eaar7191.

Ardui, S., Ameur, A., Vermeesch, J. R. et Hestand, M. S. (2018). Single molecule real-time (smrt) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research*, *46*(5), 2159–2168.

Ataei, S., Ahmadi, J., Marashi, S.-A. et Abolhasani, I. (2024). Amir-p3: An ai-based microrna prediction pipeline in plants. *Plos one*, *19*(8), e0308016.

Atchison, R. W., Casto, B. C. et Hammon, W. M. (1965). Adenovirus-associated defective virus particles. *Science*, *149*(3685), 754–756.

Auer, P. L. et Doerge, R. (2010). Statistical design and analysis of rna sequencing data. *Genetics*, *185*(2), 405–416.

Axtell, M. J. (2013a). Classification and comparison of small rnas from plants. *Annual review of plant biology*, *64*, 137–159.

Axtell, M. J. (2013b). Shortstack: comprehensive annotation and quantification of small rna genes. *Rna*, *19*(6), 740–751.

Axtell, M. J., Jan, C., Rajagopalan, R. et Bartel, D. P. (2006). A two-hit trigger for sirna biogenesis in plants. *Cell*, *127*(3), 565–577.

Axtell, M. J. et Meyers, B. C. (2018). Revisiting criteria for plant microrna annotation in the era of big data. *The Plant Cell*, *30*(2), 272–284.

Bajan, S. et Hutvagner, G. (2020). Rna-based therapeutics: From antisense oligonucleotides to mirnas. *Cells*, *9*(1), 137.

Baldrich, P., Kakar, K., Siré, C., Moreno, A. B., Berger, A., García-Chapa, M., López-Moya, J. J., Riechmann, J. L. et San Segundo, B. (2014). Small rna profiling reveals regulation of arabidopsis mir168 and heterochromatic sirna415 in response to fungal elicitors. *BMC genomics*, *15*, 1–16.

Banimfreg, B. H. (2023). A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics. *Healthcare Analytics*, *3*, 100190.

Bartel, B. et Bartel, D. P. (2003). Micrornas: at the root of plant development? *Plant physiology*, *132*(2), 709–717.

Batut, B., van den Beek, M., Doyle, M. A. et Soranzo, N. (2021). Rna-seq data analysis in galaxy. In *Rna Bioinformatics* 367–392. Springer.

Baulcombe, D. (2004). Rna silencing in plants. *Nature*, *431*(7006), 356–363.

Beg, M. S., Brenner, A. J., Sachdev, J., Borad, M., Kang, Y.-K., Stoudemire, J., Smith, S., Bader, A. G., Kim, S. et Hong, D. S. (2017). Phase i study of mrx34, a liposomal mir-34a mimic, administered twice weekly in patients with advanced solid tumors. *Investigational new drugs*, *35*, 180–188.

Bej, S. et Basak, J. (2014). Micrornas: the potential biomarkers in plant stress response. *American Journal of Plant Sciences*, *2014*.

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Bergman, N. H. (2007). *Comparative genomics, Chapter 9 BLAST QuickStart*. Springer Science & Business Media.

Biegert, A., Mayer, C., Remmert, M., Söding, J. et Lupas, A. N. (2006). The mpi bioinformatics toolkit for protein sequence analysis. *Nucleic acids research*, *34*(suppl_2), W335–W339.

Billi, A. C., Fischer, S. E. et Kim, J. K. (2018). Endogenous rnai pathways in c. elegans. In *WormBook: The Online Review of C. elegans Biology [Internet]*. WormBook.

Bolser, D. M., Staines, D. M., Perry, E. et Kersey, P. J. (2017). Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In *Plant Genomics Databases* 1–31. Springer.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y. *et al.* (2013). Blast: a more efficient report with usability improvements. *Nucleic acids research*, *41*(W1), W29–W33.

Borthakur, D. *et al.* (2008). Hdfs architecture guide. *Hadoop apache project*, *53*(1-13), 2.

Bradshaw, S., Brazil, E. et Chodorow, K. (2019). *MongoDB: the definitive guide: powerful and scalable data storage*. " O'Reilly Media, Inc.".

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Budak, H. et Akpinar, B. A. (2015). Plant mirnas: biogenesis, organization and origins. *Functional & integrative genomics*, *15*(5), 523–531.

Buschmann, D., Haberberger, A., Kirchner, B., Spornraft, M., Riedmaier, I., Schelling, G. et Pfaffl, M. W. (2016). Toward reliable biomarker signatures in the age of liquid biopsies-how to standardize the small rna-seq workflow. *Nucleic Acids Research*, *44*(13), 5995–6018.

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B., Dworkin, J. P., Lupisella, M. L., Smith, D. J., Botkin, D. J. *et al.* (2017). Nanopore dna sequencing and genome assembly on the international space station. *Scientific reports*, *7*(1), 18022.

Catalanotto, C., Cogoni, C. et Zardo, G. (2016). Microrna in control of gene expression: an overview of nuclear functions. *International journal of molecular sciences*, *17*(10), 1712.

Chadeuf, G., Ciron, C., Moullier, P. et Salvetti, A. (2005). Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their in vivo persistence after vector delivery. *Molecular Therapy*, *12*(4), 744–753.

Chakravarthy, S., Sternberg, S. H., Kellenberger, C. A. et Doudna, J. A. (2010). Substrate-specific kinetics of dicer-catalyzed rna processing. *Journal of molecular biology*, *404*(3), 392–402.

Chang, C.-C. et Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1–27.

Chao, Y.-S., Wu, C.-J., Lai, Y.-C., Hsu, H.-T., Cheng, Y.-P., Wu, H.-C., Huang, S.-Y. et Chen, W.-C. (2022). Diagnostic accuracy of symptoms for an underlying disease: a simulation study. *Scientific Reports*, *12*(1), 13810.

Chen, C., Li, J., Feng, J., Liu, B., Feng, L., Yu, X., Li, G., Zhai, J., Meyers, B. C. et Xia, R. (2021a). srnaanno—a database repository of uniformly annotated small rnas in plants. *Horticulture Research*, *8*. Accessed on February 20, 2024. Récupéré de `http://www.plantsrnas.org`

Chen, H.-M., Chen, L.-T., Patel, K., Li, Y.-H., Baulcombe, D. C. et Wu, S.-H. (2010). 22-nucleotide rnas trigger secondary sirna biogenesis in plants. *Proceedings of the National Academy of Sciences*, *107*(34), 15269–15274.

