

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉVALUATION DES MÉTHODES D'ANALYSES D'EXPRESSION
DIFFÉRENTIELLE DES PETITS ARNS

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR
NDIAYE DIOP

JUILLET 2019

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire 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.07-2011). 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.»

DÉDICACES

Mes premières pensées se tournent vers mon père et mon grand-père, des hommes généreux et avides de connaissances. Ces deux grands hommes m'ont éduqué dans le droit chemin et m'ont aidé à fréquenter l'école. Je regrette juste qu'ils ne soient plus là aujourd'hui pour voir le résultat de mon travail. En honneur à leur mémoire, je tiens à leur dédier ce présent manuscrit. À ma très chère mère que Dieu la garde en bonne santé. À toute ma famille et mes amis. À tous ceux qui sont proches de mon cœur et dont je n'ai pas cité leur nom. À toutes personnes malades. À mon village natal. je dédie ce modeste travail.

Ndiaye DIOP

REMERCIEMENTS

Je tiens à remercier d'abord mon directeur de recherche Monsieur Abdoulaye Banniré DIALLO qui a consacré beaucoup de temps pour répondre à toutes mes questions et m'a aidé à résoudre les problèmes que j'ai rencontrés pour rédiger ce manuscrit. Je le remercie aussi pour son apport scientifique et ses précieux conseils tout au long de mon mémoire. Je tiens à remercier Madame Halima ELbiaze qui m'a beaucoup aidé sur les démarches administratives et durant ma première session de maîtrise à l'UQAM. Mes sincères gratitude à tous les membres du laboratoire bioinformatique de l'UQAM surtout à Nadia Tahiri et Etienne Lord. Mes remerciements s'adressent à tous les membres du jury pour l'honneur qu'ils m'ont fait en acceptant d'évaluer ce travail. Ma gratitude va également à mes oncles, frères, soeurs, tantes, etc pour leur soutien considérable durant mes études. Je tiens à remercier aussi Mouhamed Dieye, Mamadou Cissé, mes camarades, amis étudiants Sénégalais vivant à Montréal pour leur intégration. J'adresse aussi mes sincères remerciements à toute ma famille surtout à mes parents avec une mention très spéciale à ma mère pour sa patience et à : Ibrahima Diop, Mamadou Diop, Mor S. Diop , Mor Cissé, Ndiaye Diop, Meissa Mbaye, Serigne Mor Diao, Serigne Oumar Faroukh Diop, Ndèye Mbengue, S. M. Gueye, Mamadou Moustapha Sarr... qui m'ont soutenu durant mes années d'étude. Ma liste n'est pas exhaustive, j'oublie certainement des personnes. Je tiens seulement à ce qu'elles sachent que si leur nom n'est pas cité sur le papier, il est présent dans mes souvenirs.

Ndiaye DIOP

TABLE DES MATIÈRES

LISTE DES FIGURES	xi
LISTE DES TABLEAUX	xiii
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES . .	xv
RÉSUMÉ	xvii
CHAPITRE I	
INTRODUCTION	1
1.1 Mise en contexte	1
1.2 Motivation	2
1.3 Problématique	2
1.4 Méthodologie	3
1.5 Contribution et organisation du mémoire	4
CHAPITRE II	
CONCEPTS GÉNOMIQUES ET MICROARNS	7
2.1 Introduction	7
2.2 Biologie des acides ribonucléiques (ARNs)	7
2.3 Biologie des micro acides ribonucléiques (microARNs)	10
2.3.1 Biogenèse des micro acides ribonucléiques (microARNs)	11
2.3.2 Mode d'action des microARNs	12
2.4 Régulation par les microARNs	13
2.5 Fonctions des microARNs	15
2.5.1 Rôle des microARNs chez les plantes	15
2.5.2 Rôle des microARNs chez les animaux	17
2.5.3 Bilan des microARNs et leurs fonctions	22
2.6 Détermination de leur profil d'expression	23
2.7 Conclusion	25

CHAPITRE III	
REVUE DE LA LITTÉRATURE	27
3.1 Introduction	27
3.2 Méthodes paramétriques	27
3.2.1 DESeq2	28
3.2.2 Sleuth	28
3.2.3 EdgeR	29
3.2.4 Limma et Voom	30
3.2.5 BaySeq	30
3.2.6 DESeq	31
3.3 Méthodes non paramétriques	32
3.3.1 NOIseq	32
3.3.2 SAMseq	32
3.4 Autres méthodes	33
3.4.1 EBSeq	33
3.4.2 Cuffdiff2	34
3.4.3 Approche intégrée	35
3.5 Observations sur les méthodes	35
3.6 Type de normalisation des approches	37
3.7 Conclusion	38
CHAPITRE IV	
PROPOSITION D'UNE APPROCHE D'EXPRESSION DIFFÉRENTIELLE	41
4.1 Introduction	41
4.2 Modèle du pipeline DESeq2	42
4.3 Implémentation	43
4.3.1 Principaux objectifs de notre approche	43
4.3.2 Structure du Workflow DESeq2	44
4.4 Données de simulation	47

4.5 Conclusion	49
CHAPITRE V	
PRÉSENTATION DES RÉSULTATS	53
5.1 Introduction	53
5.2 Résultats de simulation	53
5.2.1 Résultats de normalisation	54
5.2.2 Résultats d'estimation des dispersions	57
5.2.3 Résultats des valeurs de dispersions	63
5.2.4 Résultats des méthodes et tests d'expression différentielle . . .	66
5.3 Discussion et interprétation	73
5.4 Limites de notre approche	75
5.5 Conclusion	76
CHAPITRE VI	
CONCLUSION ET PERSPECTIVES	77
APPENDICE A	
CODE RSTUDIO IMPLÉMENTÉ	79
BIBLIOGRAPHIE	83

LISTE DES FIGURES

Figure	Page
2.1 Éléments structuraux de la structure secondaire de l'ARN (Dardel et Képès, 2002)	9
2.2 Exemple de représentation d'une structure secondaire d'un ARN (Dardel et Képès, 2002)	10
2.3 Biogenèse des microARNs (BRET et SCHVED, 2009)	12
2.4 Mode d'action des microARNs (BRET et SCHVED, 2009)	13
2.5 Approche d'identification des gènes cibles et des fonctions des microARNs chez les animaux (Wienholds et Plasterk, 2005)	19
2.6 Différents niveaux de régulation de la tumorigenèse par les microARNs (Kong <i>et al.</i> , 2012)	22
4.1 Adaptation d'un Workflow d'expression différentielle basé sur DE-Seq2	45
4.2 Visualisation de la base de données : On observe dans cette base de données contient 15 colonnes dont les deux premières (Mirnas et Cluster) sont des variables qualitatives. Les 13 autres colonnes sont des variables quantitatives continues, normalisées sur un intervalle de [0,1] en fonction du temps	48
4.3 Illustration des périodes de prélèvement des microARNs durant la période allant de septembre à Novembre. La température de l'air et du sol sont donnés à titre indicatif en fonction des temps de prélèvement. (Li <i>et al.</i> , 2018)	49
5.1 Histogramme normalisé au temps T1	56
5.2 Histogramme normalisé au temps T5	56
5.3 Histogramme normalisé aux temps T1 et T5	57
5.4 Estimation des dispersions au temps T1	58

5.5	Estimation des dispersions au temps T1	58
5.6	Estimation des dispersions au temps T1	59
5.7	Estimation des dispersions au temps T1	59
5.8	Estimation des dispersions au temps T5	60
5.9	Estimation des dispersions au temps T5	60
5.10	Estimation des dispersions aux temps T1 et T5	61
5.11	Estimation des dispersions aux temps T1 et T5	61
5.12	Estimation des dispersions aux temps T1 et T5	62
5.13	Estimation des dispersions aux temps T1 et T5	62
5.14	Estimation des dispersions aux temps T1 et T5	63
5.15	Estimation des dispersions aux temps T1 et T5	63

LISTE DES TABLEAUX

Tableau	Page
2.1 Les différentes espèces d'ARN présentes dans la cellule et leurs rôles respectifs (Wienholds et Plasterk, 2005)	8
2.2 Fonctions biologiques des microARNs chez les animaux et les maladies (Wienholds et Plasterk, 2005)	18
2.3 Exemples de microARN dont la fonction biologique et/ou les cibles ARNm sont caractérisées (Cavaillé, 2004)	23
3.1 Résumé 1 des principales observations des méthodes (Wienholds et Plasterk, 2005)	36
3.2 Résumé 2 des principales observations des méthodes (Wienholds et Plasterk, 2005)	37
3.3 Méthodes et normalisation d'expression différentielle des RNA-Seq (Costa-Silva <i>et al.</i> , 2017)	38
5.1 Résultats des 10 premières observations des valeurs de dispersions temps T1. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T1-1 à T1-4)	64
5.2 Résultats des 10 premières observations des valeurs de dispersions temps T5. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T5-1 et T5-2)	65
5.3 Résultats des 10 premières observations des valeurs de dispersions temps T1 et T5. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T1-T5-1 à T1-T5-6)	65
5.4 Tableau des variables	67
5.5 Résultats des 10 premières observations statistiques et expression différentielle temps T1-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-1)	67

5.6	Résultats des 10 premières observations statistiques et expression différentielle temps T1-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-2)	68
5.7	Résultats des 10 premières observations statistiques et expression différentielle temps T1-3. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-3)	68
5.8	Résultats des 10 premières observations statistiques et expression différentielle temps T1-4. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-4)	69
5.9	Résultats des 10 premières observations statistiques et expression différentielle temps T5-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T5-1)	69
5.10	Résultats des 10 premières observations statistiques et expression différentielle temps T5-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T5-2)	70
5.11	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-1)	70
5.12	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-2)	71
5.13	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-3. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-3)	71
5.14	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-4. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-4)	72
5.15	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-5. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-5)	72
5.16	Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-6. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-6)	73

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

3'-UTR three prime Untranslated Transcribed Region

ADNc Acide désoxyribonucléique complémentaire

ADN Acide Désoxyribonucléique

ARN Acide Ribonucléique

BN Binomiale Négative

DCL1 Dicer-Like1

DE Differential Expression

EP Epithelial

ES Embryonic Stem

FDR False Discovery Rate

GLM Generalized linear model

Hid Head involution defective

log2FC logarithmic 2 Fold Change

PCR Polymerase Chain Reaction

RLE Relative Log Expression

RPKM Read Per Kilobase per million Mapped

TMM Trimmed Mean of M-values

TPR True Positive Rate

RÉSUMÉ

Les microARNs appartiennent à la famille des petits ARNs non-codants et agissent comme inhibiteurs des ARN messagers. Dans les dernières années, de nombreux microARNs et leurs cibles ont été découverts chez les mammifères et les plantes. La bioinformatique joue un rôle important dans ce domaine, et des programmes informatiques de découvertes de cibles ont été mis à la disposition de la communauté scientifique. Les microARNs peuvent réguler chacun des centaines de gènes, et les profils d'expression de ces derniers peuvent servir comme classificateurs de certaines maladies. Des avancées technologiques de séquençage à haut débit comme RNA-Seq, chIP-Seq, HT-Seq, fournissent des données quantitatives sous forme de comptage. La croissance des données séquencées est volumineuse, d'où une analyse plus quantitative pourrait être utile pour extraire des connaissances. Pour analyser l'expression différentielle, des tests et des modèles statistiques sont nécessaires pour estimer la variabilité différentielle dans le comptage des données. Des outils de mesure d'expression de gènes ne nécessitent pas une connaissance a priori profonde sur les données de séquences génomiques ont été développés. Ces approches permettent de faire une analyse d'expression différentielle génique des données plus particulièrement celles des microARNs. Dans ce rapport, nous présentons une approche basée sur la distribution binomiale négative (NB) dont la variance et la moyenne sont liées par une régression locale. Nous développons une implémentation dans DESeq2 comme package R/ Bioconductor. Ce dernier intègre des avancées méthodologiques avec plusieurs nouvelles fonctionnalités pour faciliter une analyse plus quantitative des données comparatives de microARNs. Ce processus d'analyse différentielle inclut trois étapes : la normalisation, l'estimation des dispersions et les tests d'expression différentielle. La pertinence de cette approche réside de son applicabilité à tous les types de données génomiques, protéiques, etc. Elle est aussi utilisable aux données de petites échelles avec un taux faible de réplicats. Cette méthode proposée modélise les comptages et tient compte aussi des sur-dispersions et des valeurs aberrantes que contiennent ces données de microARNs.

Mots clés : microARNs, expression différentielle, DESeq2, distribution binomiale négative

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte

L'expression différentielle permet d'identifier des gènes les plus significativement affectés par des variables explicatives contrôlées. Les données utilisées ont souvent une taille spectaculaire à haut débit. On parle aussi de données OMICS, en référence à des données recueillies à l'échelle du génome (genomics), du transcriptome (transcriptomics), du protéome (proteomics), du métabolome (metabolomics), etc. L'étude de l'expression génique consiste à caractériser et quantifier les produits d'expression de l'ADN (les ARN messagers : le transcriptome) de manière à identifier, dans un tissu, dans un état et à un moment donné du développement, les séquences actives et donc à révéler ainsi le niveau d'expression des gènes dont elles sont issues. Le séquençage à haut débit de fragments d'ARN est utilisé dans une gamme d'essais quantitatifs. Une caractéristique commune entre ces différentes plates-formes technologiques est qu'elles séquencent de grandes quantités de fragments d'ADN. En règle générale, ces lectures sont attribuées à une classe en fonction de leur correspondance avec une région commune du génome cible, chaque classe représentant un transcrit cible, dans le cas d'ARN-Seq. Le niveau d'expression est une donnée quantitative, et donc exploitable statistiquement. Une statistique récapitulative importante est le nombre de lectures dans une classe pour l'ARN-Seq. Ce nombre de lectures a une bonne approximation et est linéairement

lié à l'abondance du transcrit cible.

1.2 Motivation

L'intérêt de ce sujet recherche est basée sur la comparaison des profils d'expression différentielle entre des types de microARNs du blé dans des conditions biologiques différentes. La comparaison est effectuée séparément entre les variables de la base de données aux temps T1 et T5 durant les années 2010 et 2013. Nous mettons en œuvre des tests statistiques basés sur DESeq2 pour déterminer si, pour un microARN donné, une différence observée dans le nombre de lectures est significative, c'est-à-dire s'il est supérieur à celui attendu simplement en raison de la variation aléatoire naturelle. Les lectures et le comptage des données de microARNs suivent une distribution linéaire généralisée, qui peut être approximée par une distribution de poisson ou binomiale négative. En conséquence, cette dernière distribution a été utilisée pour tester l'expression différentielle des données microARNs du blé. Cette distribution possède des paramètres importants pour trouver l'expression différentielle, qui est uniquement déterminé par sa moyenne ; sa variance et toutes les autres propriétés en découlent en particulier. Cependant, il a été noté que la distribution binomiale prédit des variations très importantes dans ces genres d'étude de données microARNs. Par conséquent, les tests statistiques résultants de cette approche de DESeq2 contrôle bien l'erreur de type I (False Discovery Rate) annoncée dans l'analyse des données de microARNs.

1.3 Problématique

La détection de microARNs différentiellement exprimés met souvent en jeu des tests statistiques classiques. Cependant, le volume des données peut poser problème en termes de temps de calcul, de fiabilité statistique des résultats ainsi que de leur lisibilité. Des adaptations des outils d'analyse de données sont par

conséquent mises en œuvre afin de pallier à ces problèmes. Dans cette étude, nous avons quatre variétés de blé dans la base donnée : **Manitou**, **Winter**, **Isopring** et **Norstar**. Ces données de microARNs ont été produites aux années **2010** et **2013** entre deux variables temporelles **T1** et **T5**. La question de recherche est déterminer les microARNs qui sont différentiellement exprimés sur ces quatre catégories de blé. Nous verrons aussi la variabilité et le comportement que possèdent les séquences des microARNs du blé d'une période temporelle à une autre. L'extraction de toutes ces connaissances mène notre réflexion à faire une étude comparative pour évaluer des méthodes d'analyse d'expression différentielle l'expression différentielle des petits ARNS du blé dans ce projet de recherche.

1.4 Méthodologie

Nous présentons une approche basée sur DESeq2 , successeur de la méthode ancienne DESeq. Le DESeq2 intègre les avancées méthodologiques et plusieurs nouvelles fonctionnalités pour faciliter une analyse plus quantitative des données comparatives d'ARN-seq, en utilisant des estimateurs pour la dispersion et l'expression des de microARNs. Nous démontrons les avantages des nouvelles fonctionnalités de DESeq2 en décrivant un certain nombre d'entre eux avec des Folds Changes et leurs estimations de l'erreur type 1, permettant un classement et une visualisation améliorée des microARNs, des tests d'hypothèses au-dessus et au-dessous d'un seuil et la transformation en logarithme régularisé pour la qualité des données. Une évaluation et un regroupement des données de comptage sur-dispersés sont mise en exergue dans notre proposition. Cette dernière dont son pouvoir statistique est très important par rapport aux autres outils existants, en révélant que notre méthodologie présente une sensibilité et une précision élevée, tout en contrôlant le taux de faux positifs rencontré au niveau des microARNs.

L'hypothèse nulle est qu'il n'y a pas de différence significative entre les gènes,

c'est-à-dire que notre modèle DESeq2 plus simple suffit à expliquer la variation de l'expression des gènes entre les échantillons. Si la valeur p ajustée pour un gène dépasse un seuil de signification (p. Ex. $P_{adj} < 0,05$), nous devrions envisager d'utiliser un autre modèle plus complexe pour ce gène. En pratique, nous choisirions habituellement modèle DESeq2 et l'appliquerions à tous les microARNs.

1.5 Contribution et organisation du mémoire

Pour mener bien cette question de recherche centrale sur l'analyse d'expression différentielle des petits ARNs du blé, nous proposons de modéliser les données de comptage avec une distribution binomiale négative (NB). Notre approche basée sur le package DESeq2 permet d'analyser les données de microARNs. La distribution NB a des paramètres qui sont uniquement déterminés par la moyenne et la variance. Cependant, le nombre de répétitions dans les ensembles de données d'intérêt est souvent trop petit pour estimer de manière fiable les deux paramètres, moyenne et variance pour chaque microARN du blé. Dans cet manuscrit, nous améliorons notre modèle en permettant des relations de variance et de moyenne plus générales, basées sur les données. Nous fournissons une méthode efficace pour ajuster le modèle aux données et montrons qu'elle offre de meilleurs ajustements des paramètres et résultats d'expression différentielle (**chapitre 5**). En conséquence, une sélection plus équilibrée de microARN exprimés de manière différentielle sur toute la plage dynamique des données peut être obtenue (**section résultats statistiques et test d'expression différentielle**). Nous implémentons notre méthode en l'appliquant à des données de microARNs du blé (**Section Données de simulation**) et nous discutons de la façon dont elle se compare à d'autres approches (**Section Conclusions**). Ce manuscrit est divisé comme suit :

— **Chapitre 1** nous entamons ce rapport par une introduction pour définir

le contexte, les motivations, la problématique et la méthodologie de notre sujet de recherche.

- **Chapitre 2** décrit les concepts biologiques des ARNs et leur rôle essentiel dans les espèces végétales et animales. Une biogenèse des microARNs pour mieux comprendre les principes génomiques de base.
- **Chapitre 3** est un état de l'art sur les principales approches faites pour analyser et interpréter de l'expression différentielle. Les avantages et les inconvénients des approches ont été montrés dans cette revue de la littérature.
- **Chapitre 4** est la partie contribution dans laquelle on retrouve notre workflow du pipeline DESeq2 sur l'expression différentielle. La description et la nature des données microARNs d'expérimentation sont définies.
- **Chapitre 5** est consacré à la présentation des résultats de notre approche. Les résultats de la normalisation, de l'expression différentielle et des tests statistiques de chaque type de comparaison entre les variables sont présentés dans cette partie. La discussion et l'interprétation des résultats suivent ces trois phases importantes afin de pour voir les points forts et faibles de notre approche.
- **Chapitre 6** nous terminons ce rapport par une conclusion pour dégager les points importants qui ont été développés dans les chapitres du rapport, compare nos résultats par rapport aux autres approches d'expression différentielle existantes et sortir des perspectives ultérieures.

CHAPITRE II

CONCEPTS GÉNOMIQUES ET MICROARNS

2.1 Introduction

L'organisme vivant est défini par son génome, qui représente l'ensemble du matériel génétique d'un individu ou d'une espèce particulière, codé sous forme d'ADN ou d'ARN. Dans ce chapitre, nous essayerons de définir les principes génomiques et microARNs sur les espèces animales et végétales. La biogenèse est un facteur intrinsèque pour bien connaître la régulation et les fonctions des microARNs. Une étape importante que nous allons développer est la détermination de profil d'expression des microARNs.

2.2 Biologie des acides ribonucléiques (ARNs)

L'ARN comme l'ADN double brin, les ARNs sont composés de longues chaînes nucléotidiques. Chaque nucléotide possède une base azotée, un sucre ribose, et un groupement phosphate. La production de l'ARN est faite à partir de la phase de transcription de l'ADN dans le noyau des organismes. Cet ARN est composé des mêmes bases azotées que l'ADN hormis l'uracile qui prend la place de la thymine. La base d'uracile contient des propriétés de liaisons envers l'adénine que la thymine. L'ARN peut se replier sur lui-même et prend des conformations différentes grâce à des liaisons chimiques et des forces physiques, qui engendrent des struc-

tures secondaires et tertiaires lors de la prédiction. On peut noter plusieurs types d'ARNs. Les ARNs messagers qui codent pour les protéines et sont communs à tous les organismes vivants et dans certains virus. Ils existent aussi d'autres ARNs impliqués dans la régulation post-transcriptionnelle de l'expression génique, de la réplication de l'ADN et dans d'autres fonctions régulatrices (**Tableau 2.1**).

Tableau 2.1 Les différentes espèces d'ARN présentes dans la cellule et leurs rôles respectifs (Wienholds et Plasterk, 2005)

Type d'ARN	Fonction
ARN messenger(ARNm)	ARN codant, support de l'information pour la synthèse des protéines
ARN ribosomal (ARNr)	À la base de la structure des ribosomes, rôle catalytique au cours de la synthèse protéique
ARN de transfert (ARNt)	Rôle d'adaptateurs entre l'ARNm et les acides aminés constituant les protéines
Petits ARN nucléaires	Impliqués au cours de différents processus nucléaires dont l'épissage alternatif
Petits ARN nucléolaires	ARN capables de modifier la structure des ARNr ou des ARNsn par méthylation ou pseudouridylation
Petits ARN des corps de Cajal	ARN présents dans les nucléoles guidant les modifications des ARNsn ou des ARNsno (les corps de Cajal sont des organites nucléaires)
MicroARN (miARN)	Petites molécules d'ARN non codant capables de réguler l'expression des gènes
Autres ARN non codants	Rôles divers, par exemple : rôle de XIST (X [Inactive]-Specific Transcript) dans l'inactivation du chromosome X, rôle de 7SLRNA dans le complexe ribonucléoprotéique cytoplasmique

Le repliement de la molécule d'ARN forme une structure stable maintenue par des ponts hydrogènes entre les paires de bases A-U, G-U, C-G. Une nomenclature a été faite par ces paires de bases : G-C et A-U pour Watson-Crick (Watson et Crick, 1953) et la paire G-U est appelée Wobble. Pour la stabilité de la molécule d'ARN, certaines liaisons comme G-c sont plus fortes que celles de A-U qui à leur tour sont plus robustes que G-U. On peut noter dans la structure d'ARN des boucles

internes , des boucles terminales (loop), les empilements, les hernies (bulges) et les boucles multiples (**figure 2-1**).

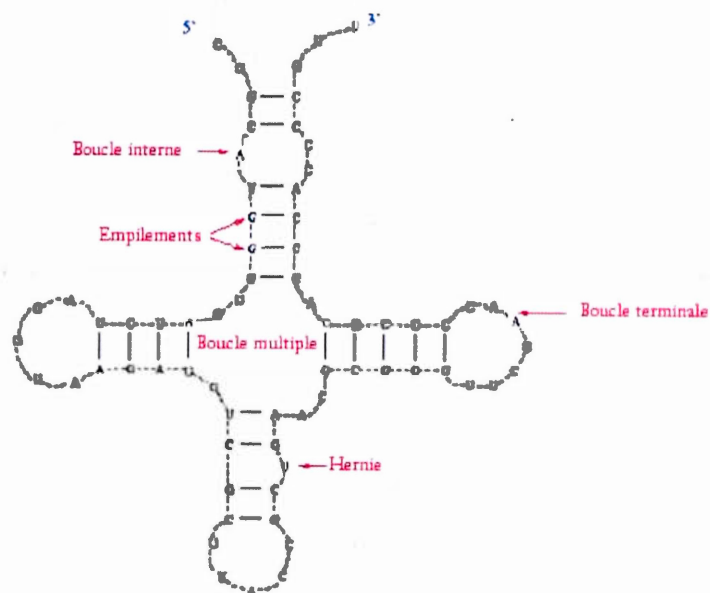


Figure 2.1 Éléments structuraux de la structure secondaire de l'ARN (Dardel et Képès, 2002)

La molécule d'ARN peut être représentée de façon visuelle ou encodée en format parenthèse. Voici les deux cas de formats définis :

1er cas : une paire de parenthèse signifie une liaison entre deux nucléotides.

2em cas : un point correspond à un nucléotide qui n'a pas d'appariement (**figure 2.2**).

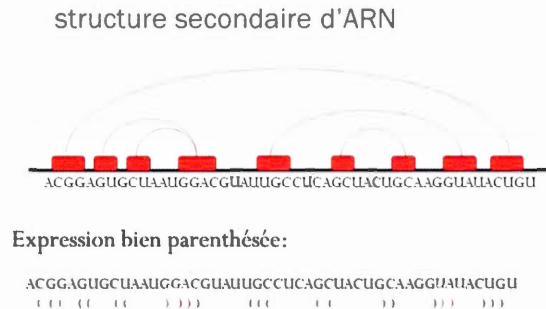


Figure 2.2 Exemple de représentation d’une structure secondaire d’un ARN (Dardel et Képès, 2002)

2.3 Biologie des micro acides ribonucléiques (microARNs)

Les microARNs sont une catégorie particulière de petites molécules d'ARN endogène capables de moduler l'expression des gènes au niveau transcriptionnel ou au niveau transductionnel. On les retrouve ces microARNs aux niveaux des espèces végétales et animales (Bartel, 2004). Les microARNs sont des molécules d'ARN simple brin, de près de 22 nucléotides de longueur de chaîne. Ils sont donc de petite taille, mais se révèlent être relativement abondantes dans les cellules (BRET et SCHVED, 2009). Des centaines de microARNs ont été répertoriées chez l'homme. Il existe une banque de données permettant les recense : la banque miRBase ([http :microrna.sanger.ac.uk/sequences/](http://microrna.sanger.ac.uk/sequences/)). Dans cette banque ou on dépose ces microARNs avec un numéro d'accession, le nom de l'espèce en question. Par exemple : "hsa-miR-101" pour microARN 101 chez l'homosapiens). Certaines molécules de microARNs ont été conservées au cours de l'évolution, permettant ainsi leur rôle essentiel en physiologie, notamment au cours du développement. Ces microARNs régulent l'expression génique en s'appariant avec des ARNm cibles dont ils sont partiellement complémentaires (Hartmann *et al.*, 2004). Ce proces-

sus d'hybridation réprime la traduction de la protéine correspondante ou clive l'ARNm cible au milieu du site de fixation du microARN. Ce dernier mécanisme est très similaire à celui mis en oeuvre lors de l'inférence par l'ARN.

2.3.1 Biogenèse des micro acides ribonucléiques (microARNs)

Cette synthèse des microARNs commence par la transcription d'un long précurseur primaire, appelé "pri-microARN", grâce à l'action de l'ARN polymérase 2. Ce précurseur va être clivé par une ribonucléase, Drosha, et son cofacteur Pasha, pour donner naissance au "pré-miARN", en épingle de cheveux possédant 70 à 100 nucléotides. Ce pré-microARN est exporté dans le cytoplasme via l'exportine 5 et est prise en charge par Dicer, une autre ribonucléase interagissant avec les protéines de la famille Argonautes. Ce processus permet de donner une structure double brin d'environ 22 nucléotides. Le microARN mature est incorporé par la suite dans le complexe RISC (RNA-Induced Silencing Complex), qui va être guidé jusqu'aux ARNm cibles (**figure 2-3**). Dans (Hartmann *et al.*, 2004), nous notons qu'aucun pri-microARN n'a été identifié durant le processus de biogenèse. Un précurseur, généralement plus long que celui observé chez les animaux (60 à 300 nucléotides), est divisé dans le noyau par DCL1 (homologue de la protéine Dicer). La structure double brin du microARN est activement exportée dans le cytoplasme où elle subit une nouvelle étape de maturation qui fait également intervenir une hélicase. La forme simple brin mature est ensuite chargée sur le complexe miRNP/RISC. La fixation par hybridation presque parfaite du microARN sur la séquence codante de l'ARNm cible est responsable du clivage de l'ARNm cible et de l'arrêt de la synthèse de la protéine correspondante.

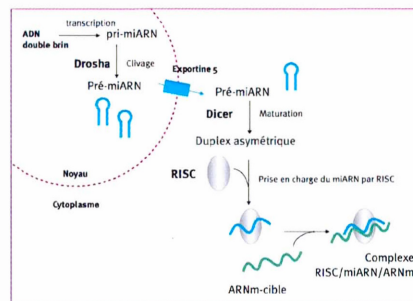


Figure 2.3 Biogenèse des microARNs (BRET et SCHVED, 2009)

2.3.2 Mode d'action des microARNs

On a noté au niveau des microARNs, une parfaite complémentarité de séquence avec les molécules d'ARNm. Nous distinguons aussi deux mécanismes distincts en fonction du degré de complémentarité entre la séquence microARN et celle de l'ARNm cible (**figure 2-3**). La complémentarité est parfaite (végétaux), le complexe RISC joue un rôle d'interaction entre le microARN et l'ARNm, permet l'action de ribonucléases capables de cliver, et donc de dégrader, la molécule d'ARNm. Cette étape agit au niveau transcriptionnel. Dans le cas d'une complémentarité partielle (les mammifères), on constate que l'ARN n'est pas clivé, mais la liaison du microARN aux régions sont traduites en 3' (séquences UTR [Untranslated Regions] est à l'origine d'une inhibition de la traduction en protéine. Dans ces deux types de mécanismes, la conséquence est une extinction de l'expression du gène correspondant (BRET et SCHVED, 2009).

On peut signaler que la capacité pour une molécule de microARN de se lier à l'ARN selon une complémentarité de séquence imparfaite. Cette dernière autorise une multiplicité de cibles possibles. Des algorithmes pouvant prédire jusqu'à 200 gènes de l'expression serait potentiellement contrôlée par une molécule de microARN donnée, ont été développés par les bioinformaticiens. (Esquela-Kerscher

et Slack, 2006), un tiers du génome pourrait être régulé par ces petites molécules d'ARN non codant, ce qui montre l'importance de ce mécanisme de contrôle pour la cellule (**figure 2.4**).

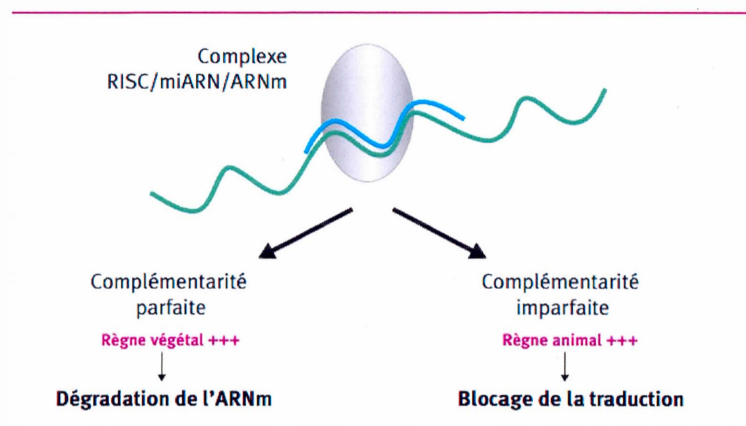


Figure 2.4 Mode d'action des microARNs (BRET et SCHVED, 2009)

2.4 Régulation par les microARNs

Le processus de régulation des microARNs est basé sur une complémentarité plus ou moins parfaite entre la séquence du microARN et celle de sa cible. La partie complémentaire de la cible est généralement située dans la région 3'-UTR de l'ARN. Une limitation de cette complémentarité de 8 à 10 nucléotides en 5' du microARN est notée chez la plupart des animaux. Chez les plantes, cette complémentarité entre le microARN et sa cible est généralement parfaite. Les microARNs, une fois chargés dans le complexe RISC, ont plusieurs méthodes pour réprimer l'expression de leur ARNm cible. Dans (Rogers et Chen, 2013), cette voie de régulation utilisée dépend de la protéine d'Argonaute (AGO) impliquée du niveau de complémentarité entre la séquence du microARN et de sa cible. Chez les plantes, la régulation la plus importante est le clivage de l'ARNm cible.