Chen, H.-M., Li, Y.-H. et Wu, S.-H. (2007). Bioinformatic prediction and experimental validation of a microrna-directed tandem trans-acting sirna cascade in arabidopsis. *Proceedings of the National Academy of Sciences*, *104*(9), 3318–3323.

Chen, J., Wang, S., Wu, F., Wei, M., Li, J. et Yang, F. (2022). Genome–wide identification and functional characterization of auxin response factor (arf) genes in eggplant. *International Journal of Molecular Sciences*, *23*(11), 6219.

Chen, X., Ding, Y., Yang, Y., Song, C., Wang, B., Yang, S., Guo, Y. et Gong, Z. (2021b). Protein kinases in plant responses to drought, salt, and cold stress. *Journal of integrative plant biology*, *63*(1), 53–78.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. et Hwang, C.-C. (2013). Effects of gc bias in next-generation-sequencing data on de novo genome assembly. *PloS one*, *8*(4), e62856.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. et Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, *38*(6), 1767–1771.

Conesa, A. et Götz, S. (2008). Blast2go: a comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, *2008*.

Consortium, G. O. *et al.* (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, *32*(suppl 1), D258–D261.

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506–D515.

Consortium, U. (2021). Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, *49*(D1), D480–D489.

Dai, X. et Zhao, P. X. (2011). psrnatarget: a plant small rna target analysis server. *Nucleic acids research*, *39*(suppl 2), W155–W159.

Dai, X., Zhuang, Z. et Zhao, P. X. (2018). psrnatarget: a plant small rna target analysis server (2017 release). *Nucleic acids research*.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. *et al.* (2021). Twelve years of samtools and bcftools. *Gigascience*, *10*(2), giab008.

Darty, K., Denise, A. et Ponty, Y. (2009). Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, *25*(15), 1974.

de Castro, M. R., dos Santos Tostes, C., Dávila, A. M., Senger, H. et da Silva, F. A. (2017). Sparkblast: scalable blast processing using in-memory operations. *BMC bioinformatics*, *18*(1), 318.

de Felippes, F. F., Marchais, A., Sarazin, A., Oberlin, S. et Voinnet, O. (2017). A single mir390 targeting event is sufficient for triggering tas3-tasirna biogenesis in arabidopsis. *Nucleic acids research*, *45*(9), 5539–5554.

de Oliveira, L. F. V., Christoff, A. P., de Lima, J. C., de Ross, B. C. F., Sachetto-Martins, G., Margis-Pinheiro, M. et Margis, R. (2014). The wall-associated kinase gene family in rice genomes. *Plant Science*, *229*, 181–192.

De Paoli, E., Dorantes-Acosta, A., Zhai, J., Accerbi, M., Jeong, D.-H., Park, S., Meyers, B. C., Jorgensen, R. A. et Green, P. J. (2009). Distinct extremely abundant sirnas associated with cosuppression in petunia. *Rna*, *15*(11), 1965–1970.

Dean, J., Ghemawat, S. *et al.* (2004). Mapreduce: simplified data processing on large clusters. Dans *osdi*, volume 4, p. 5. USA.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M. *et al.* (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, *43*(5), 491–498.

Devore, J. L., Berk, K. N., Carlton, M. A. *et al.* (2012). *Modern mathematical statistics with applications*, volume 285. Springer.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. et Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, *35*(4), 316–319.

Diao, P., Chen, C., Zhang, Y., Meng, Q., Lv, W. et Ma, N. (2020). The role of nac transcription factor in plant cold response. *Plant Signaling & Behavior*, *15*(9), 1785668.

Dobrowsky, T., Gianni, D., Pieracci, J. et Suh, J. (2021). Aav manufacturing for clinical use: insights on current challenges from the upstream process perspective. *Current Opinion in Biomedical Engineering*, *20*, 100353.

Doench, J. G., Petersen, C. P. et Sharp, P. A. (2003). sirnas can function as mirnas. *Genes & development*, *17*(4), 438–442.

Dong, C.-H., Agarwal, M., Zhang, Y., Xie, Q. et Zhu, J.-K. (2006). The negative regulator of plant cold responses, hos1, is a ring e3 ligase that mediates the ubiquitination and degradation of ice1. *Proceedings of the National Academy of Sciences*, *103*(21), 8281–8286.

Dong, C.-H. et Pei, H. (2014). Over-expression of mir397 improves plant tolerance to cold stress in arabidopsis thaliana. *Journal of Plant Biology*, *57*(4), 209–217.

Dotto, M. C., Petsch, K. A., Aukerman, M. J., Beatty, M., Hammell, M. et Timmermans, M. C. (2014). Genome-wide analysis of leafbladeless1-regulated and phased small rnas underscores the importance of the tas3 ta-sirna pathway to maize development. *PLoS genetics*, *10*(12), e1004826.

Du, Z., Lin, W., Yu, B., Zhu, J. et Li, J. (2022). Integrated metabolomic and transcriptomic analysis of the flavonoid accumulation in the leaves of cyclocarya paliurus at different altitudes. *Frontiers in plant science*, *12*, 794137.

Fahlgren, N., Montgomery, T. A., Howell, M. D., Allen, E., Dvorak, S. K., Alexander, A. L. et Carrington, J. C. (2006). Regulation of auxin response factor3 by tas3 ta-sirna affects developmental timing and patterning in arabidopsis. *Current biology*, *16*(9), 939–944.

Fang, Y., Xie, K. et Xiong, L. (2014). Conserved mir164-targeted nac genes negatively regulate drought resistance in rice. *Journal of experimental botany*, *65*(8), 2119–2135.

Farazi, T. A., Juranek, S. A. et Tuschl, T. (2008). The growing catalog of small rnas and their association with distinct argonaute/piwi family members. *Development*, *135*(7), 1201–1214.

Fei, Q., Xia, R. et Meyers, B. C. (2013). Phased, secondary, small interfering rnas in posttranscriptional regulatory networks. *The Plant Cell*, *25*(7), 2400–2415.

Finkel, R. S., Mercuri, E., Darras, B. T., Connolly, A. M., Kuntz, N. L., Kirschner, J., Chiriboga, C. A., Saito, K., Servais, L., Tizzano, E. *et al.* (2017). Nusinersen versus sham control in infantile-onset spinal muscular atrophy. *New England Journal of Medicine*, *377*(18), 1723–1732.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. et Mello, C. C. (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *nature*, *391*(6669), 806–811.

Freund, Y. et Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Fukudome, A. et Fukuhara, T. (2017). Plant dicer-like proteins: double-stranded rna-cleaving enzymes for small rna biogenesis. *Journal of Plant Research*, *130*, 33–44.