Après leur clivage de l'ARNm de leur complexe RISC, dû à une dégradation par des exoribonucléases comme XRN1 et XRN4. Ces dernières sont impliquées dans la dégradation des produits de clivage en 3'. La partie 5' est poly-adenylée ou poly-uracilée à son extrémité 3' et est dirigée vers l'exosome pour sa dégradation (Ibrahim *et al.*, 2006). La biogenèse des siARNs s'amorce par le clivage de leur transcrit par un complexe RISC chargé d'un microARN de longueur de 22 nucléotides (Cuperus *et al.*, 2010). Après libération de l'ARNm complémentaire, le complexe RISC est de nouveau apte à s'associer à une nouvelle cible et à introduire sa dégradation.

Chez les animaux, on observe une complémentarité entre le microARN et sa cible est limitée à la région 5' du microARN. L'ARNm cible est destabilisé par le complexe RISC qui induit une dégradation de sa queue poly-adenylée. La protéine Decapping2 (DCP2) catalyse le retrait de la coiffe par un autre complexe Decapping1 (DCP1) et Varicose (VCS) (Xu *et al.*, 2006). En absence de coiffe et de queue poly-adenylée, l'ARNm est moins stable et est susceptible d'être dégradé par des nucléases. Le processus de dégradation de l'ARNm cible est précédé d'une inhibition de sa traduction, un autre mode de régulation des microARNs chez les animaux. La plupart des microARNs chez les animaux utilise cette voie de régulation, et est associée avec un faible niveau de complémentarité avec l'ARNm cible. Par contre chez les plantes, cette voie fait que l'expression de certains ARNm cible ne diminue pas avec la sur-expression de leur microARNs tandis que l'expression de la protéine est diminuée (Chen, 2004) et (Brodersen *et al.*, 2008). Certains microARNs sont capables de déclencher la méthylation de l'ADN et sont à l'origine de la régulation transcriptionnelle. Par exemple chez le riz, des microARNs produits par DCL3 et associés à AGO4 induisent la méthylation des gènes MIR (Wu *et al.*, 2010).

2.5 Fonctions des microARNs

L'un des défis majeurs de la recherche de nouveaux microARNs est d'exploiter leur faculté de cibler de l'ARNm pour altérer le mécanisme de fonctionnement des espèces (par exemple améliorer la résistance du blé au froid). Dans cette partie, le blocage post-transcriptionnel lors de la régulation, induit de nombreuses fonctions biologiques chez les animaux et les plantes. Nous nous limitons de quelques fonctions des microARNs dans cette section.

2.5.1 Rôle des microARNs chez les plantes

Chez les plantes, les microARNs jouent un rôle important dans plusieurs aspects notamment pour le développement des organes, le changement des phases, la traduction du signal et la réponse au stress environnemental. (Reinhart *et al.*, 2002), (Chen, 2005). Ces microARNs ont un haut degré de complémentarité de séquence avec leurs ARNm cibles. Cela a permis de faciliter la prédiction bioinformatique des gènes cibles de microARN (Allen *et al.*, 2005), dont certains ont ensuite été validés expérimentalement pour être de véritables cibles des microARNs. On note que les microARNs de plantes régulent les gènes codant pour différents types de protéines. Une catégorie majeure de gènes cibles de microARN comprend des facteurs de transcription ou d'autres protéines régulatrices qui jouent un rôle dans le développement des plantes (Adai *et al.*, 2005).

2.5.1.1 Signalisation d'auxine

L'auxine est essentielle à la forme de la plante et à sa réponse à l'environnement. Durant des dernières années, nous savons que le rôle de l'auxine dans la formation des motifs décrit dans (Fleming, 2005) et des mécanismes de signalisation de l'auxine a considérablement progressé. La concentration locale en auxine, établie par le transport de l'auxine polaire, semble tracer l'axe de l'embryon, établir

les cellules souches des racines et contrôler la croissance des primordia à partir des méristèmes, de la formation des racines latérales et des réponses gravitropes. L'auxine entraîne la dégradation d'une classe de protéines de répression de la transcription, appelées protéines Aux / IAA, par la voie du protéasome de l'ubiquitine. Les protéines Aux / IAA se hétérodimérisent avec des membres de l'auxin response factor (ARF) de la famille des activateurs et des répresseurs de transcription (ARF), et inhibent les activités des ARF activateurs. L'auxine est liée par TIR1, dans (Dharmasiri *et al.*, 2005) une protéine de la F-box dans la protéine ligase d'ubiquitine SCFTIR1. L'auxine favorise l'interaction entre SCFTIR1 et ses substrats, les protéines Aux / IAA, afin d'entraîner leur dégradation protéolytique.

2.5.1.2 Identité des organes et développement de la reproduction

Les organes floraux sont initiés dans les verticilles du méristème floral. Les identités des primordiums des organes floraux sont spécifiées par les activités combinatoires de trois grandes classes de gènes homéotiques floraux connus sous le nom de gènes A, B et C (Jack, 2004). Les gènes A et C spécifient les identités respectives du périanthe et des organes reproducteurs. Les mutations de perte de fonction du gène AGAMOUS (AG) de classe C entraînent le remplacement des organes reproducteurs par les organes du périanthe, tandis que les mutations de perte de fonction du gène APETALA2 (AP2) de classe A entraînent l'effet inverse, ce qui suggère que les gènes A et les gènes C agissent de manière antagoniste pour se limiter à leurs domaines d'activité dans le méristème floral. En fait, l'ARNm de l'AG ne se trouve que dans les deux tours internes du méristème floral, qui deviendront plus tard les organes de reproduction, tandis que l'ARNm d'APETALA1 d'un gène de classe A, ne se trouve que dans les deux tours externes du méristème qui plus tard deviennent les organes du périanthe. Cependant aussi, la transcription du gène AP2 de classe A est présente dans tout le méristème floral.

2.5.1.3 Croissance des feuilles

Le gène CINCINNATA (CIN) est nécessaire à la régulation différentielle de la division cellulaire pendant la morphogenèse de la feuille pour obtenir une feuille plate (Nath *et al.*, 2003). Les mutants de cin ont des feuilles froissées. Le CIN appartient à la famille des facteurs de transcription (TCP). Chez Arabidopsis, la surexpression de miR319 / Jaw conduit à des feuilles froissées et à une réduction des niveaux de cinq gènes TCP contenant des sites de liaison miR319 / Jaw (Palatnik *et al.*, 2003).

2.5.1.4 Rôles prédictifs des autres microARNs

Les cibles miARN codant pour les facteurs de transcription sont les mieux caractérisées à ce jour, les microARNs cibles codant pour des protéines ayant diverses autres fonctions cellulaires, telles que la dégradation des protéines, le métabolisme et les réponses au stress, ont été prédites et certaines ont été validées à l'aide des critères de clivage *in vivo* dans (Sunkar et Zhu, 2004) et (Jones-Rhoades et Bartel, 2004). Il est donc attendu que les microARNs régulent divers processus cellulaires.

2.5.2 Rôle des microARNs chez les animaux

Diverses fonctions biologiques ont été notées chez les animaux (**tableau 2.2**). Ces fonction des microARNs chez les animaux ont été obtenues par plusieurs approches (**Fig.4.5**)ci-dessous. Il existe une méthode qui consiste à identifier par criblage génétique la perte et le gain de fonction chez *C. elegans* et la drosophile. Étant donné que ces microARNs ont été trouvés en fonction de leurs fonctions biologiques, les fonctions décrites sont probablement les plus précises. Les autres fonctions des microARNs génétiquement attribuées proviennent d'approches génétiques inverses.

Tableau 2.2 Fonctions biologiques des microARNs chez les animaux et les maladies (Wienholds et Plasterk, 2005)

miRNA	Target(s)	Function(s)
lin-4	lin-14, lin-28	Early developmental timing
let-7	lin-41, hbl-1, daf-12 pha-4, ras	Late developmental timing
lsy-6	cog-1	Left/right neuronal asymmetry
miR-273	die-1	Left/right neuronal asymmetry
bantam	hid	Programmed cell death
miR-14	Drice ?	Programmed cell death and fat metabolism
miR-7	Notch targets ?	Notch signaling
miR-430	?	Brain morphogenesis
miR-196	Hoxb8	Developmental patterning
miR-181	?	Hematopoietic lineage differentiation
miR-1	Hand2	Cardiomyocyte differentiation and proliferation
miR-375	Mtpn	Insulin secretion
miR-16	Several	AU-rich element-mediated mRNA instability
miR-32	Retrovirus PFV-1	Antiviral defense
miR-143	Erk5 ?	Adipocyte differentiation
SVmiRNAs	SV40 viral mRNAs	Susceptibility to cytotoxic T cells
miR-15- miR-16	?	Downregulated in B-cell chronic lymphocyte leukemia
miR-143, miR-145	?	Downregulated in colonic adenocarcinoma
miR- 155/BIC	?	Upregulated in diffuse large B-cell lymphoma
let-7	Ras ?	Downregulated in lung cell carcinoma
miR-17-92	?	Upregulated in B-cell lymphoma

Ces approches incluent l'inactivation de microARN (Lee *et al.*, 2005) et des études de sur-expression de microARNs spécifiques. La plupart des gènes cibles de microARNs proviennent de prédictions informatiques. Cependant, une petite partie de ces cibles prédites ont été validées expérimentalement chez des cellules de la Drosophile et de l'humaine (Bentwich, 2005), (Lai, 2004). Plusieurs gènes ont de nombreux sites cibles pour un ou plusieurs microARNs différents. Dans

certains cas, beaucoup de sites cibles sont essentiels pour une régulation adéquate in vitro ou in vivo (Krek *et al.*, 2005), (Brennecke *et al.*, 2005). La régulation combinatoire par différents microARNs ajoute une autre couche de complexité qui peut entraîner une activité plus précise dans divers types de cellules, similaire à la régulation combinatoire par différents facteurs de transcription. Nous pouvons citer quelques fonctions importantes des microARNs chez les animaux.

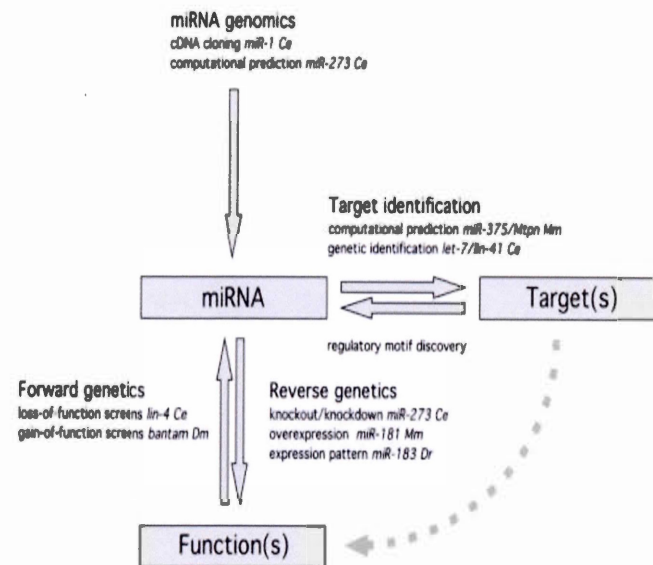


Figure 2.5 Approche d'identification des gènes cibles et des fonctions des microARNs chez les animaux (Wienholds et Plasterk, 2005)

2.5.2.1 Apoptose chez les mouches

Chez la drosophile, deux microARNs bantam et miR-14 ont été génétiquement impliqués dans la régulation de la mort cellulaire programmée (apoptotique). Le bantam a été identifié dans un crible d'insertion d'élément EP(épithélial) pour

les gènes qui affectent la prolifération cellulaire. La sur-expression provoque une prolifération tissulaire et inhibe l'apoptose induite par la prolifération (Brennecke *et al.*, 2003). En revanche, la perte de la fonction du microARN bantam est mortelle. Dans (Brennecke *et al.*, 2003), Bantam régule négativement l'expression de la protéine Hid in vivo et bloque ainsi l'apoptose induite par cette maladie dans l'œil. Le miR-14 a été identifié par un dépistage génétique des inhibiteurs de la mort cellulaire programmée. Les animaux avec des allèles de perte de fonction miR-14 ont une mort cellulaire accrue induite par la protéine Reaper, alors que la sur-expression de miR-14 supprime l'apoptose. L'absence de miR-14 se caractérise également par une semi-sénilité, une durée de vie réduite et une sensibilité au stress.

2.5.2.2 Développement précoce des vertébrés

Plusieurs observations montrent que les microARNs sont très importants pour le développement normal des mammifères. Dans les travaux de (Suh *et al.*, 2004) et (Houbaviy *et al.*, 2003), les cellules ES (Embryonic Stem) de souris et humaines expriment un ensemble spécifique de microARN qui sont régulés négativement lors de la différenciation en corps embryoïdes. Ces mêmes cellules ES déficientes en dicer sont viables mais ne forment pas de microARNs matures et donc ne parviennent pas à se différencier in vitro et in vivo (Kanellopoulou *et al.*, 2005). On constate que dans (Bernstein *et al.*, 2003), les embryons de souris mutantes dicer meurent avant la formation de l'axe au cours de la gastrulation et présentent une perte de cellules ES.

2.5.2.3 Moment du développement chez les vers

La génétique a permis découvrir que les deux microARNs les mieux étudiés, lin-4 et let-7 agissaient dans la voie hétérochronique pour réguler le développement de

C. elegans (Lee *et al.*, 1993). (Reinhart *et al.*, 2000), le Let-7 a été identifié comme un suppresseur des mutants *lin-14*, indiquant qu'ils agissent dans le même chemin. Les mutations de perte de fonction de *lin-4* et de *let-7* provoquent toutes les deux un retard de développement mais à différents stades de développement (Reinhart *et al.*, 2000). Alors que les mutants de *lin-4* "nuls" réitèrent les destins spécifiques du premier stade larvaire aux stades ultérieurs. Les mutants *let-7* "nuls" réitèrent les destins de cellules larvaires au stade adulte. Inversement, la sur-expression de *let-7* donne un développement précoce, le phénotype hétérochronique opposé (Reinhart *et al.*, 2000). La chronologie des phénotypes correspond à l'apparition de l'expression de *lin-4* et de *let-7*, qui est précoce et tardive dans le développement (Lee *et al.*, 1993). Ce qui permet d'en déduire que, *lin-4* est une minuterie de développement précoce et *let-7* une minuterie de type tardif.

2.5.2.4 MicroARNs et cancer

La moitié des gènes de microARNs humains sont localisés sur des sites connus pour être impliqués dans des cancers, tels que des sites fragiles, des régions minimales de perte d'hétérozygotie. Dans l'article de (Calin *et al.*, 2004), de tels emplacements suggèrent que certains microARNs sont impliqués dans la tumorigénèse. Les profils d'expression des microARNs d'un grand nombre d'échantillons de tumeurs humaines montrent que les microARNs sont parfois régulés positivement, mais généralement négativement dans les tumeurs (Lu *et al.*, 2005). En plus, presque tous les microARNs sont différentiellement exprimés dans différents cancers. Comme les profils microARNs reflètent la lignée du développement et l'état de différenciation des tumeurs. Ces profils peuvent être utilisés pour classer les tumeurs peu différenciées (Lu *et al.*, 2005). Les microARNs régulent les voies métaboliques impliquées dans la cancérogenèse en ciblant de nombreux oncogènes et gènes suppresseurs de tumeur. L'augmentation ou la perte d'activité d'un microARN pourrait participer à la croissance tumorale (Kong *et al.*, 2012). Les rôles

de certains microARNs ont été étudiés plus en détail dans le tableau (tableau 2.3).

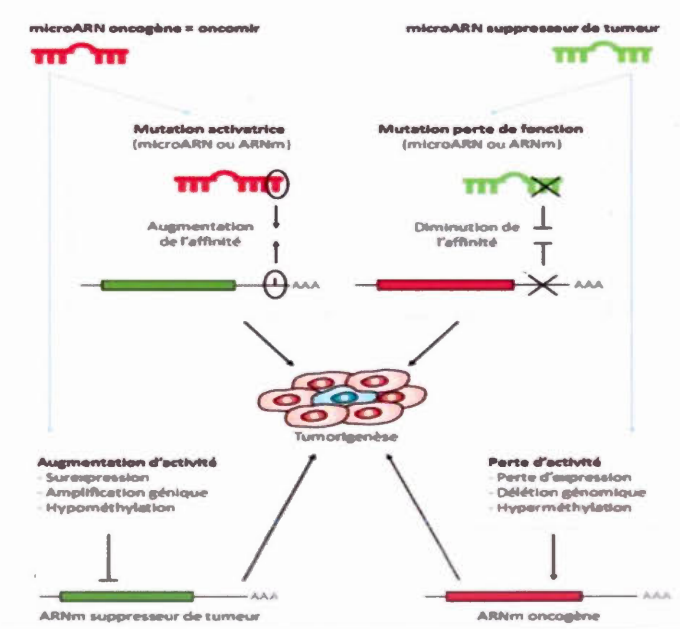


Figure 2.6 Différents niveaux de régulation de la tumorigenèse par les microARNs (Kong *et al.*, 2012)

2.5.3 Bilan des microARNs et leurs fonctions

Ce tableau ci-dessous résume l'ensemble des fonctions contrôlées des microARNs dans chez les plantes et les animaux.

Tableau 2.3 Exemples de microARN dont la fonction biologique et/ou les cibles ARNm sont caractérisées (Cavaillé, 2004)

microARN	organisme	Genes cible	Fonctions contrôlées
lin-4	<i>C. elegans</i>	lin-14, lin-28	Mise en place des stades larvaires
let-7	id	lin-41, lin-57/hbl-1	Mise en place des stades larvaires
lsy-6	id	cog-1	Asymétrie neuronale droite/gauche
Bantam	<i>D. melanogaster</i>	hid	Apoptose, prolifération cellulaire
miR-14	id	?	Apoptose, métabolisme des lipides
miR-2	id	reaper/grim/sickle	Apoptose ?
miR-2	id	reaper/grim/sickle	Apoptose ?
miR-7	id	ε(spl), Brd, hairy	Signalisation Notch ?
miR-277	id	Multiples gènes	Voies de dégradation de la Valine, Leucine et Isoleucine ?
miR-165/-166	<i>A.thaliana</i>	Phabulosa, Phavoluta	Morphogenèse des feuilles
Jaw	id	gènes TCP	Morphogenèse des feuilles
miR-172	id	APATA2	Développement floral
miR-127/-136	Mammifères	Rtl1 (rétrotransposon-like)	Extinction de Rtl ? Empreinte parentale ?
miR-181	id	?	Différenciation hématopoïétique
miR-15/-16/142/-155	id	?	Oncogenèse ?

2.6 Détermination de leur profil d'expression

Dans (van Rooij, 2011), la même approche traditionnelle pour visualiser l'expression des ADNs, le buvardage de type northern a été appliqué aux microARNs. Cette approche a besoin de purifier une fraction de petits ARNs, de la séparer par électrophorèse sur gel d'acrylamide dénaturant et de la transférer sur une membrane. Une sonde radiomarquée est utilisée pour détecter la présence d'un microARN sur la membrane. La sensibilité de cette technique est directement liée à l'affinité entre la sonde et le microARN ciblé. Mais on peut signaler qu'il existe plusieurs types de sondes et de marquages qui ont été développés. Du fait de sa

simplicité et de son coût moins cher, on utilise des sondes d'ADN, mais elles ont une affinité plus faible pour la cible par rapport aux sondes d'ARNs. Les sondes incorporant des LNA (Locked Nucleic Acid) dans leurs séquences sont théoriquement les plus efficaces. Une autre approche a été développée et consiste à incorporer une plus grande quantité d'isotopes radioactifs dans la sonde, produisant un signal plus fort. Comme alternatif, on peut également remplacer la radioactivité par une étiquette fluorescente quelque soit le type d'acide nucléique utilisé. Le buvardage de northern est employé systématiquement pour confirmer les résultats obtenus par séquençage à haut débit. Il faut noter que l'hybridation effectuée peut durer plus longue deux jours et nécessite une quantité considérable d'ARN (10 à 15 micro-gramme) ou de petits ARNs (2 à 5 micro-gramme) (van Rooij, 2011). Il est possible d'appliquer ces principes et d'utiliser une sonde non-radioactive pour effectuer une hybridation in situ. Cette technique permet de visualiser le type cellulaire où est présent le microARN étudié. Une autre approche utilise le PCR (Polymerase Chain Reaction) quantitative en temps réel pour établir un profil d'expression des microARNs. De nombreuses techniques différentes sont utilisées pour la détermination de profil d'expression. Les deux problèmes rencontrés avec cette approche de PCR est la taille des microARNs et la limitation technique qu'elle englobe. La majeure partie des protocoles utilisent une amorce universelle à laquelle est liée le microARN. Cette approche permet d'étendre l'ensemble de la population d'ARN présents dans l'échantillon. Elle permet aussi de fournir une séquence capable de faire la transcription inverse en ADNc. Du fait de la présence d'une séquence commune, la sensibilité et la spécificité de cette approche sont hautement dépendantes du design de l'amorce spécifique au microARN. De plus la forte homologie entre les microARNs d'une même famille, ne variant pas généralement que d'un ou deux nucléotides, rend leur identification plus difficile. Cette approche est cependant plus sensible et requiert moins de matériel biologique que le buvardage de type northern. La spécificité de la liaison avec l'amorce

étant dépendante du contenu en GC du microARN, il est très difficile de comparer l'abondance relative de plusieurs microARNs par le PCR quantitative en temps réel. Donc pour ce faire, le séquençage à haut débit est beaucoup plus efficace (van Rooij, 2011).

2.7 Conclusion

Les systèmes microARNs chez les plantes et les animaux jouent un rôle fondamental dans le développement et semblent exercer principalement leur influence en contrôlant les gènes régulateurs. Cependant, il existe également de nombreuses différences de fonctions entre ces deux systèmes. Chez les animaux, la première étape de la biogenèse des microARNs concerne Drosha, mais ce rôle est joué par DCL1 chez les plantes. Chez les plantes, les microARNs régulent principalement leurs cibles par clivage dans la région codante de l'ARN, tandis les microARNs chez les animaux fonctionnent principalement par répression traductionnelle à l'aide de cibles situées au niveau de la région 3'-UTR. Bien que ces différences entre les systèmes de microARNs soient claires. Une des raisons possibles des différences générales entre les deux systèmes est qu'ils ont évolué séparément, bien que probablement à partir de composants anciens communs. L'abondance des ces microARNs dans les banques de données, a permis aux bioinformaticiens de proposer des modèles paramétriques et non paramétriques basés sur d'analyse d'expression différentielle des gènes.

CHAPITRE III

REVUE DE LA LITTÉRATURE

3.1 Introduction

Dans cette partie d'état d'art, nous verrons les types d'approches d'analyse de l'expression différentielle : les modèles paramétriques, les modèles non-paramétriques et les autres. Ces méthodes sont basées sur des outils spécifiques pour expliquer l'expression différentielle des ARN-seq et ou des microARNs sur différentes types d'espèces. Plusieurs modèles ont été proposés parmi lesquels nous pouvons en citer : DESeq2, EdgeR, NOISeq, Bayseq, NOIseq, etc. Cette liste n'est pas exhaustive, mais nous essayerons de voir les approches les plus utilisées dans la littérature sur l'analyse et l'interprétation de l'expression différentielle. Leur contexte d'utilisation est aussi un facteur important pour connaître le profil d'expression des gènes dans chaque approche.

3.2 Méthodes paramétriques

Les approches paramétriques sauvegardent toutes les informations sur les données contenues dans les paramètres. Il est possible de prédire la valeur de données inconnues en observant le modèle adopté et ses paramètres. Lorsque des méthodes paramétriques sont appliquées à l'expression génique différentielle, supposons que, généralement après une phase de normalisation, chaque valeur d'expression d'un

gène donné est cartographiée dans une distribution particulière connue, telle que Poisson (Marioni *et al.*, 2008), (Hardcastle et Kelly, 2010) ou binomiale négative (Anders et Huber, 2010), (Robinson et Smyth, 2007a).

3.2.1 DESeq2

L'approche implémentée dans DESeq2 commence d'abord par construire un modèle avec les comptages observés. Ensuite, il convient d'utiliser la même méthode depuis le DESeq d'origine ou entre deux étapes. La première étape consiste à rechercher la valeur du paramètre qui rend la probabilité la plus grande, appelée estimation maximale de vraisemblance. La deuxième étape est de prendre toutes les valeurs du gène et déplace ces valeurs vers une valeur moyenne. Le DESeq2 utilise le théorème de Bayes basé sur des probabilités pour déterminer la quantité de mouvement de chaque gène : si l'information relative au gène est faible, sa valeur est proche de la moyenne, si l'information relative au gène est élevée, sa valeur est très peu déplacée. Ainsi, les valeurs déplacées sont utiles pour évaluer différents ensembles de gènes ainsi que pour appliquer un seuil pour déterminer les gènes différentiellement exprimés (Love *et al.*, 2014).

3.2.2 Sleuth

Dans (Bray *et al.*, 2016), la méthode de sleuth commence par filtrer des transcrits de faible abondance. Il s'en suit l'application de deux normalisations successives, puis de l'estimation des paramètres pour le modèle de chaque transcrit. Ce qui entraîne la régularisation de la variance biologique contribuant à la variance de l'abondance du transcrit à travers les échantillons, et enfin à une estimation de la variance totale globale pour chaque transcrit. Sleuth reçoit une sortie de quantification provenant de l'outil kallisto. Ce dernier est un programme de quantification d'ARN-seq deux fois plus rapide que les précédentes et d'une précision similaire.

Le Kallisto pseudoaligns lit une référence en produisant une liste de transcriptions compatibles avec chaque lecture, tout en évitant l'alignement de bases individuelles. kallisto permet d'analyser 30 millions de lectures d'ARN par paires de bases appariées en moins de 10 minutes sur un ordinateur portable standard. Ce qui entraîne une suppression de goulot d'étranglement informatique majeur dans l'analyse d'ARN-seq (Bray *et al.*, 2016).

3.2.3 EdgeR

Il est un modèle de super dispersion suivant une distribution de Poisson à un paramètre. Dans leurs travaux (Robinson *et al.*, 2010), EdgeR est utilisé pour prendre en compte les variations techniques et biologiques que pourraient engendrer les données d'analyse différentielle. On lui applique en même temps la méthode empirique bayésienne pour modérer le degré de dispersion excessive par rapport aux transcrits. Le modèle edgeR estime les dispersions géniques par la méthode du maximum de vraisemblance conditionnelle, en fonction du nombre total de gènes reçus (Smyth et Verbyla, 1996). Une procédure de la méthode de Bayes empirique est utilisée pour réduire les dispersions vers une valeur consensuelle, empruntant ainsi des informations entre gènes (Robinson et Smyth, 2007a). L'expression différentielle est évaluée pour chaque gène en utilisant un test exact similaire au test de Fisher, mais adapté aux données sur-dispersées (Robinson et Smyth, 2007b). Dans le cas des expériences avec répliquions techniques, les données peuvent être bien ajustées par la distribution de Poisson, comme l'a montré (Marioni *et al.*, 2008) dans son article. Puisque la distribution de Poisson est un cas particulier de la distribution de binomiale négative ($\phi = 0$), edgeR peut effectuer une analyse différentielle basée sur la méthode de Poisson.

3.2.4 Limma et Voom

Il fait parti des modèles linéaires, ce modèle a été conçu au début pour faire l'analyse des données d'un micro-réseau mais son extension a permis l'analyse d'ARN-seq (Law *et al.*, 2014). Dans son guide (Smyth *et al.*, 2002), Limma exige l'utilisation de la normalisation TMM du paquetage edgeR associée à l'utilisation de la conversion de voom, qui transforme essentiellement les comptages normalisés en logarithmes à base 2 et estime la relation moyenne-variance pour déterminer le poids de chaque observation faite initialement par un modèle linéaire. Dans cet article de (Law *et al.*, 2014), nous distinguons deux manières d'intégrer la relation moyenne-variance dans l'analyse de l'expression différentielle. La première consiste à modifier la procédure empirique de Bayes dans Limma afin d'intégrer une tendance de variance moyenne. La deuxième méthode incorpore la tendance de la variance-moyenne dans un poids de précision pour chaque observation individuelle normalisée. Une nomenclature a été faite en appelant la première méthode limma-trend et la seconde méthode voom, un acronyme pour «modélisation de la variance au niveau observationnel». Limma-trend applique la relation de variance-moyenne au niveau des gènes alors que voom l'applique au niveau des observations individuelles.

3.2.5 BaySeq

BaySeq utilise l'approche empirique bayésienne pour estimer la probabilité à posteriori de chaque ensemble de modèles, ce qui permet de définir des modèles d'expression différentielle pour chaque t-uplet (Hardcastle et Kelly, 2010). BaySeq estime ensuite la probabilité à posteriori de chaque modèle pour chacun des gènes de l'ensemble de données. Cette approche commence par définir chacun de nos modèles en termes de similarité et de différence entre les échantillons. Pour un modèle donné, nous cherchons à définir quels échantillons se comportent de manière simi-

laire et pour quels ensembles d'échantillons il existe des différences identifiables. Pour évaluer les probabilités à posteriori de chaque modèle pour chaque t-uplet, (Hardcastle et Kelly, 2010) considère une distribution pour le t-uplet définie par un ensemble de paramètres sous-jacents pour lesquels une distribution à priori existe. Les échantillons se comportant de manière similaire doivent posséder la même distribution préalable sur les paramètres sous-jacents du t-uplet, alors que les échantillons se comportant différemment doivent posséder des distributions à priori différentes. Les informations provenant de l'ensemble des gènes sont utilisées pour former une distribution empirique préalable des paramètres du modèle de binomiale négative (Soneson et Delorenzi, 2013). L'avantage important de cette méthode est que l'évaluation de la probabilité à posteriori pour plusieurs modèles est simplement réalisable.

3.2.6 DESeq

Dans (Anders et Huber, 2010), les auteurs proposent une approche basée sur une distribution binomiale négative, avec variance et moyenne liée par régression locale. L'approche basée sur DESeq adopte une approche conservatrice en définissant la dispersion d'un gène comme la plus grande des valeurs obtenues à partir de l'ajustement et de l'estimation de la dispersion individuelle du gène (Soneson et Delorenzi, 2013). La régression locale pour trouver la relation moyenne-variance. Il décrit une approche consistante permettant de sélectionner la plus grande valeur ajustée et l'estimation de la dispersion individuelle pour chaque gène. L'approche de régression locale a également été utilisée dans la transformation stabilisante de la variance fournie par DESeq (notée VST en anglais)

3.3 Méthodes non paramétriques

Les approches non-paramétriques peuvent garder plus de détails sur la distribution des données, autrement dit ne pas imposer de modèle rigide à adapter. Cela est possible parce que les modèles non paramétriques prennent en compte le fait que la distribution de données ne peut pas être définie à partir d'un ensemble fini de paramètres. Ainsi, la quantité d'informations sur les données peut augmenter avec son volume. Nous distinguons dans la revue de littérature deux types modèles non paramétriques. Car ces derniers ne sont pas nombreux pour déterminer l'expression différentielle.

3.3.1 NOIseq

C'est un modèle qui est applicable à toutes sortes de données et est non paramétrique (Tarazona *et al.*, 2015). Il modélise de manière empirique le bruit dans les données de comptage et permet l'analyse des données sans réplication (Tarazona *et al.*, 2011). NOIseq explore la distribution des fold-changes et les différences d'expression absolue entre les conditions contrastées pour les données observées. Il compare cette distribution à la distribution correspondante obtenue en comparant des paires d'échantillons appartenant à la même condition (on parle de «distribution de bruit»). (Soneson et Delorenzi, 2013), dans cette approche calcule, pour chaque gène, une statistique (notée qNOIseq) définie comme la fraction de points de la distribution du bruit correspondant à un fold-change plus faible et à une différence d'expression absolue inférieure à celle du gène d'intérêt dans les données d'origine.

3.3.2 SAMseq

Il est aussi une approche non paramétrique avec ré-échantillonnage pour séquencer des comptages avec des profondeurs différentes (Li et Tibshirani, 2013). Il

peut être appliqué aux données avec des résultats quantitatifs, à deux classes ou à plusieurs classes. SAMseq repose principalement sur une statistique de Wilcoxon, moyennée sur plusieurs ré-échantillonnages des données observées. Il utilise une stratégie de permutation d'échantillon pour estimer le false discovery rate (FDR) pour différentes valeurs de seuil pour cette statistique. Les estimations sont ensuite utilisées pour définir une q-valeur pour chaque gène. Il fait des comparaisons couplées et non appariées à deux groupes, des comparaisons multi-groupes et d'analyse de survie aussi (Soneson et Delorenzi, 2013). La méthode SAMseq donne des taux vrais positifs élevés dans tous les paramètres de simulation, apparemment sans FDR (false discovery rate) élevé. Cependant, pour les plus petites tailles d'échantillons, cette méthode ne trouve pas de gènes significativement et différentiellement exprimés, ce qui n'est pas surprenant en raison de sa nature non paramétrique et de sa dépendance aux permutations des échantillons (Soneson et Delorenzi, 2013).