Fukunaga, R., Han, B. W., Hung, J.-H., Xu, J., Weng, Z. et Zamore, P. D. (2012). Dicer partner proteins tune the length of mature mirnas in flies and mammals. *Cell*, *151*(3), 533–546.

Gasciolli, V., Mallory, A. C., Bartel, D. P. et Vaucheret, H. (2005). Partially redundant functions of arabidopsis dicer-like enzymes and a role for dcl4 in producing trans-acting sirnas. *Current Biology*, *15*(16), 1494–1500.

Giusti, L., Mica, E., Bertolini, E., De Leonardis, A. M., Faccioli, P., Cattivelli, L. et Crosatti, C. (2017). micrornas differentially modulated in response to heat and drought stress in durum wheat cultivars with contrasting water use efficiency. *Functional & integrative genomics*, *17*(2-3), 293–309.

Green, M. R. et Sambrook, J. (2012). Molecular cloning. *A Laboratory Manual 4th*, *448*.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J. et Team, B. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, *15*(7), 475–476.

Guleria, P., Mahajan, M., Bhardwaj, J. et Yadav, S. K. (2011). Plant small rnas: biogenesis, mode of action and their roles in abiotic stresses. *Genomics, proteomics & bioinformatics*, *9*(6), 183–199.

Guo, C., Li, L., Wang, X. et Liang, C. (2015). Alterations in sirna and mirna expression profiles detected by deep sequencing of transgenic rice with sirna-mediated viral resistance. *PLoS One*, *10*(1), e0116175.

Guo, G., Liu, X., Sun, F., Cao, J., Huo, N., Wuda, B., Xin, M., Hu, Z., Du, J., Xia, R. *et al.* (2018). Wheat mir9678 affects seed germination by generating phased sirnas and modulating abscisic acid/gibberellin signaling. *The Plant Cell*, *30*(4), 796–814.

Guo, Q., Qu, X. et Jin, W. (2014). Phasetank: genome-wide computational identification of phasirnas and their regulatory cascades. *Bioinformatics*, *31*(2), 284–286.

Hajiahmadi, Z., Abedi, A., Wei, H., Sun, W., Ruan, H., Zhuge, Q. et Movahedi, A. (2020). Identification, evolution, expression, and docking studies of fatty acid desaturase genes in wheat (triticum aestivum l.). *BMC genomics*, *21*, 1–20.

Hamilton, A. J. et Baulcombe, D. C. (1999). A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, *286*(5441), 950–952.

Hammond, R. K., Gupta, P., Patel, P. et Meyers, B. C. (2023). mirador: a fast and precise tool for the prediction of plant mirnas. *Plant Physiology*, *191*(2), 894–903.

Han, R., Jian, C., Lv, J., Yan, Y., Chi, Q., Li, Z., Wang, Q., Zhang, J., Liu, X. et Zhao, H. (2014). Identification and characterization of micrornas in the flag leaf and developing seed of wheat (triticum aestivum l.). *BMC genomics*, *15*(1), 289.

Hara, N., Kikuchi, M., Miyashita, A., Hatsuta, H., Saito, Y., Kasuga, K., Murayama, S., Ikeuchi, T. et Kuwano, R. (2017). Serum microrna mir-501-3p as a potential biomarker related to the progression of alzheimer's disease. *Acta neuropathologica communications*, *5*, 1–9.

Hastie, T., Tibshirani, R., Friedman, J. H. et Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Health Canada (2024). Biologics, radiopharmaceuticals and genetic therapies. https://www.canada.ca/en/health-canada/services/drugs-health-products/biologics-radiopharmaceuticals-genetic-therapies.html. Accessed April 2025.

Hermonat, P. L. et Muzyczka, N. (1984). Use of adeno-associated virus as a mammalian dna cloning vector: transduction of neomycin resistance into mammalian tissue culture cells. *Proceedings of the National Academy of Sciences*, *81*(20), 6466–6470.

Hoggan, M. D., Blacklow, N. R. et Rowe, W. (1966). Studies of small dna viruses found in various adenovirus preparations: physical, biological, and immunological characteristics. *Proceedings of the National Academy of Sciences*, *55*(6), 1467–1474.

Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., Kasschau, K. D. et Carrington, J. C. (2007). Genome-wide analysis of the rna-dependent rna polymerase6/dicer-like4 pathway in arabidopsis reveals dependency on mirna-and tasirna-directed targeting. *The Plant Cell*, *19*(3), 926–942.

Hu, R., Qi, G., Kong, Y., Kong, D., Gao, Q. et Zhou, G. (2010). Comprehensive analysis of nac domain transcription factor gene family in populus trichocarpa. *BMC plant biology*, *10*(1), 1–23.

Hu, Z., Yu, Y., Wang, R., Yao, Y., Peng, H., Ni, Z. et Sun, Q. (2011). Expression divergence of tambd2 homoeologous genes encoding methyl cpg-binding domain proteins in wheat (triticum aestivum l.). *Gene*, *471*(1-2), 13–18.

Hutvágner, G. et Zamore, P. D. (2002). A microrna in a multiple-turnover rnai enzyme complex. *Science*, *297*(5589), 2056–2060.

Islam, M. N., Masud, M. K., Haque, M. H., Hossain, M. S. A., Yamauchi, Y., Nguyen, N.-T. et Shiddiky, M. J. (2017). Rna biomarkers: diagnostic and prognostic potentials and recent developments of electrochemical biosensors. *Small Methods*, *1*(7), 1700131.

Jeong, D.-H., Thatcher, S. R., Brown, R. S., Zhai, J., Park, S., Rymarquis, L. A., Meyers, B. C. et Green, P. J. (2013). Comprehensive investigation of microrna enhanced by analysis of sequence variants, expression patterns, argonaute loading, and target cleavage. *Plant physiology*, *162*(3), 1225–1245.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. et Marks, D. S. (2004). Human microrna targets. *PLoS biology*, *2*(11), e363.

Johnson, N. R., Yeoh, J. M., Coruh, C. et Axtell, M. J. (2016). Improved placement of multi-mapping small rnas. *G3: Genes, Genomes, Genetics*, *6*(7), 2103–2111.

Jones-Rhoades, M. W. et Bartel, D. P. (2004). Computational identification of plant microrna and their targets, including a stress-induced mirna. *Molecular cell*, *14*(6), 787–799.

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M. et Sarnow, P. (2005). Modulation of hepatitis c virus rna abundance by a liver-specific microrna. *science*, *309*(5740), 1577–1581.