3.4 Autres méthodes

Dans la revue de la littérature, nous n'avons pas vu une classification de ces méthodes sur les approches paramétriques et les approches non paramétrique. Ces autres types de méthodes sont basées sur la détection de transcription. Elles ont été développées afin d'identifier des transcrits ou des isoformes inconnus et peuvent également être appliquées à l'identification de l'expression différentielle des gènes, telles que EBSeq (Leng *et al.*, 2013) et Cuffdiff2 (Trapnell *et al.*, 2013).

3.4.1 EBSeq

Cette approche a été développée par (Leng *et al.*, 2013) dans le but principal d'identifier des isoformes exprimées de manière différentielle. EBSeq est également robuste dans l'identification des expressions des gènes différentiellement ex-

primés. EBSeq nécessite des comptages de gènes ou des estimations de l'expression des isoformes, mais n'est pas spécifique à une méthode d'estimation particulière comme les autres modèles (paramétriques et les approches non paramétrique). Il ressemble à baySeq, qui applique l'approche empirique bayésienne. EBSeq utilise une approche similaire, mais suppose une forme paramétrique de la distribution à priori des paramètres, avec des hyperparamètres partagés entre tous les gènes et estimés à partir des données (Soneson et Delorenzi, 2013). Le taux de vrais positifs d'EBSeq n'étant pas en grande partie pas affecté par la taille de l'échantillon, mais le FDR diminue au et à mesure que la taille de l'échantillon augmente. Son principal avantage sur les autres approches réside dans sa capacité à identifier les isoformes de l'expression différentielle, les résultats de simulation d'études de cas de (Leng *et al.*, 2013) ont démontré également de bonnes performances pour l'identification des gènes différentiellement exprimés.

3.4.2 Cuffdiff2

Cuffdiff2 estime l'expression à une résolution au niveau de la phase transcriptionnelle. Il contrôle la variabilité évidente des gènes à travers des replicats biologiques. Cuffdiff 2 identifie aussi de manière robuste les transcrits et les gènes exprimés de manière différentielle et révèle les changements d'épissage différentiel et de préférence des promoteurs (Trapnell *et al.*, 2013). Il suppose que l'expression d'une transcription dans chaque condition peut être mesurée en comptant le nombre de fragments générés par celle-ci. Dans l'article de (Trapnell *et al.*, 2013), une mutation dans le niveau d'expression d'un transcrit est mesuré en comparant le nombre de fragments dans chaque condition. Si la probabilité de voir un changement dans ce compte est suffisamment petite avec un modèle statistique approprié de la variabilité inhérente à ce comptage, la transcription est réputée de manière significativement et différentielle exprimée. Comme le modèle suit une distribution de Poisson, la variabilité est estimée en calculant le nombre moyen de répétitions, ce

qui permet de calculer une p-value pour tout changement observé dans le nombre de fragments d'un transcrit.

3.4.3 Approche intégrée

Une approche intégrée pour identifier le microARNome de blé hexaploïde associé au développement et à la tolérance au stress abiotique a été développée par (Agharbaoui *et al.*, 2015). Les approches actuellement utilisées pour identifier les microARN dans ce système complexe polyploïde se concentrent sur les microARN conservés et hautement exprimés, en évitant régulièrement ceux qui sont souvent spécifiques à une lignée, à une condition ou qui sont apparus récemment en évolution. En outre, de nombreux facteurs environnementaux et biologiques affectant l'expression des microARN n'étaient pas encore pris en compte, ce qui aboutissait à un répertoire incomplet de microARN du blé (Agharbaoui *et al.*, 2015). Les auteurs ont développé une technique indépendante de la conservation basée sur une approche intégrée combinant un apprentissage automatique, des outils bio-informatiques, des informations biologiques sur les profils d'expression de microARN connus et des critères universels de microARN de plantes pour identifier les microARN avec plus de confiance (Agharbaoui *et al.*, 2015). Leur pipeline développé peut potentiellement identifier de nouveaux microARN de blé qui partagent des caractéristiques communes à plusieurs espèces ou qui sont spécifiques à une espèce ou à un clade.

3.5 Observations sur les méthodes

Parmi ces approches d'expression différentielle, nous pouvons faire un résumé des principales observations et des caractéristiques pour chacune des méthodes évaluées dans le tableau ci-dessous.

Tableau 3.1 Résumé 1 des principales observations des méthodes (Wienholds et Plasterk, 2005)

Nom	Observations
DESeq	- Conservative with default settings. Becomes more conservative when outliers are introduced - Generally low TPR - Poor FDR control with 2 samples/-condition, good FDR control for larger sample sizes, also with outliers - Medium computational time requirement, increases slightly with sample size
edgeR	- Slightly liberal for small sample sizes with default settings. Becomes more liberal when outliers are introduced - Generally high TPR- Poor FDR control in many cases, worse with outliers - Medium computational time requirement, largely independent of sample size
Lima et voom	- Good type I error control, becomes more conservative when outliers are introduced - Low power for small sample sizes. Medium TPR for larger sample sizes. - Good FDR control. Largely unaffected by introduction of outliers - Computationally fast

Tableau 3.2 Résumé 2 des principales observations des méthodes (Wienholds et Plasterk, 2005)

Nom	Observations
baySeq	- Highly variable results when all DE genes are regulated in the same direction. Less variability when the DE genes are regulated in different directions. - Low TPR. Largely unaffected by outliers - Poor FDR control with 2 samples/condition, good for larger sample sizes in the absence of outliers. Poor FDR control in the presence of outliers. - Computationally slow, but allows parallelization
EBSeq	- TPR relatively independent of sample size and presence of outliers.- Poor FDR control in most situations, relatively unaffected by outliers - Medium computational time requirement, increases slightly with sample size.
NOIseq	- Not clear how to set the threshold for qNOI-Seq to correspond to a given FDR threshold - Performs well, in terms of false discovery curves, when the dispersion is different between the conditions (see supplementary material). - Computational time requirement highly dependent on sample size
SAMseq	- Low power for small sample sizes. High TPR for large enough sample sizes - Performs well also for simulation study - Largely unaffected by introduction of outliers - Computational time requirement highly dependent on sample size

3.6 Type de normalisation des approches

Ce tableau ci-dessous extrait de l'article de (Costa-Silva *et al.*, 2017) montre le type de normalisation utilisée pour déterminer l'expression différentielle des gènes.

Tableau 3.3 Méthodes et normalisation d'expression différentielle des RNA-Seq
(Costa-Silva *et al.*, 2017)

Name	Version	Normalisation
baySeq	2.4.1	Scaling factors (quantile/ TMM/ total)
DESeq	1.22.1	DESeq size factors
EBSseq	1.12.0	DESeq median normalization
edgeR	3.12.1	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)
limma + voom	3.26.9	TMM
NOIseq	2.14.1	RPKM / TMM / Upper quartile
SAMseq (samr)	2.0	Based on the read count mean over the null features of data set
DESeq2	1.10.1	DESeq size factors
sleuth	0.29.0	DESeq size factors (with slight modifications)

3.7 Conclusion

Dans ce chapitre, nous avons présenté une revue approfondie des méthodes d'identification des gènes exprimés de manière différentielle ou des transcrits. Parmi les méthodes évaluées, nous ne pouvons pas affirmer avec certitude qu'un modèle est meilleur par rapport à un autre. Les tableaux **3.1** et **3.2** résument les principales constatations et observations des approches. Aucune méthode évaluée n'est optimale dans toutes les circonstances, et par conséquent, la méthode de choix dans une situation particulière dépend des conditions expérimentales et des types données. Toutes ces approches susmentionnées sont basées sur une distribution binomiale négative ou de poisson. Partant de ces différents types de normalisation (tableau **3.3**), nous introduirons les deux autres étapes fondamentales (estimation des dispersions et les test d'analyse de l'expression différentielle) de l'expression différentielle dans le chapitre suivant.

CHAPITRE IV

PROPOSITION D'UNE APPROCHE D'EXPRESSION DIFFÉRENTIELLE

4.1 Introduction

De nombreux packages R pour la détection de gènes exprimés de manière différentielle ont été développés à partir de données biologiques. Le processus d'analyse comporte trois parties essentielles, la normalisation, l'estimation de la dispersion et le test d'expression différentielle. Les méthodes de normalisation et les modèles statistiques ne sont pas échangeables entre les packages sans transformations adéquates dont les utilisateurs ne connaissent pas souvent. L'un des points majeurs de l'analyse des données de microARNs ou ARN-seq est la détection des gènes différentiellement exprimés. La détection des gènes différentiellement exprimés met souvent en jeu des tests statistiques classiques. Cependant, le volume des données génomiques peut poser problème en termes de temps de calcul, de fiabilité statistique des résultats ainsi que de leur lisibilité. Des adaptations des outils comme DESeq2 et tant d'autres sont par conséquent mises en œuvre afin de pallier à ces problèmes. L'approche DESeq2 fournit des méthodes pour tester l'analyse d'expression différentielle en utilisant des modèles linéaires généralisés en suivant une distribution binomiale négative. Les estimations des dispersions et les LFC (Logarithmic Fold Change) intègrent des distributions à priori basées sur des données. Cette approche fournit aussi des informations sur les nombres d'échantillon et de lecture nécessaires pour obtenir des résultats significatifs sur le plan biologique.

4.2 Modèle du pipeline DESeq2

L'analyse d'expression différentielle dans DESeq2 débute par une matrice de comptage K avec une ligne pour chaque gène i et une colonne pour chaque échantillon j . Les paramètres d'entrées de la matrice K_{ij} montrent le nombre de lectures de séquençage qui ont été cartographiées sans ambiguïté sur un gène dans un échantillon. La modélisation des comptages K_{ij} suit une distribution binomiale négative (appelée souvent une distribution gamma-Poisson) avec une moyenne ajustée μ_{ij} et un paramètre de dispersion spécifique au gène α_i (Anders et Huber, 2010), (Love *et al.*, 2014) .

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i) \quad (4.1)$$

Avec la moyenne μ_{ij} qui s'écrit sous sa forme générale comme :

$$\mu_{ij} = s_{ij}q_{ij} \quad (4.2)$$

La moyenne prend comme une quantité q_{ij} , proportionnelle à la concentration de fragments d'ADNc du gène dans l'échantillon, mise à l'échelle par un facteur de normalisation s_{ij} dans (4.2). Dans plusieurs applications, la même constante s_j peut être utilisée pour tous les gènes d'un échantillon, ce qui explique les différences de profondeur de séquençage entre les échantillons (Love *et al.*, 2014).

Nous utilisons des GLM sous forme logarithmique pour déterminer la concentration réelle attendue de fragments pour l'échantillon j :

$$\log_2(q_{ij}) = \sum_r x_{jr}\beta_{ir} \quad (4.3)$$

Dans l'équation (4.3), les coefficients β_{ir} donnent les log2 fold change pour le gène i pour chaque colonne r de la matrice du modèle et x_{jr} sont les éléments de la matrice

Le paramètre de dispersion α_i définit la relation entre la variance du comptage observé et sa valeur moyenne. En d'autres termes, le comptage observé par rapport

à la valeur moyenne, qui dépend à la fois du facteur de taille s_{ij} et de la partie dépendante de la covariable q_{ij} définie ci-dessus.

$$\text{Var}(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2 \quad (4.4)$$

En pratique, nous ne connaissons pas les paramètres de la moyenne et de la variance et nous devons les estimer à partir des données. Comme, le nombre de répétitions est faible et il est nécessaire de formuler d'autres hypothèses de modélisation afin d'obtenir des estimations utiles (Anders et Huber, 2010).

4.3 Implémentation

Un modèle de script R (**section Appendices A**) est associé au package DESeq2 de R et permet d'exécuter le pipeline afin de faire les tests de gènes différentiellement exprimés. Les étapes successives de l'analyse d'expression différentielle et les fonctions DESeq2 associées sont présentées dans la figure ci-dessous du workflow. Nos implémentations ont été faites via la plateforme **Rstudio** avec la version **1.1.383**.

4.3.1 Principaux objectifs de notre approche

Les trois objectifs d'apprentissage fixés de notre approche sont les suivants :

1. Comprendre les différentes étapes d'une analyse d'expression différentielle dans le contexte de DESeq2
2. Créer des tableaux de résultats pour la comparaison des différentes variables
3. Résumer les microARNs significativement exprimés de manière différentielle pour chaque comparaison

4.3.2 Structure du Workflow DESeq2

La première étape de toute analyse consiste à importer les données dans un format prêt à l'analyse. Ce dernier dépend des exigences du package utilisé pour l'analyse. De nombreux packages R existent pour la détection de gènes exprimés de manière différentielle à partir de données d'ARN-Seq. Le processus d'analyse comprend trois étapes principales, à savoir la normalisation, l'estimation de la dispersion et le test d'expression différentielle. Nous présentons DESeq2, une méthode d'analyse différentielle des données de comptage, utilisant l'estimation du retrait pour les dispersions et les folds changes afin d'améliorer la stabilité et l'interprétabilité des estimations. Cela permet une analyse plus quantitative axée sur la force plutôt que sur une simple présence d'une expression différentielle. Le package DESeq2 est disponible sur Bioconductor. Le DESeq2 fournit des méthodes pour tester l'expression différentielle en utilisant des modèles linéaires généralisés. Il suit une loi binomiale négative qui est une distribution de probabilité discrète. DESeq2 intègre des des fonctions et des paramètres à estimer importants dans son fonctionnement pour faciliter une analyse plus quantitative des données comparatives de microARNs. Cette distribution binomiale négative basée sur DESeq2 effectue une analyse par défaut à travers les étapes suivantes :

- estimation des facteurs de taille : **estimationSizeFactors**
- estimation de la dispersion : **estimationDispersions**
- Ajustement binomial négatif et statistique de Wald : **nbinomWaldTest**

Les estimations de la dispersion et les tests statistiques incorporent des distributions a priori pilotées par les données. Les étapes du workflow permettant de décrire notre approche est :

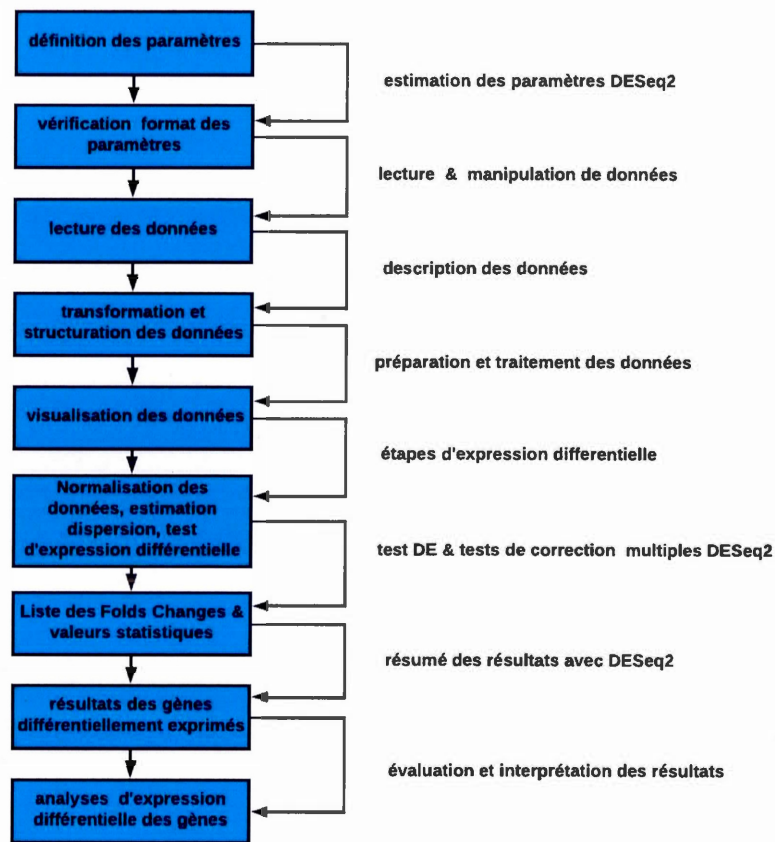


Figure 4.1 Adaptation d'un Workflow d'expression différentielle basé sur DESeq2

Les étapes réalisées de ce workflow (**figure 4.1**) sont la normalisation des données qui est une phase importante dans le processus. Les estimations des dispersions aussi servent à calculer les valeurs des gènes dispersés. La phase finale est le test de l'expression différentielle permettant de voir les gènes exprimés et non exprimés.

— Normalisation des données, estimation des dispersion et test d'expression différentielle

Cette étape comporte trois parties : La normalisation permet de standardiser le jeu de données afin d'obtenir une distribution normale ou gaussienne. Les valeurs aberrantes, les valeurs manquantes et les outliers dans la base de

données. Ce nettoyage permet de cette partie de normalisation sert aussi à rendre les résultats explicites. Dans cette étape, il est impossible de fournir à notre workflow une matrice de comptage RPKM (arrondis) en entrée ou toute matrice de valeurs déjà normalisées, car certains modèles statistiques n'acceptent que les comptages bruts en entrée. Les deux dernières parties restantes : les estimations des dispersions et les tests d'expression permettent de visualiser et d'obtenir les microARNs qui sont sur exprimés ou sous exprimés dans notre workflow. Voici un exemple explicatif de la fonction R-cran pour la normalisation. la fonction **DESeq()** avec le paramètre **normalized** réalisent la normalisation et l'analyse diff en une seule étape. La normalisation des données de comptage se fait par la méthode "effective library size". Pour l'estimation des dispersions, nous illustrons cette étape pour ce code ci-dessous : La fonction **dispersions()** de R-cran permet d'effectuer une estimation de la dispersion des gènes et Valeurs de dispersion. Pour ce qui est de la partie du test d'expression différentielle, nous avons utilisée la fonction **p.adjust** qui calcul des p-values ajustées selon la méthode de bonferroni et de la procédure de Benjamini-Hochberg(BH).

— Liste des folds changes et valeurs statistiques

Cette étape permet de déterminer les gènes régulés de haut en bas à partir des valeurs de log2.

— Résultats des gènes différentiellement exprimés

Des tableaux avec plusieurs paramètres statistiques ont été faits dans cette partie. Les résultats sont statistiquement significatifs pour différencier les microARNs exprimées. les gènes différentiellement exprimés au seuil de 10% après ajustement des tests multiples. La fonction permettant de réaliser cette étape est **plotMA** de DESeq2.

— Analyse de l'expression différentielle des gènes

La dernière étape du workflow consiste à évaluer et identifier tous les microARNs différenciés ou non. l'interprétation des résultats finaux complètent notre workflow. La fonction R-cran appelée **results** contient les résultats de l'analyse différentielle. La fonction permet aussi de sélectionner les gènes déclarés différentiellement exprimés dans une table ordonnée par p-value.

4.4 Données de simulation

Les données proviennent d'une base de données de 220 lignes et 14 colonnes. Les valeurs de la matrice de notre base de données sont des ratios des variables entre quatre catégories de microARNs du blé. Les types de microARNs du blé sont : **Manitou**, **Winter**, **Isopring** et **Norstar**. Ces microARNs ont été créés sur des périodes temporelles *T1* et *T5* pour les années 2010 et 2013. Dans la figure ci-dessous (**figure 4.2**), nous montrons un aperçu sur la base de données des microARNs du blé avec les variables en colonne et les observations en ligne. Chaque type de variable de la table possède un temps *T1* et *T5* afin de mesurer la variabilité différentielle avec les autres variables. On note également dans (**figure 5.1**), deux années 2010 et 2013 pour étudier le profil d'expression des microARNs du blé. Cette étude porte sur des données réelles de microARNs, on note des réplicats dans le tableau des données ci-dessous. Il existe des réplicats biologiques, car on effectue des mesures sur des échantillons différents. Ces types de réplicats ont pour but de limiter la variation due au processus d'échantillonnage au niveau de la base de données. Ces variables de cette base de données sont continues et les valeurs sont normalisées dans un intervalle de $[0,1]$. Les ratios de l'expression des microARNs sont obtenus dans les différents cultivars en fonctions du temps.

Mirnas	Cluster	Man_y2013_T1/Man_y2010_T1	Man_y2013_T5/Man_y2010_T5	I
TATATTATCAGGTTTATGGTCC	filtClusters3323	0,682017371	0,725694198	
TGATAATCCGGCAGGAAAGTAC	filtClusters1119	0,5412384	0,705539908	
TTTGGAAATGATAATCCGGCAC	filtClusters17877	0,905680376	0,873520745	
TTTCGTGCGGATTATCATTTTC	filtClusters20153	0,647880319	0,820827965	
GAATTCTACGGTGGAAAGCTC	filtClusters12173	0,773214347	0,547710277	
GAATTCTACGGTGGAAAGCTC	filtClusters1207	0,795765556	0,689363597	
TTTCCACCGTAGAATTCTGTGT	filtClusters14276	0,934182884	0,760842019	
ACCGTAGAATTCTGTGTGCCAAA	filtClusters1684	0,549510587	0,676033185	
AGACATATACAGTTTGGGTGTT	filtClusters12176	0,607985077	0,583686268	
ATTCAAGTCAGACATATACAGT	filtClusters3988	0,711933433	0,853024333	
ACACCCAAACTGTATATGTCTG	filtClusters3129	0,975795589	0,660383122	
ATGATTATAACCCCAACTGT	filtClusters18741	0,895291877	0,734373404	
ACCCAAACTGTATATGTCTGAC	filtClusters189	0,742020729	0,721424999	
TATCGGTGGACAAATATGT	filtClusters12885	0,880823426	0,717890842	
CATATTATTCACITGGACCTA	filtClusters4692	0,747944132	0,829211158	
CATATTGTTCACITGGACCTA	filtClusters12892	0,731106494	1,002860606	
TTATTCACTTGGACCTAAACCC	filtClusters2158	0,842156249	0,772605468	
TTTGAGAGACCAAGTCAGCT	filtClusters20703	0,77649907	0,868355167	
CCGGTCACATTCAGTCAGAC	filtClusters1677	0,683968844	0,807981221	

Figure 4.2 Visualisation de la base de données : On observe dans cette base de données contient 15 colonnes dont les deux premières (Mirnas et Cluster) sont des variables qualitatives. Les 13 autres colonnes sont des variables quantitatives continues, normalisées sur un intervalle de $[0,1]$ en fonction du temps

Cependant, des prélèvements ont été faites sur des microARNs du blé, portant sur ces réponses au froid sont principalement menées dans des environnements contrôlés qui ne prennent en compte que les réponses à des variables environnementales uniques. (Li *et al.*, 2018).

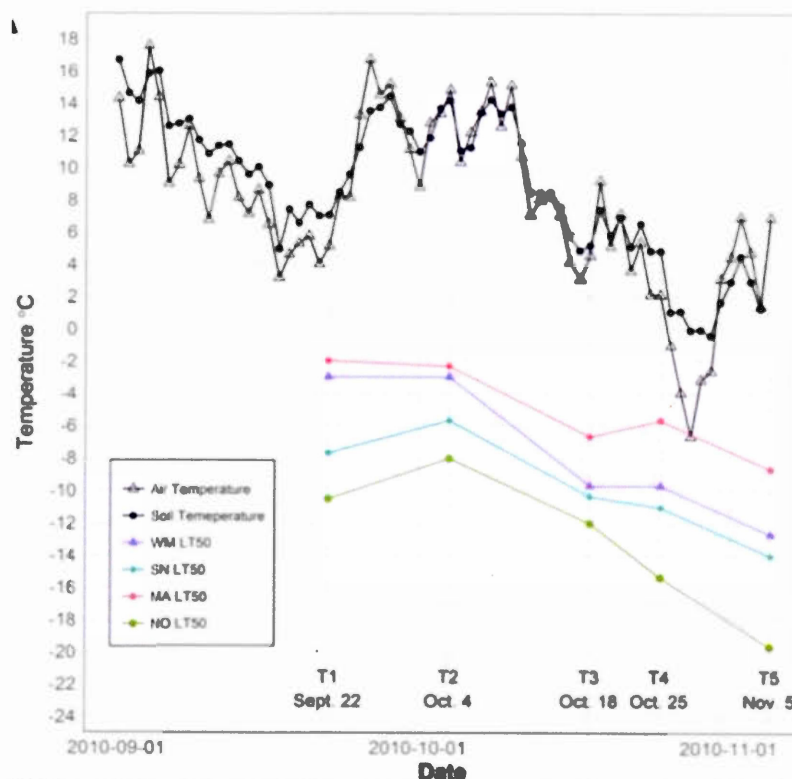


Figure 4.3 Illustration des périodes de prélèvement des microARNs durant la période allant de septembre à Novembre. La température de l'air et du sol sont donnés à titre indicatif en fonction des temps de prélèvement. (Li *et al.*, 2018)

4.5 Conclusion

L'approche DESeq2 offre une solution complète et générale pour l'analyse au niveau des gènes des données RNA-seq. Cette méthode possède des performances constantes sur des types de données et est applicable à de petites études sur des microARNs ou des gènes en particulier. Une avancée critique sur les estimateurs et les tests pour l'analyse d'expression différentielle sont bien décrits dans cette approche. Les estimateurs de DESeq2 fournissent des solutions solides et statisti-

quement bien fondées aux problèmes d'expression dans les études de microARNs. Cette approche tient en compte de plusieurs modèles d'ajustement (**Bonferroni** et **Benjamini-Hocheberg**) pour faire les multiples tests de significativité statistique sur les données. L'exécution de notre workflow (**figure 4.1**) permet de distinguer les trois principales étapes de notre approche proposée pour étudier les microARNs du blé. Les résultats découlant de cette analyse d'expression différentielle seront traités dans le chapitre suivant de présentation des résultats. Ces derniers obtenus identifient tout le processus d'une d'analyse d'expression différentielle de notre proposition.

CHAPITRE V

PRÉSENTATION DES RÉSULTATS

5.1 Introduction

Ce chapitre est consacré uniquement sur la présentation des résultats de notre approche d'expression différentielle basée sur DESeq2. Les simulations ont permis d'obtenir des résultats importants sur les trois étapes d'identification d'expression des microARNs de quatre catégories de blé. Les résultats de la normalisation des données, des estimations des dispersions et des tests d'expression différentielle sont les plus importants dans cette partie. Nous avons réalisé plus des tests comparatifs entre ces types de microARNs du blé pour évaluer notre approche d'expression différentielle proposée dans le **chapitre 4**. Les tests d'analyse d'expression obtenus permettent de évaluer et interpréter les microARNs différentiellement exprimés.

5.2 Résultats de simulation

Pour alléger le document, nous avons agrégé nos résultats afin de sortir ceux qui montrent les variabilités de l'expression différentielle les plus importantes entre les périodicités aux temps **T1** ou **T5** et **T1** et **T5**. Car nous avons vu que les résultats découlants de ces ratios des variables permettent de mieux analyser les microARNs du blé qui sont différentiellement exprimés aux années **2010** et **2013** entre Manitou, Norstar, Isopring et Winter. Pour les étapes de nos résultats pré-

sentés (normalisation, estimation des dispersions, et tests statistiques d'expression différentielle). Vu le nombre important de résultats, nous avons pris les rations des variables les plus pertinentes suivantes :

Années 2010 et 2013 au temps T1

- $\text{Nor-y}_{2010\text{-T1}} \cdot \text{Man-y}_{2010\text{-T1}} - \text{X8S-y}_{2013\text{-T1}} \cdot \text{11W-y}_{2013\text{-T1}}$
- $\text{X8S-y}_{2010\text{-T1}} \cdot \text{11W-y}_{2010\text{-T1}} - \text{Nor-y}_{2010\text{-T1}} \cdot \text{Man-y}_{2010\text{-T1}}$
- $\text{X8S-y}_{2010\text{-T1}} \cdot \text{11W-y}_{2010\text{-T1}} - \text{Nor-y}_{2013\text{-T1}} \cdot \text{Man-y}_{2013\text{-T1}}$
- $\text{X8S-y}_{2013\text{-T1}} \cdot \text{11W-y}_{2013\text{-T1}} - \text{Nor-y}_{2013\text{-T1}} \cdot \text{Man-y}_{2013\text{-T1}}$

Années 2010 et 2013 au temps T5

- $\text{X8S-y}_{2010\text{-T5}} \cdot \text{11W-y}_{2010\text{-T5}} - \text{Nor-y}_{2013\text{-T5}} \cdot \text{Man-y}_{2013\text{-T5}}$
- $\text{X8S-y}_{2013\text{-T5}} \cdot \text{11W-y}_{2013\text{-T5}} - \text{Nor-y}_{2013\text{-T5}} \cdot \text{Man-y}_{2013\text{-T5}}$

Années 2010 et 2013 au temps T1 et T5

- $\text{Nor-y}_{2010\text{-T1}} \cdot \text{Man-y}_{2010\text{-T1}} - \text{X8S-y}_{2013\text{-T5}} \cdot \text{11W-y}_{2013\text{-T5}}$
- $\text{X8S-y}_{2010\text{-T1}} \cdot \text{11W-y}_{2010\text{-T1}} - \text{Nor-y}_{2013\text{-T5}} \cdot \text{Man-y}_{2013\text{-T5}}$
- $\text{X8S-y}_{2010\text{-T5}} \cdot \text{11W-y}_{2010\text{-T5}} - \text{Nor-y}_{2010\text{-T1}} \cdot \text{Man-y}_{2010\text{-T1}}$
- $\text{X8S-y}_{2010\text{-T5}} \cdot \text{11W-y}_{2010\text{-T5}} - \text{Nor-y}_{2013\text{-T1}} \cdot \text{Man-y}_{2013\text{-T1}}$
- $\text{X8S-y}_{2013\text{-T1}} \cdot \text{11W-y}_{2013\text{-T1}} - \text{Nor-y}_{2013\text{-T5}} \cdot \text{Man-y}_{2013\text{-T5}}$
- $\text{X8S-y}_{2013\text{-T5}} \cdot \text{11W-y}_{2013\text{-T5}} - \text{Nor-y}_{2013\text{-T1}} \cdot \text{Man-y}_{2013\text{-T1}}$

Remarque

Pour homogénéiser nos simulations, nous avons utilisé ces mêmes variables ci-dessus tout au long de la présentation des différents résultats.

5.2.1 Résultats de normalisation

Pour le temps T1, les histogrammes de normalisation (**figure 5.2**), nous permettent de voir les normalisations des quatre catégories de blé au temps T1

ci-dessus. Dans cette figure, on voit une similarité entre les différentes normalisations des variables (X8S-y2010-T1..11W-y2010-T1, Nor-y2010-T1..Man-y2010-T1 et X8S-y2013-T1..11W-y2013-T1, Nor-y2013-T1..Man-y2013-T1) aux années 2010 et 2013 pendant la période T1. Pour l'histogramme (Nor-y2010-T1..Man-y2010-T1 et X8S-y2013-T5..11W-y2013-T5), on note une légère augmentation de la distribution des données normalisées (**figure 5.1**). Dans la base de données des microARNs du blé étudiée dans ce rapport, les valeurs des variables sont quantitatives continues. Les valeurs sont ensuite normalisées dans un intervalle [0,1]. En plus les types de microARNs ont été cultivés dans différentes conditions en fonction du temps T1 et T2. Ces histogrammes obtenus servent à s'approcher d'une distribution normale pour mieux réduire les biais dans les données **figure (5.1,..., 5.3)**.

Au Temps T5, nous avons des histogrammes normalisés beaucoup plus hétérogènes. Dans la figure (**figure 5.3**), on observe une grande différence au temps T5 entre les variables X8S-y2010-T5..11W-y2010-T5 et Nor-y2013-T5..Man-y2013-T5 d'une part et d'autre part entre X8S-y2013-T5..11W-y2013-T5 et Nor-y2013-T5..Man-y2013-T5. Les normalisations observées au temps T5 sont très différentes par rapport à ceux obtenus au temps T1.

En prenant les histogrammes normalisés aux temps T1 et T5 pour les années 2010 et 2013, nous constatons que les données normalisées sont différentes d'une variable à une autre (**figure 5.4**). Chaque histogramme à une forme particulière dans les six cas de figures. Cette technique de normalisation aux différents temps T1 ou T5 et aux temps T1-T5, sert à nettoyer les données afin qu'on puisse les rendre exploitables. Cette normalisation permet aussi de comparer les quatre variétés de blé pour pouvoir faire l'analyse d'expression différentielle. Cette phase de normalisation permet d'identifier et de corriger les biais techniques de telle sorte que les différences d'expression ne soient dues qu'à des effets biologiques.