Joshi, T. et Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC genomics*, *8*(1), 1–10.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* (2021). Highly accurate protein structure prediction with alphafold. *nature*, *596*(7873), 583–589.

Kakrana, A., Li, P., Patel, P., Hammond, R., Anand, D., Mathioni, S. et Meyers, B. (2017). Phasis: A computational suite for de novo discovery and characterization of phased, sirna-generating loci and their mirna triggers. *BioRxiv*, p. 158832.

Kakrana, A., Mathioni, S. M., Huang, K., Hammond, R., Vandivier, L., Patel, P., Arikit, S., Shevchenko, O., Harkess, A. E., Kingham, B. *et al.* (2018). Plant 24-nt reproductive phasirnas from intramolecular duplex mrnas in diverse monocots. *Genome research*, *28*(9), 1333–1344.

Kal, A. J., van Zonneveld, A. J., Benes, V., van den Berg, M., Koerkamp, M. G., Albermann, K., Strack, N., Ruijter, J. M., Richter, A., Dujon, B. *et al.* (1999). Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular biology of the cell*, *10*(6), 1859–1872.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. et Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361.

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., Hatzigeorgiou, A., Sartzetakis, S. et Sellis, T. (2014). Mr-microt: a mapreduce-based microrna target prediction method. Dans *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, p. 47. ACM.

Kantar, M., Lucas, S. J. et Budak, H. (2011). mirna expression patterns of triticum dicoccoides in response to shock drought stress. *Planta*, *233*(3), 471–484.

Kazemi-Shahandashti, S.-S. et Maali-Amiri, R. (2018). Global insights of protein responses to cold stress in plants: Signaling, defence, and degradation. *Journal of plant physiology*, *226*, 123–135.

Kenan-Eichler, M., Leshkowitz, D., Tal, L., Noor, E., Melamed-Bessudo, C., Feldman, M. et Levy, A. A. (2011). Wheat hybridization and polyploidization results in deregulation of small rnas. *Genetics*, *188*(2), 263–272.

Khraiwesh, B., Zhu, J.-K. et Zhu, J. (2012). Role of mirnas and sirnas in biotic and abiotic stress responses of plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1819*(2), 137–148.

Kim, D.-H., Behlke, M. A., Rose, S. D., Chang, M.-S., Choi, S. et Rossi, J. J. (2005). Synthetic dsrna dicer substrates enhance rnai potency and efficacy. *Nature biotechnology*, *23*(2), 222.

Koch, A., Kumar, N., Weber, L., Keller, H., Imani, J. et Kogel, K.-H. (2013). Host-induced gene silencing of cytochrome p450 lanosterol c14α-demethylase–encoding genes confers strong resistance to fusarium species. *Proceedings of the National Academy of Sciences*, *110*(48), 19324–19329.

Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T. et Winter, D. R. (2018). A beginner's guide to analysis of rna sequencing data. *American journal of respiratory cell and molecular biology*, *59*(2), 145–157.

Kornbrot, D. (2014). Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*.

Köster, J. et Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522.

Kozomara, A., Birgaoanu, M. et Griffiths-Jones, S. (2018). mirbase: from microrna sequences to function. *Nucleic acids research*, *47*(D1), D155–D162.

Kozomara, A., Birgaoanu, M. et Griffiths-Jones, S. (2019). mirbase: from microrna sequences to function. *Nucleic acids research*, *47*(D1), D155–D162. Accessed on May 7, 2024. Récupéré de `https://mirbase.org`

Kruskal, W. H. et Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Kulkarni, J. A., Cullis, P. R. et Van Der Meel, R. (2018). Lipid nanoparticles enabling gene therapies: from concepts to clinical utility. *Nucleic acid therapeutics*, *28*(3), 146–157.

Kurtoglu, K. Y., Kantar, M., Lucas, S. J. et Budak, H. (2013). Unique and conserved micrornas in wheat chromosome 5d revealed by next-generation sequencing. *PLoS One*, *8*(7), e69801.

Langmead, B., Trapnell, C., Pop, M. et Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, *10*(3), R25.

Lebatteux, D., Remita, A. M. et Diallo, A. B. (2019). Toward an alignment-free method for feature extraction and accurate classification of viral sequences. *Journal of Computational Biology*, *26*(6), 519–535.

Leclercq, M., Diallo, A. B. et Blanchette, M. (2013). Computational prediction of the localization of micrornas within their pre-mirna. *Nucleic acids research*, *41*(15), 7200–7211.

Leclercq, M., Diallo, A. B. et Blanchette, M. (2017). Prediction of human mirna target genes using computationally reconstructed ancestral mammalian sequences. *Nucleic acids research*, *45*(2), 556–566.

Lecomte, E., Tournaire, B., Cogné, B., Dupont, J.-B., Lindenbaum, P., Martin-Fontaine, M., Broucque, F., Robin, C., Hebben, M., Merten, O.-W. *et al.* (2015). Advanced characterization of dna molecules in raav vector preparations by single-stranded virus next-generation sequencing. *Molecular Therapy-Nucleic Acids*, *4*, e260.

Lee, R. C., Feinbaum, R. L. et Ambros, V. (1993). The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *cell*, *75*(5), 843–854.

Lei, J. et Sun, Y. (2014). mir-prefer: an accurate, fast and easy-to-use plant mirna prediction tool using small rna-seq data. *Bioinformatics*, *30*(19), 2837–2839.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, Q., Byrns, B., Badawi, M. A., Diallo, A. B., Danyluk, J., Sarhan, F., Laudencia-Chingcuanco, D., Zou, J. et Fowler, D. B. (2018). Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant physiology*, *176*(3), 2376–2394.

Li, S., Castillo-González, C., Yu, B. et Zhang, X. (2017). The functions of plant small rna s in development and in stress responses. *The Plant Journal*, *90*(4), 654–670.

Li, Y., Li, C., Ding, G. et Jin, Y. (2011a). Evolution of mir159/319 microrna genes and their post-transcriptional regulatory link to sirna pathways. *BMC evolutionary biology*, *11*(1), 1–18.

Li, Y., Li, C., Xia, J. et Jin, Y. (2011b). Domestication of transposable elements into microrna genes in plants. *Plos one*, *6*(5), e19212.

Liu, H.-H., Tian, X., Li, Y.-J., Wu, C.-A. et Zheng, C.-C. (2008). Microarray-based analysis of stress-regulated microrrnas in arabidopsis thaliana. *Rna*, *14*(5), 836–843.