— Normalisation années 2010 et 2013 au temps T1

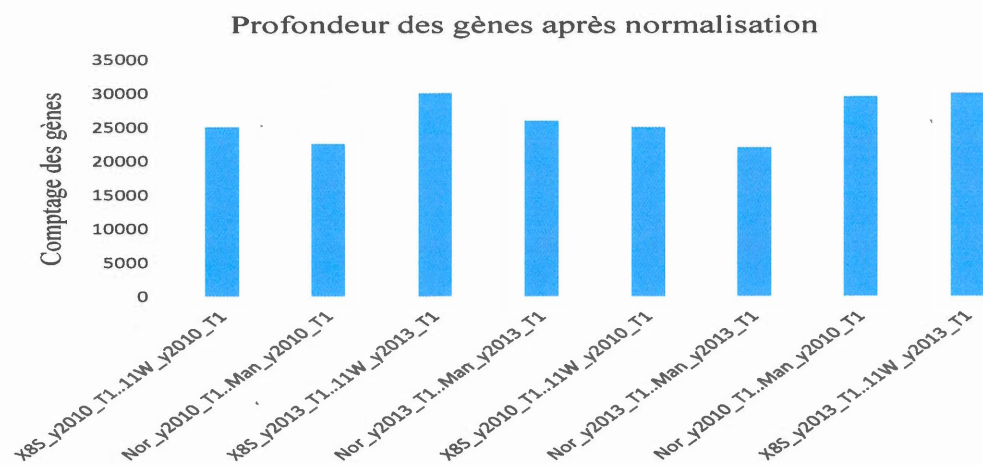


Figure 5.1 Histogramme normalisé au temps T1

— Normalisation années 2010 et 2013 au temps T5

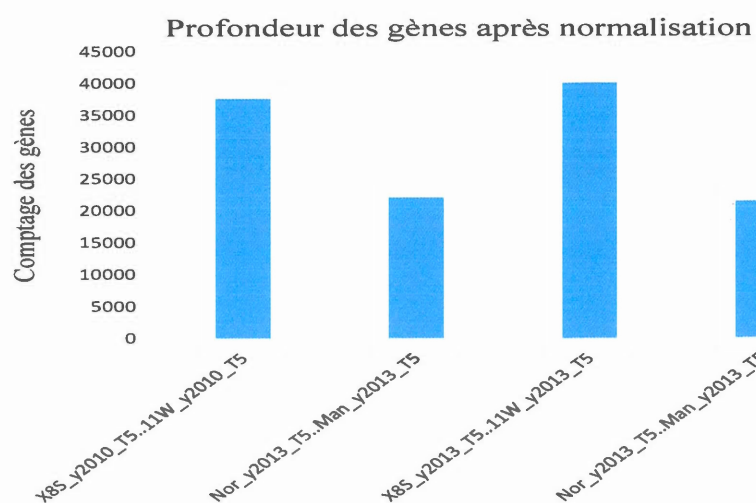


Figure 5.2 Histogramme normalisé au temps T5

— Normalisation années 2010 et 2013 aux temps T1-T5

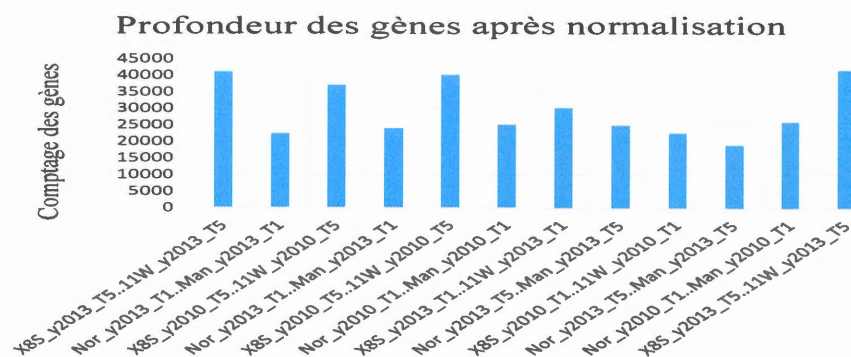


Figure 5.3 Histogramme normalisé aux temps T1 et T5

5.2.2 Résultats d'estimation des dispersions

En se basant sur les mêmes variables ci-dessus aux années 2010 et 2013, nous pouvons constater que les dispersions des microARNs sont différentes au temps T1 ou T5 et T1-T5. Si nous prenons les estimations des dispersions au temps T1 (figure 5.5 ... figure 5.8), nous voyons que les dispersions sont différentes du fait que de la variabilité notée entre les variables du blé. Les droites d'ajustement obtenus diffèrent aussi entre ces différentes figures au temps T1. Cet ajustement permet de centrer les valeurs finales vers les vraies valeurs.

Au temps T5, on voit des estimations très peu similaires entre les variables utilisées (figure 5.9 et figure 5.10). Par compte dans la figure 5.10 on observe plus de valeurs aberrantes. Les estimations des dispersions sont plus dispersées par rapport à la figure 5.9. Pour les estimations des dispersions au temps T1 et T5 (figures 5.11 ... figure 5.16), on constate que les dispersions sont différentes entre les six figures obtenues. Les valeurs aberrantes sont très peu observées dans les données

de microARNs. Les estimations finales des microARNs sont centrées autour de la droite d'ajustement de chaque figure. Nous notons aussi que dans ces résultats de dispersions les droites d'ajustement ont des formes particulières dans nos résultats au temps T1-T5.

— Estimation des dispersions années 2010 et 2013 au temps T1

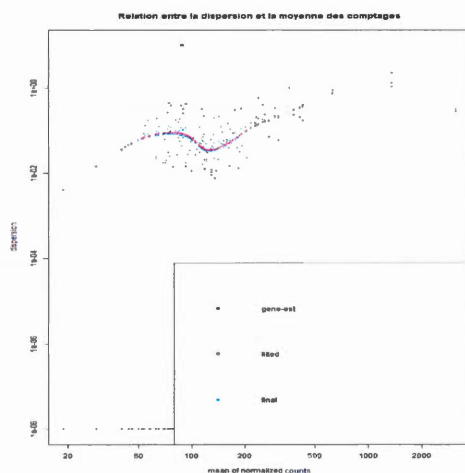


Figure 5.4 Estimation des dispersions au temps T1

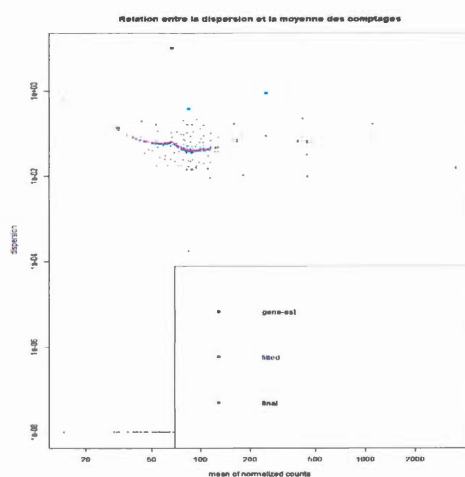


Figure 5.5 Estimation des dispersions au temps T1

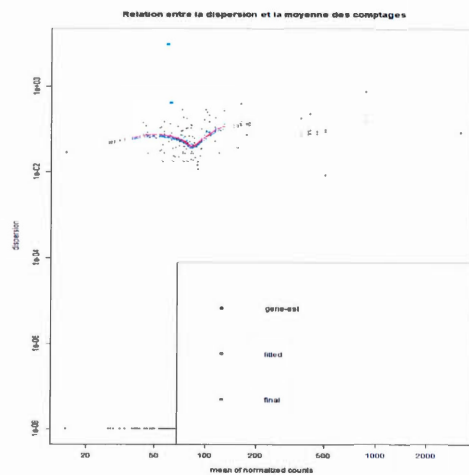


Figure 5.6 Estimation des dispersions au temps T1

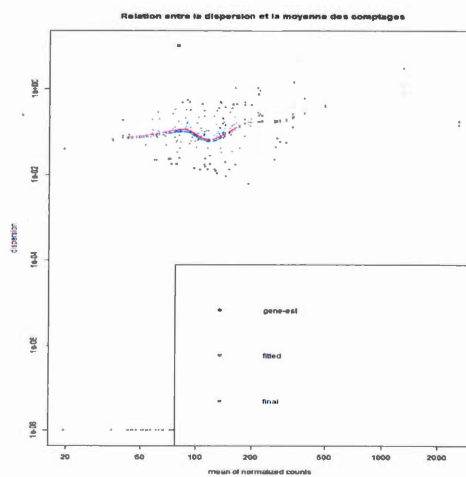


Figure 5.7 Estimation des dispersions au temps T1

— Estimation des dispersions années 2010 et 2013 au temps T5

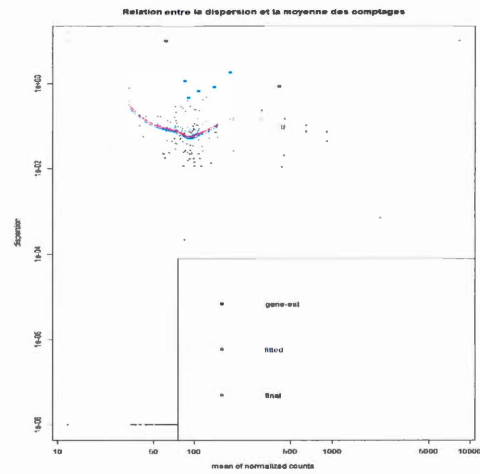


Figure 5.8 Estimation des dispersions au temps T5

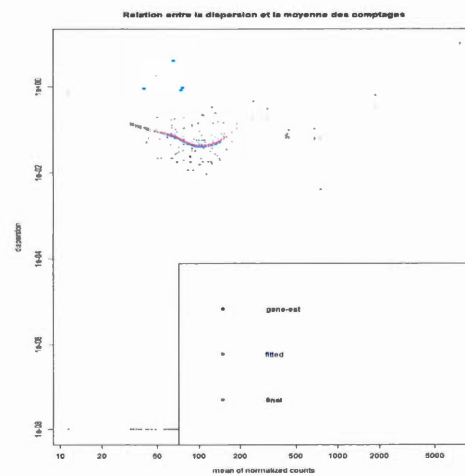


Figure 5.9 Estimation des dispersions au temps T5

— Estimation des dispersions années 2010 et 2013 aux temps T1 et T5

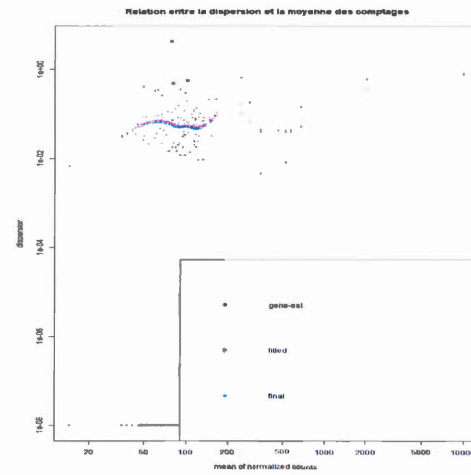


Figure 5.10 Estimation des dispersions aux temps T1 et T5

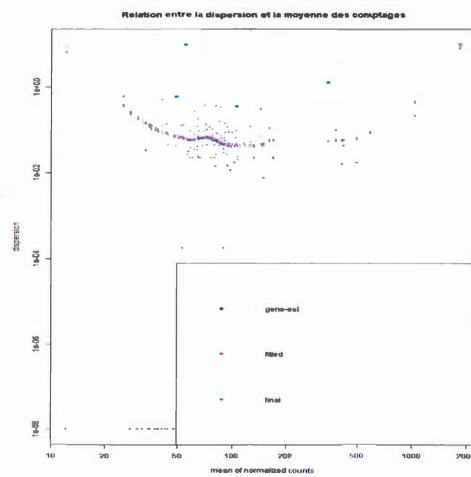


Figure 5.11 Estimation des dispersions aux temps T1 et T5

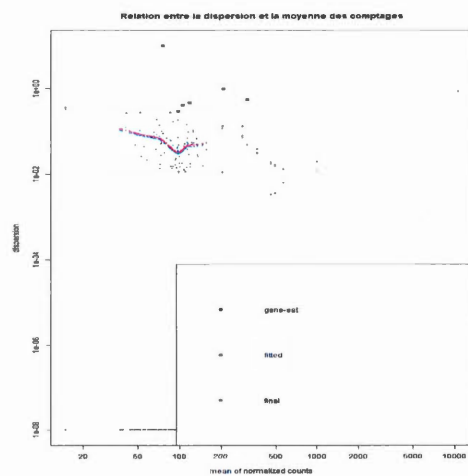


Figure 5.12 Estimation des dispersions aux temps T1 et T5

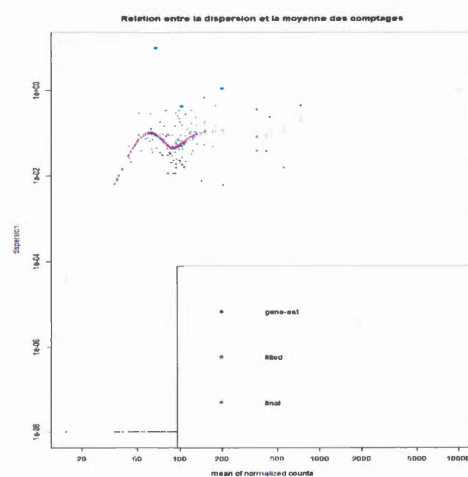


Figure 5.13 Estimation des dispersions aux temps T1 et T5

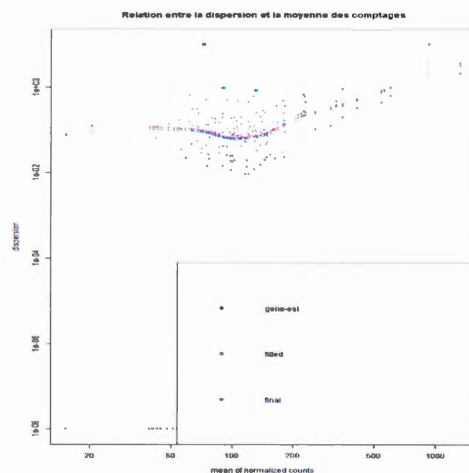


Figure 5.14 Estimation des dispersions aux temps T1 et T5

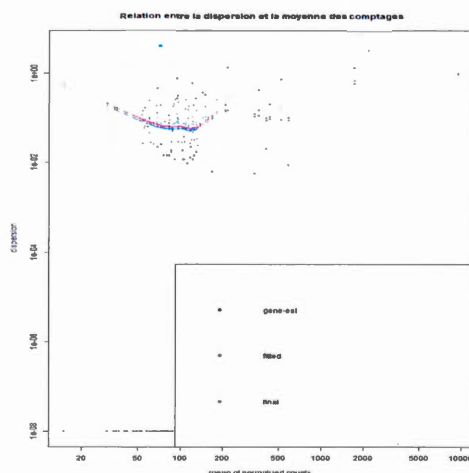


Figure 5.15 Estimation des dispersions aux temps T1 et T5

5.2.3 Résultats des valeurs de dispersions

Les colonnes des dispersions de 1 à 4 du **tableau 5.1** représentent les résultats des dispersions des microARNs du blé. Les dix premières lignes sont seulement affichées car le nombre d'observation de la base de données est de 220 lignes (**Chap**

4. section Données de simulation). Par analogie, nous procédons de la même manière aux tableaux 5.2 et tableau 5.3. Mais les variables et les temps dans toutes les dispersions sont différentes.