Liu, Y., Teng, C., Xia, R. et Meyers, B. C. (2020). Phasirnas in plants: their biogenesis, genic sources, and roles in stress responses, development, and reproduction. *Plant Cell*, *32*(10), 3059–3080.

Liu, Y.-X., Wang, M. et Wang, X.-J. (2014). Endogenous small rna clusters in plants. *Genomics, Proteomics & Bioinformatics*, *12*(2), 64–71.

Llave, C. (2010). Virus-derived small interfering rnas at the core of plant–virus interactions. *Trends in plant science*, *15*(12), 701–707.

Locati, M. D., Terpstra, I., de Leeuw, W. C., Kuzak, M., Rauwerda, H., Ensink, W. A., van Leeuwen, S., Nehrdich, U., Spaink, H. P., Jonker, M. J. *et al.* (2015). Improving small rna-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization. *Nucleic acids research*, *43*(14), e89–e89.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. et Hofacker, I. L. (2011). Viennarna package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.

Loukoianov, A., Yan, L., Blechl, A., Sanchez, A. et Dubcovsky, J. (2005). Regulation of vrn-1 vernalization genes in normal and transgenic polyploid wheat. *Plant Physiology*, *138*(4), 2364–2373.

Love, M. I., Huber, W. et Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, *15*(12), 550.

Lu, D. et Thum, T. (2019). Rna-based diagnostic and therapeutic strategies for cardiovascular disease. *Nature Reviews Cardiology*, *16*(11), 661–674.

Lucas, S. J. et Budak, H. (2012). Sorting the wheat from the chaff: identifying mirnas in genomic survey sequences of triticum aestivum chromosome 1al. *PLoS One*, *7*(7), e40859.

Lück, S., Kreszies, T., Strickert, M., Schweizer, P., Kuhlmann, M. et Douchkov, D. (2019). sirna-finder (si-fi) software for rnai-target design and off-target prediction. *Frontiers in Plant Science*, *10*, 1023.

Lukasik, A., Wójcikowski, M. et Zielenkiewicz, P. (2016). Tools4mirs–one place to gather all the tools for mirna analysis. *Bioinformatics*, *32*(17), 2722–2724.

Lunardon, A., Johnson, N. R., Hagerott, E., Phifer, T., Polydore, S., Coruh, C. et Axtell, M. J. (2020). Integrated annotations and analyses of small rna–producing loci from 47 diverse plants. *Genome research*, *30*(3), 497–513.

Lv, D.-K., Bai, X., Li, Y., Ding, X.-D., Ge, Y., Cai, H., Ji, W., Wu, N. et Zhu, Y.-M. (2010). Profiling of cold-stress-responsive mirnas in rice by microarrays. *Gene*, *459*(1-2), 39–47.

Lv, X., Lan, S., Guy, K. M., Yang, J., Zhang, M. et Hu, Z. (2016). Global expressions landscape of nac transcription factor family and their responses to abiotic stresses in citrullus lanatus. *Scientific Reports*, *6*(1), 1–14.

Ma, X., Xin, Z., Wang, Z., Yang, Q., Guo, S., Guo, X., Cao, L. et Lin, T. (2015). Identification and comparative analysis of differentially expressed mirnas in leaves of two wheat (triticum aestivum l.) genotypes during dehydration stress. *BMC Plant Biology*, *15*(1), 21.

MacRae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D. et Doudna, J. A. (2006). Structural basis for double-stranded rna processing by dicer. *Science*, *311*(5758), 195–198.

Mallory, A. C. et Vaucheret, H. (2006). Functions of micrornas and related small rnas in plants. *Nature genetics*, *38*(6), S31–S36.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B., Steuernagel, B., Mayer, K. F., Olsen, O.-A. *et al.* (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *science*, *345*(6194).

Martens-Uzunova, E., Jalava, S., Dits, N., Van Leenders, G., Møller, S., Trapman, J., Bangma, C., Litman, T., Visakorpi, T. et Jenster, G. (2012). Diagnostic and prognostic signatures from the small non-coding rna transcriptome in prostate cancer. *Oncogene*, *31*(8), 978–991.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp–10.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, *23*(2), 143–149.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*.

Meijer, A., De Meyer, T., Vandepoele, K. et Kyndt, T. (2022). Spatiotemporal expression profile of novel and known small rnas throughout rice plant development focussing on seed tissues. *BMC genomics*, *23*(1), 1–17.

Merkel, D. *et al.* (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux j*, *239*(2), 2.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., Cao, X., Carrington, J. C., Chen, X., Green, P. J. *et al.* (2008). Criteria for annotation of plant micrornas. *The Plant Cell*, *20*(12), 3186–3190.

Meyers Lab, D. D. P. S. C. (2019). MPSS database: List of genome-specific PHAS loci. `https://mpss.danforthcenter.org/phas_index.php`. Accessed: 2019-11-27.

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A. et Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and applications in bioinformatics. *Computational and structural biotechnology journal*, *23*, 2289–2303.

Morgulis, A., Gertz, E. M., Schäffer, A. A. et Agarwala, R. (2006). A fast and symmetric dust implementation to mask low-complexity dna sequences. *Journal of Computational Biology*, *13*(5), 1028–1040.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.-L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. *et al.* (2008). Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem cells. *Genome research*, *18*(4), 610–621.

Nakano, M., McCormick, K., Demirci, C., Demirci, F., Gurazada, S. G. R., Ramachandruni, D., Dusia, A., Rothhaupt, J. A. et Meyers, B. C. (2020). Next-generation sequence databases: Rna and genomic informatics resources for plants. *Plant physiology*, *182*(1), 136–146.

Naso, M. F., Tomkowicz, B., Perry III, W. L. et Strohl, W. R. (2017). Adeno-associated virus (aav) as a vector for gene therapy. *BioDrugs*, *31*(4), 317–334.

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., Voinnet, O. et Jones, J. D. (2006). A plant mirna contributes to antibacterial resistance by repressing auxin signaling. *Science*, *312*(5772), 436–439.

Nellore, A., Collado-Torres, L., Jaffe, A. E., Alquicira-Hernández, J., Wilks, C., Pritt, J., Morton, J., Leek, J. T. et Langmead, B. (2016). Rail-rna: Scalable analysis of rna-seq splicing and coverage. *Bioinformatics*, p. btw575.

Nelson, D. L. et Cox, M. M. (2000). *Lehninger Principles of Biochemistry* (3rd éd.). New York: Worth Publishers.

O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. et Baltimore, D. (2007). Microrna-155 is induced during the macrophage inflammatory response. *Proceedings of the National Academy of Sciences*, *104*(5), 1604–1609.

Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P. et Ragoussis, J. (2020). Methodologies for transcript profiling using long-read technologies. *Frontiers in genetics*, *11*, 606.

Pan, R.-H., Tseng, L.-Y., Liao, I.-E., Chan, C.-L., Lai, K. R. et Lin, K.-B. (2015). Design of an ngs microrna predictor using multilayer hierarchical mapreduce framework. Dans *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1–8. IEEE.

Pan, Y., Li, Y., Liu, Z., Zou, J. et Li, Q. (2022). Computational genomics insights into cold acclimation in wheat. *Frontiers in Genetics*, *13*.

Pantano, L. (2022). Degreport: Report of deg analysis. r package version 1.34.0. `http://lpantano.github.io/DEGreport/`.

Pargent, F., Pfisterer, F., Thomas, J. et Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, *37*(5), 2671–2692.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. et Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, K., Tian, Y., Sun, X., Song, C., Ren, Z., Bao, Y., Xing, J., Li, Y., Xu, Q., Yu, J. *et al.* (2021). tae-mir399-ubc24 module enhances freezing tolerance in winter wheat via a cbf signaling pathway. *Journal of Agricultural and Food Chemistry*, *69*(45), 13398–13415.

Piriyapongsa, J. et Jordan, I. K. (2008). Dual coding of sirnas and mirnas by plant transposable elements. *Rna*, *14*(5), 814–821.

Place, R. F., Li, L.-C., Pookot, D., Noonan, E. J. et Dahiya, R. (2008). Microrna-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences*, *105*(5), 1608–1613.

Pola-Sánchez, E., Hernández-Martínez, K. M., Pérez-Estrada, R., Sélem-Mójica, N., Simpson, J., Abraham-Juárez, M. J., Herrera-Estrella, A. et Villalobos-Escobedo, J. M. (2024). Rna-seq data analysis: A practical guide for model and non-model organisms. *Current Protocols*, *4*(5), e1054.

Potter, M., Lins, B., Mietzsch, M., Heilbronn, R., Van Vliet, K., Chipman, P., Agbandje-McKenna, M., Cleaver, B. D., Clément, N., Byrne, B. J. *et al.* (2014). A simplified purification protocol for recombinant adeno-associated virus vectors. *Molecular Therapy-Methods & Clinical Development*, *1*, 14034.

Praat, M., De Smet, I. et van Zanten, M. (2021). Protein kinase and phosphatase control of plant temperature responses. *Journal of Experimental Botany*, *72*(21), 7459–7473.

Qu, J., Ye, J. et Fang, R. (2007). Artificial microrna-mediated virus resistance in plants. *Journal of virology*, *81*(12), 6690–6699.

Rahim, A. A., Uzair, M., Rehman, N., Rehman, O. U., Zahra, N. et Khan, M. R. (2022). Genome-wide identification and characterization of receptor-like protein kinase 1 (rpk1) gene family in triticum aestivum under drought stress. *Frontiers in Genetics*, *13*.

Rahmdani, F., Remita, A. M., Wu, C.-J. et Diallo, A. B. (2020). Assessing the evidence of published mirnas in wheat (triticum aestivum). Dans *The 24th Annual International Conference on Research in Computational Molecular Biology, June 22-25, 2020, Virtual*, p. poster 415. RECOMB.

Ray, K. K., Wright, R. S., Kallend, D., Koenig, W., Leiter, L. A., Raal, F. J., Bisch, J. A., Richardson, T., Jaros, M., Wijngaard, P. L. *et al.* (2020). Two phase 3 trials of inclisiran in patients with elevated ldl cholesterol. *New England journal of medicine*, *382*(16), 1507–1519.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. et Bartel, D. P. (2002). Micrornas in plants. *Genes & development*, *16*(13), 1616–1626.

Remita, M. A., Lord, E., Agharbaoui, Z., Leclercq, M., Badawi, M. A., Sarhan, F. et Diallo, A. B. (2016). A novel comprehensive wheat mirna database, including related bioinformatics software. *Current Plant Biology*, *7*, 31–33.

Ronquist, F., Huelsenbeck, J. P. et Teslenko, M. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Rorbach, G., Unold, O. et Konopka, B. M. (2018). Distinguishing mirtrons from canonical mirnas with data exploration and machine learning methods. *Scientific reports*, *8*(1), 7560.

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., Voelkerding, K. V. *et al.* (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. *The Journal of Molecular Diagnostics*, *20*(1), 4–27.

Sawicki, M., Rondeau, M., Courteaux, B., Rabenoelina, F., Guerriero, G., Gomès, E., Soubigou-Taconnat, L., Balzergue, S., Clément, C., Ait Barka, E. *et al.* (2019). On a cold night: transcriptomics of grapevine flower unveils signal transduction and impacted metabolism. *International journal of molecular sciences*, *20*(5), 1130.

Schumacher, A., Pireddu, L., Niemenmaa, M., Kallio, A., Korpelainen, E., Zanetti, G. et Heljanko, K. (2014). Seqpig: simple and scalable scripting for large sequencing data sets in hadoop. *Bioinformatics*, *30*(1), 119–120.

Shirane, K., Toh, H., Kobayashi, H., Miura, F., Chiba, H., Ito, T., Kono, T. et Sasaki, H. (2013). Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-cpg methylation and role of dna methyltransferases. *PLoS genetics*, *9*(4), e1003439.

Shivaprasad, P. V., Chen, H.-M., Patel, K., Bond, D. M., Santos, B. A. et Baulcombe, D. C. (2012). A microrna superfamily regulates nucleotide binding site–leucine-rich repeats and other mrnas. *The Plant Cell*, *24*(3), 859–874.

So, J. B. Y., Kapoor, R., Zhu, F., Koh, C., Zhou, L., Zou, R., Tang, Y. C., Goo, P. C., Rha, S. Y., Chung, H. C. *et al.* (2021). Development and validation of a serum microrna biomarker panel for detecting gastric cancer in a high-risk population. *Gut*, *70*(5), 829–837.

Solanke, A. U. et Sharma, A. K. (2008). Signal transduction during cold stress in plants. *Physiology and Molecular biology of plants*, *14*(1), 69–79.

Song, G., Zhang, R., Zhang, S., Li, Y., Gao, J., Han, X., Chen, M., Wang, J., Li, W. et Li, G. (2017). Response of microrna s to cold treatment in the young spikes of common wheat. *BMC genomics*, *18*(1), 212.

Song, X., Wang, D., Ma, L., Chen, Z., Li, P., Cui, X., Liu, C., Cao, S., Chu, C., Tao, Y. *et al.* (2012). Rice rna-dependent rna polymerase 6 acts in small rna biogenesis and spikelet development. *The Plant Journal*, *71*(3), 378–389.

Spark News Archive (2016). Spark wins CloudSort Benchmark as the most efficient engine. `https://spark.apache.org/news/spark-wins-cloudsort-100tb-benchmark.html`. Accessed: March 19, 2025.

Stief, A., Brzezinka, K., Lämke, J. et Bäurle, I. (2014). Epigenetic responses to heat stress at different time scales and the involvement of small rnas. *Plant Signaling & Behavior*, *9*(10), e970430.

Sun, F., Guo, W., Du, J., Ni, Z., Sun, Q. et Yao, Y. (2013). Widespread, abundant, and diverse te-associated sirnas in developing wheat grain. *Gene*, *522*(1), 1–7.

Sun, Z., Evans, J., Bhagwate, A., Middha, S., Bockol, M., Yan, H. et Kocher, J.-P. (2014). Cap-mirseq: a comprehensive analysis pipeline for microrna sequencing data. *BMC genomics*, *15*, 1–10.

Sunkar, R., Kapoor, A. et Zhu, J.-K. (2006). Posttranscriptional induction of two cu/zn superoxide dismutase genes in arabidopsis is mediated by downregulation of mir398 and important for oxidative stress tolerance. *The Plant Cell*, *18*(8), 2051–2065.

Sunkar, R., Li, Y.-F. et Jagadeeswaran, G. (2012). Functions of microrna s in plant stress responses. *Trends in plant science*, *17*(4), 196–203.

Tai, P. W., Xie, J., Fong, K., Seetin, M., Heiner, C., Su, Q., Weiand, M., Wilmot, D., Zapp, M. L. et Gao, G. (2018). Adeno-associated virus genome population sequencing achieves full vector genome resolution and reveals human-vector chimeras. *Molecular Therapy-Methods & Clinical Development*, *9*, 130–141.

Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M. et Watanabe, Y. (2008). The mechanism selecting the guide strand from small rna duplexes is different among argonaute proteins. *Plant and cell physiology*, *49*(4), 493–500.

Tang, Z., Zhang, L., Xu, C., Yuan, S., Zhang, F., Zheng, Y. et Zhao, C. (2012). Uncovering small rna-mediated responses to cold stress in a wheat thermosensitive genic male-sterile line by deep sequencing. *Plant physiology*, *159*(2), 721–738.

Tiwari, B., Habermann, K., Arif, M. A., Weil, H. L., Garcia-Molina, A., Kleine, T., Mühlhaus, T. et Frank, W. (2020). Identification of small rnas during cold acclimation in arabidopsis thaliana. *BMC plant biology*, *20*(1), 1–25.

Tran, N. T., Heiner, C., Weber, K., Weiand, M., Wilmot, D., Xie, J., Wang, D., Brown, A., Manokaran, S., Su, Q. *et al.* (2020). Aav-genome population sequencing of vectors packaging crispr components reveals design-influenced heterogeneity. *Molecular Therapy-Methods & Clinical Development*, *18*, 639–651.

Tran, N. T., Lecomte, E., Saleun, S., Namkung, S., Robin, C., Weber, K., Devine, E., Blouin, V., Adjali, O., Ayuso, E. *et al.* (2022). Human and insect cell-produced recombinant adeno-associated viruses show differences in genome heterogeneity. *Human Gene Therapy*, *33*(7-8), 371–388.

Tribolet, L., Kerr, E., Cowled, C., Bean, A. G., Stewart, C. R., Dearnley, M. et Farr, R. J. (2020). Microrna biomarkers for infectious diseases: from basic research to biosensing. *Frontiers in Microbiology*, *11*, 1197.

Tuteja, N., Singh, S. et Tuteja, R. (2012). Helicases in improving abiotic stress tolerance in crop plants. *Improving crop resistance to abiotic stress*, 435–449.

Valinezhad Orang, A., Safaralizadeh, R., Kazemzadeh-Bavili, M. *et al.* (2014). Mechanisms of mirna-mediated gene regulation from common downregulation to mrna-specific upregulation. *International journal of genomics*, *2014*.

Van der Auwera, G. A. et O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Vaucheret, H. et Voinnet, O. (2024). The plant sirna landscape. *The Plant Cell*, *36*(2), 246–275.

Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S. I. et Moazed, D. (2004). Rnai-mediated targeting of heterochromatin by the rits complex. *Science*, *303*(5658), 672–676.

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T. et Hatzigeorgiou, A. G. (2015). Diana-mirpath v3. 0: deciphering microrna function with experimental support. *Nucleic acids research*, p. gkv403.

Vo, K. T. X., Kim, C.-Y., Chandran, A. K. N., Jung, K.-H., An, G. et Jeon, J.-S. (2015). Molecular insights into the function of ankyrin proteins in plants. *Journal of plant biology*, *58*(5), 271–284.

Wang, B., Sun, Y.-f., Song, N., Wei, J.-p., Wang, X.-j., Feng, H., Yin, Z.-y. et Kang, Z.-s. (2014). Micrornas involving in cold, wounding and salt stresses in triticum aestivum l. *Plant physiology and biochemistry*, *80*, 90–96.

Wang, D., Tai, P. W. et Gao, G. (2019a). Adeno-associated virus vector as a platform for gene therapy delivery. *Nature reviews Drug discovery*, *18*(5), 358–378.

Wang, H., Niu, H., Liang, M., Zhai, Y., Huang, W., Ding, Q., Du, Y. et Lu, M. (2019b). A wall-associated kinase gene cawakl20 from pepper negatively modulates plant thermotolerance by reducing the expression of aba-responsive genes. *Frontiers in plant science*, p. 591.

Wang, Q., Liu, N., Yang, X., Tu, L. et Zhang, X. (2016). Small rna-mediated responses to low-and high-temperature stresses in cotton. *Scientific reports*, *6*, 35558.

Wang, Y., Hu, Z., Yang, Y., Chen, X. et Chen, G. (2009). Function annotation of an sbp-box gene in arabidopsis based on analysis of co-expression networks and promoters. *International journal of molecular sciences*, *10*(1), 116–132.

Wilusz, J. E., Sunwoo, H. et Spector, D. L. (2009). Long noncoding rnas: functional surprises from the rna world. *Genes & development*, *23*(13), 1494–1504.

Wright, J. F. (2008). Manufacturing and characterizing aav-based vectors for use in clinical studies. *Gene therapy*, *15*(11), 840–848.

Wu, C.-J., Danyluk, J. et Diallo, A. B. (2020). Methodes d'analyse fonctionnelle des petits arns – siwalk: Pipeline ngs de prédiction de siarn chez les plantes. Dans *Journee de la recherche de UQAM 2e edition, March 20, 2020, Montreal*, p. posterC5. UQAM.

Wu, C.-J., Remita, A. M. et Diallo, A. B. (2019). Mirlibspark: A scalable ngs plant microrna prediction pipeline for multi-library functional annotation. Dans *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 669–674. ACM.

Xie, F., Stewart Jr, C. N., Taki, F. A., He, Q., Liu, H. et Zhang, B. (2014). High-throughput deep sequencing shows that micro rna s play important roles in switchgrass responses to drought and salinity stress. *Plant Biotechnology Journal*, *12*(3), 354–366.

Xie, M. et Yu, B. (2015). sirna-directed dna methylation in plants. *Current genomics*, *16*(1), 23–31.

Xu, S., Liu, N., Mao, W., Hu, Q., Wang, G. et Gong, Y. (2016). Identification of chilling-responsive microrna s and their targets in vegetable soybean (glycine max l.). *Scientific reports*, *6*, 26619.

Yadav, S. K. (2010). Cold stress tolerance mechanisms in plants. a review. *Agronomy for sustainable development*, *30*(3), 515–527.

Yang, L., Wu, K., Gao, P., Liu, X., Li, G. et Wu, Z. (2014). Gslrpk, a novel cold-activated leucine-rich repeat receptor-like protein kinase from glycine soja, is a positive regulator to cold stress tolerance. *Plant Science*, *215*, 19–28.

Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., Miao, J. et Dong, Y. (2020). High-throughput transcriptome profiling in drug and biomarker discovery. *Frontiers in genetics*, *11*, 19.

Yang, X. et Li, L. (2011). mirdeep-p: a computational tool for analyzing the microrna transcriptome in plants. *Bioinformatics*, *27*(18), 2614–2615.

Yao, Y., Ni, Z., Peng, H., Sun, F., Xin, M., Sunkar, R., Zhu, J.-K. et Sun, Q. (2010). Non-coding small rnas responsive to abiotic stress in wheat (triticum aestivum l.). *Functional & integrative genomics*, *10*, 187–190.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R. *et al.* (2020). Ensembl 2020. *Nucleic acids research*, *48*(D1), D682–D688.

Yu, H.-F., Huang, F.-L. et Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*, 41–75.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S. et Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Dans *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, 2–2. USENIX Association.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S. et Stoica, I. (2010). Spark: Cluster computing with working sets. Dans *2nd USENIX workshop on hot topics in cloud computing (HotCloud 10)*.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J. *et al.* (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65.

Zeng, X., Xu, Y., Jiang, J., Zhang, F., Ma, L., Wu, D., Wang, Y. et Sun, W. (2018). Identification of cold stress responsive micrornas in two winter turnip rape (brassica rapa l.) by high throughput sequencing. *BMC Plant Biology*, *18*, 1–13.

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V. et Meyers, B. C. (2015). Spatiotemporally dynamic, cell-type–dependent premeiotic and meiotic phasirnas in maize anthers. *Proceedings of the National Academy of Sciences*, *112*(10), 3146–3151.

Zhang, C., Li, G., Wang, J. et Fang, J. (2012a). Identification of trans-acting sirnas and their regulatory cascades in grapevine. *Bioinformatics*, *28*(20), 2561–2568.

Zhang, H., Wu, T., Li, Z., Huang, K., Kim, N.-E., Ma, Z., Kwon, S.-W., Jiang, W. et Du, X. (2021). Osgata16, a gata transcription factor, confers cold tolerance by repressing oswrky45–1 at the seedling stage in rice. *Rice*, *14*(1), 1–15.

Zhang, H. et Zhu, J.-K. (2011). Rna-directed dna methylation. *Current opinion in plant biology*, *14*(2), 142–147.

Zhang, J., Xu, Y., Huan, Q. et Chong, K. (2009). Deep sequencing of brachypodium small rnas at the global genome level identifies micrornas involved in cold stress response. *BMC genomics*, *10*(1), 449.

Zhang, X., Xia, J., Lii, Y. E., Barrera-Figueroa, B. E., Zhou, X., Gao, S., Lu, L., Niu, D., Chen, Z., Leung, C. *et al.* (2012b). Genome-wide analysis of plant nat-sirnas reveals insights into their distribution, biogenesis and function. *Genome biology*, *13*(3), R20.

Zhang, Y., Huang, J., Xie, F., Huang, Q., Jiao, H. et Cheng, W. (2024). Identification of plant micrornas using convolutional neural network. *Frontiers in Plant Science*, *15*, 1330854.

Zhang, Z., Teotia, S., Tang, J. et Tang, G. (2019). Perspectives on micrornas and phased small interfering rnas in maize (zea mays l.): functions and big impact on agronomic traits enhancement. *Plants*, *8*(6), 170.

Zhao, J., Liu, Q., Hu, P., Jia, Q., Liu, N., Yin, K., Cheng, Y., Yan, F., Chen, J. et Liu, Y. (2016). An efficient potato virus x-based microrna silencing in nicotiana benthamiana. *Scientific reports*, *6*, 20573.

Zheng, Y., Wang, S. et Sunkar, R. (2014). Genome-wide discovery and analysis of phased small interfering rnas in chinese sacred lotus. *PloS one*, *9*(12), e113790.

Zhou, B., Kang, Y., Leng, J. et Xu, Q. (2019). Genome-wide analysis of the mirna–mrnas network involved in cold tolerance in populus simonii$\times$ p. nigra. *Genes*, *10*(6), 430.

Zhu, Y., Qian, W. et Hua, J. (2010). Temperature modulates plant defense responses through nb-lrr proteins. *PLoS pathogens*, *6*(4), e1000844.

Zou, Q., Li, X.-B., Jiang, W.-R., Lin, Z.-Y., Li, G.-L. et Chen, K. (2013). Survey of mapreduce frame operation in bioinformatics. *Briefings in bioinformatics*, p. bbs088.

Zuo, Z., Hu, H., Xu, Q., Luo, X., Peng, D., Zhu, K., Zhao, Q., Xie, Y. et Ren, J. (2020). Bbcancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic acids research*, *48*(D1), D789–D796.