— Valeurs de dispersions des années 2010 et 2013 au temps T1

Tableau 5.1 Résultats des 10 premières observations des valeurs de dispersions temps T1. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T1-1 à T1-4)

Indices	T1-1	T1-2	T1-3	T1-4
1	0.08	0.05	0.06	0.09
2	0.04	0.13	0.043	0.06
3	0.03	0.12	0.04	0.06
4	0.04	0.08	0.05	0.06
5	0.07	0.06	0.05	0.08
6	0.03	0.03	0.08	0.07
7	0.08	0.05	0.06	0.09
8	0.07	0.05	0.06	0.09
9	0.08	0.05	0.06	0.09
10	0.08	0.05	0.06	0.09

— Valeurs de dispersions des années 2010 et 2013 au temps T5

Tableau 5.2 Résultats des 10 premières observations des valeurs de dispersions temps T5. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T5-1 et T5-2)

Indices	T5-2	T5-2
1	0.08	0.07
2	0.25	0.13
3	0.19	0.127
4	0.21	0.11
5	0.09	0.07
6	0.06	0.03
7	0.08	0.07
8	0.07	0.06
9	0.07	0.05
10	0.08	0.05

— Valeurs de dispersions des années 2010 et 2013 aux temps T1 et T5

Tableau 5.3 Résultats des 10 premières observations des valeurs de dispersions temps T1 et T5. Les résultats complets sont sur les fichiers nommés (fichier-resultat-dispersion de T1-T5-1 à T1-T5-6)

Indices	T1-T5-1	T1-T5-2	T1-T5-3	T1-T5-4	T1-T5-5	T1-T5-6
1	0.06	0.06	0.07	0.09	0.08	0.06
2	0.03	0.23	0.10	0.01	0.10	0.15
3	0.03	0.17	0.09	0.01	0.10	0.18
4	0.03	0.11	0.10	0.07	0.10	0.14
5	0.06	0.07	0.07	0.05	0.09	0.08
6	0.04	0.03	0.04	0.08	0.06	0.05
7	0.06	0.06	0.06	0.09	0.09	0.06
8	0.05	0.06	0.06	0.08	0.07	0.05
9	0.05	0.05	0.06	0.071	0.07	0.05
10	0.05	0.05	0.07	0.10	0.08	0.05

5.2.4 Résultats des méthodes et tests d'expression différentielle

Dans le **tableau 5.5**, nous avons comme paramètres : La première colonne représente la séquence du microARN le **baseMean**, représente simplement la moyenne des valeurs de comptage normalisées, divisée par les facteurs de taille, prise sur tous les échantillons. le **log2FC**, ratio qui mesure la variation de l'expression d'un microARN entre deux conditions. Si log2 fold change (log2FC) est négatif le microARN est sous-exprimé tandis que si log2FC est positif le microARN est sous-exprimé. le **lfcSE** est l'estimation de l'erreur standard du log fold change. le **stat** est le test statistique appelé Wald statistic-test utilisé par l'approche DESeq2 La **pvalue** ,montre les proportions de chaque microARN en indiquant si une différence est significativement observée. La **padj** les P-values ajustées par la méthode DESeq2 La **padjBonf** représente les P-values ajustées par la méthode d'ajustement de Bonferroni. Cette approche est très conservatrice. Elle est peu souvent utilisée dans les études d'expression différentielle. Elle s'avère utile lorsque l'utilisateur cherche à ne détecter qu'un nombre réduit de microARNs différentiellement exprimés. Nous faisons ce même raisonnement analogue pour les autres tableaux aux temps **T5** et entre **T1 - T5**. Nous définissons les variables des résultats partiels des 10 premières observations statistiques et de l'expression différentielle.

Tableau 5.4 Tableau des variables

Noms	Variables
T1-1	Nor-y2010-T1..Man-y2010-T1 - X8S-y2013-T1..11W-y2013-T1
T1-2	X8S-y2010-T1..11W-y2010-T1 - Nor-y2010-T1..Man-y2010-T1
T1-3	X8S-y2010-T1..11W-y2010-T1 - Nor-y2013-T1..Man-y2013-T1
T1-4	X8S-y2013-T1..11W-y2013-T1 - Nor-y2013-T1..Man-y2013-T1
T5-1	X8S-y2010-T5..11W-y2010-T5 - Nor-y2013-T5..Man-y2013-T5
T5-2	X8S-y2013-T5..11W-y2013-T5 - Nor-y2013-T5..Man-y2013-T5
T1-T5-1	Nor-y2010-T1..Man-y2010-T1 - X8S-y2013-T5..11W-y2013-T5
T1-T5-2	X8S-y2010-T1..11W-y2010-T1 - Nor-y2013-T5..Man-y2013-T5
T1-T5-3	X8S-y2010-T5..11W-y2010-T5 - Nor-y2010-T1..Man-y2010-T1
T1-T5-4	X8S-y2010-T5..11W-y2010-T5 - Nor-y2013-T1..Man-y2013-T1
T1-T5-5	X8S-y2013-T1..11W-y2013-T1 - Nor-y2013-T5..Man-y2013-T5
T1-T5-6	X8S-y2013-T5..11W-y2013-T5 - Nor-y2013-T1..Man-y2013-T1

— Tests d'expression différentielle des années 2010 et 2013 au temps T1

Tableau 5.5 Résultats des 10 premières observations statistiques et expression différentielle temps T1-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-1)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	76.21	-0.01	0.37	-0.01	0.99	0.99	1
TGATAATCCGGCACGAAAGTAC	44.29	0.22	0.36	0.61	0.54	0.99	1
TTTTGAAATGATAATCCGGCAC	40.13	-0.12	0.35	-0.56	0.58	0.99	1
TTTCGTGCCGGATTATCATTTTC	46.20	-0.05	0.36	-0.15	0.88	0.99	1
GAATTCTACGGTGGAAAGCTC	62.48	0.16	0.38	0.43	0.67	0.99	1
GAATTCTACGGTGGAAAAGCTC	122.73	-0.31	0.33	-0.96	0.34	0.99	1
TTTCCACCGTAGAATTCTGGT	71.61	-0.23	0.38	-0.60	0.55	0.99	1
ACCGTAGAATTCTGGTGCCAAA	87.03	0.01	0.37	0.01	0.99	0.99	1
AGACATATACAGTTTGGGTGTT	80.96	0.09	0.38	0.26	0.79	0.99	1
ATTCAAGTCAGACATATACAGT	75.50	-0.13	0.38	-0.34	0.74	0.99	1

Tableau 5.6 Résultats des 10 premières observations statistiques et expression différentielle temps T1-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-2)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	52.86	-0.14	0.31	-0.46	0.65	0.99	1
TGATAATCCGGCACGAAAGTAC	29.73	0.01	0.30	0.01	0.99	0.99	1
TTTTGAAATGATAATCCGGCAC	30.90	-0.08	0.30	-0.26	0.79	0.99	1
TTTCGTGCCGGATTATCATTTTC	36.22	0.02	0.31	0.05	0.96	0.99	1
GAATTCTACGGTGGAAAGCTC	41.43	-0.04	0.31	-0.14	0.90	0.99	1
GAATTCTACGGTGGAAAACCTC	107.23	0.03	0.29	0.10	0.92	0.99	1
TTTCCACCGTAGAATTCTGGT	57.47	-0.10	0.31	-0.33	0.74	0.99	1
ACCGTAGAATTCTGGTGCCAAA	56.78	-0.24	0.31	-0.78	0.43	0.99	1
AGACATATACAGTTTGGGTGTT	56.87	-0.02	0.31	-0.06	0.95	0.99	1
ATTCAAGTCAGACATATACAGT	55.75	-0.15	0.31	-0.49	0.62	0.99	1

Tableau 5.7 Résultats des 10 premières observations statistiques et expression différentielle temps T1-3. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-3)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	56.60	-0.27	0.33	-0.83	0.40	0.99	1
TGATAATCCGGCACGAAAGTAC	29.30	-0.02	0.32	-0.05	0.96	0.99	1
TTTTGAAATGATAATCCGGCAC	27.40	0.03	0.32	0.09	0.93	0.99	1
TTTCGTGCCGGATTATCATTTTC	33.98	0.08	0.32	0.24	0.81	0.99	1
GAATTCTACGGTGGAAAGCTC	38.79	0.01	0.32	0.03	0.98	0.99	1
GAATTCTACGGTGGAAAACCTC	114.40	-0.11	0.33	-0.35	0.73	0.99	1
TTTCCACCGTAGAATTCTGGT	56.70	-0.12	0.32	-0.37	0.71	0.99	1
ACCGTAGAATTCTGGTGCCAAA	55.70	-0.25	0.32	-0.77	0.44	0.99	1
AGACATATACAGTTTGGGTGTT	61.85	-0.20	0.32	-0.59	0.56	0.99	1
ATTCAAGTCAGACATATACAGT	57.71	-0.24	0.32	-0.73	0.46	0.99	1

Tableau 5.8 Résultats des 10 premières observations statistiques et expression différentielle temps T1-4. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-4)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	80.95	-0.16	0.50	-0.33	0.74	0.99	1
TGATAATCCGGCACGAAAGTAC	43.62	0.23	0.47	0.49	0.62	0.99	1
TTTTGAAATGATAATCCGGCAC	35.49	0.01	0.47	0.01	0.99	0.99	1
TTTCGTGCCGGATTATCATTTTC	43.27	0.04	0.47	0.08	0.93	0.99	1
GAATTCTACGGTGGAAAGCTC	58.84	0.29	0.48	0.62	0.53	0.99	1
GAATTCTACGGTGGAAAACCTC	132.53	-0.50	0.46	-1.09	0.27	0.97	1
TTTCCACCGTAGAATTCTGGT	70.63	-0.29	0.49	-0.60	0.55	0.99	1
ACCGTAGAATTCTGGTGCCAAA	85.30	0.01	0.50	0.01	0.98	0.99	1
AGACATATACAGTTTGGGTGTT	87.38	-0.10	0.50	-0.17	0.90	0.99	1
ATTCAAGTCAGACATATACAGT	77.98	-0.25	0.50	-0.51	0.61	0.99	1

— Tests d'expression différentielle des années 2010 et 2013 au temps T5

Tableau 5.9 Résultats des 10 premières observations statistiques et expression différentielle temps T5-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T5-1)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	57.61	-0.14	0.43	-0.34	0.74	0.99	1
TGATAATCCGGCACGAAAGTAC	34.82	0.17	0.45	0.37	0.71	0.99	1
TTTTGAAATGATAATCCGGCAC	37.47	0.01	0.46	0.02	0.98	0.99	1
TTTCGTGCCGGATTATCATTTTC	36.57	-0.09	0.45	-0.21	0.83	0.99	1
GAATTCTACGGTGGAAAGCTC	53.48	-0.01	0.44	-0.02	0.98	0.99	1
GAATTCTACGGTGGAAAACCTC	111.66	0.16	0.40	0.39	0.70	0.99	1
TTTCCACCGTAGAATTCTGGT	59.90	0.06	0.43	0.15	0.88	0.99	1
ACCGTAGAATTCTGGTGCCAAA	63.97	-0.01	0.42	-0.01	0.10	0.99	1
AGACATATACAGTTTGGGTGTT	65.52	-0.05	0.42	-0.12	0.90	0.99	1
ATTCAAGTCAGACATATACAGT	63.33	-0.34	0.43	-0.79	0.43	0.99	1

Tableau 5.10 Résultats des 10 premières observations statistiques et expression différentielle temps T5-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T5-2)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	61.14	-0.11	0.33	-0.33	0.74	0.995	1
TGATAATCCGGCACGAAAGTAC	32.32	0.02	0.33	0.07	0.94	0.99	1
TTTTGAAATGATAATCCGGCAC	33.84	-0.19	0.33	-0.57	0.56	0.99	1
TTTCGTGCCGGATTATCATTTTC	36.77	-0.16	0.33	-0.48	0.63	0.99	1
GAATTCTACGGTGGAAAGCTC	60.45	0.09	0.33	0.30	0.77	0.99	1
GAATTCTACGGTGGAAAACCTC	109.54	0.01	0.31	0.03	0.98	0.99	1
TTTCCACCGTAGAATTCTGGT	62.63	0.04	0.33	0.12	0.90	0.99	1
ACCGTAGAATTCTGGTGCCAAA	67.73	0.01	0.33	0.01	0.99	0.99	1
AGACATATACAGTTTGGGTGTT	74.87	0.09	0.33	0.28	0.77	0.99	1
ATTCAAGTCAGACATATACAGT	75.48	-0.07	0.33	-0.22	0.82	0.99	1

— Tests d'expression différentielle des années 2010 et 2013 aux temps T1 et T5

Tableau 5.11 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-1. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-1)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	64.06	-0.09	0.34	-0.28	0.78	0.99	1
TGATAATCCGGCACGAAAGTAC	34.99	0.01	0.32	0.03	0.98	0.99	1
TTTTGAAATGATAATCCGGCAC	34.60	-0.24	0.33	-0.72	0.47	0.99	1
TTTCGTGCCGGATTATCATTTTC	38.08	-0.19	0.33	-0.60	0.56	0.99	1
GAATTCTACGGTGGAAAGCTC	59.27	0.26	0.34	0.77	0.45	0.99	1
GAATTCTACGGTGGAAAACCTC	120.55	-0.05	0.32	-0.17	0.86	0.99	1
TTTCCACCGTAGAATTCTGGT	70.02	-0.04	0.34	-0.12	0.90	0.99	1
ACCGTAGAATTCTGGTGCCAAA	74.53	-0.05	0.33	-0.16	0.87	0.99	1
AGACATATACAGTTTGGGTGTT	75.78	0.19	0.33	0.57	0.57	0.99	1
ATTCAAGTCAGACATATACAGT	74.41	0.06	0.33	0.19	0.85	0.99	1

Tableau 5.12 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-2. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-2)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	50.41	-0.15	0.35	-0.44	0.66	0.99	1
TGATAATCCGGCACGAAAGTAC	27.47	0.03	0.33	0.09	0.93	0.99	1
TTTTGAAATGATAATCCGGCAC	30.13	-0.10	0.34	-0.29	0.77	0.99	1
TTTCGTGCCGGATTATCATTTTC	34.87	0.01	0.35	0.01	0.99	0.99	1
GAATTCTACGGTGGAAAGCTC	42.92	-0.17	0.35	-0.49	0.62	0.99	1
GAATTCTACGGTGGAAAACCTC	97.65	0.13	0.32	0.39	0.69	0.99	1
TTTCCACCGTAGAATTCTGGT	51.30	0.00	0.35	0.00	0.99	0.99	1
ACCGTAGAATTCTGGTGCCAAA	51.32	-0.18	0.35	-0.52	0.60	0.99	1
AGACATATACAGTTTGGGTGTT	56.52	-0.09	0.34	-0.28	0.78	0.99	1
ATTCAAGTCAGACATATACAGT	56.99	-0.28	0.34	-0.81	0.42	0.99	1

Tableau 5.13 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-3. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-3)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	60.45	-0.12	0.32	-0.39	0.69	0.99	1
TGATAATCCGGCACGAAAGTAC	37.53	0.13	0.32	0.40	0.68	0.99	1
TTTTGAAATGATAATCCGGCAC	38.51	0.03	0.32	0.09	0.93	0.99	1
TTTCGTGCCGGATTATCATTTTC	37.97	-0.09	0.32	-0.27	0.78	0.99	1
GAATTCTACGGTGGAAAGCTC	52.03	0.09	0.32	0.31	0.76	0.99	1
GAATTCTACGGTGGAAAACCTC	122.52	0.04	0.30	0.14	0.89	0.99	1
TTTCCACCGTAGAATTCTGGT	66.97	-0.05	0.32	-0.17	0.86	0.99	1
ACCGTAGAATTCTGGTGCCAAA	70.46	-0.09	0.32	-0.27	0.78	0.99	1
AGACATATACAGTTTGGGTGTT	66.00	0.01	0.32	0.05	0.96	0.99	1
ATTCAAGTCAGACATATACAGT	61.92	-0.18	0.32	-0.57	0.57	0.99	1

Tableau 5.14 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-4. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-4)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	64.65	-0.29	0.42	-0.68	0.49	0.97	1
TGATAATCCGGCAGCAAAGTAC	36.94	0.22	0.33	0.68	0.49	0.97	1
TTTTGAAATGATAATCCGGCAC	34.43	0.28	0.32	0.87	0.38	0.97	1
TTTCGTGCCGGATTATCATTTTC	35.51	-0.07	0.32	-0.22	0.83	0.97	1
GAATTCTACGGTGGAAAGCTC	48.91	0.22	0.39	0.56	0.58	0.97	1
GAATTCTACGGTGGAAAACCTC	130.61	-0.12	0.41	-0.30	0.76	0.97	1
TTTCCACCGTAGAATTCTGGT	66.04	-0.08	0.42	-0.19	0.84	0.97	1
ACCGTAGAATTCTGGTGCCAAA	69.05	-0.11	0.41	-0.26	0.79	0.97	1
AGACATATACAGTTTGGGTGTT	71.59	-0.17	0.41	-0.43	0.67	0.97	1
ATTCAAGTCAGACATATACAGT	64.16	-0.30	0.42	-0.71	0.47	0.97	1

Tableau 5.15 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-5. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-5)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	72.66	0.01	0.48	0.01	0.99	0.99	1
TGATAATCCGGCAGCAAAGTAC	41.11	0.29	0.52	0.58	0.56	0.99	1
TTTTGAAATGATAATCCGGCAC	39.13	-0.22	0.51	-0.44	0.66	0.99	1
TTTCGTGCCGGATTATCATTTTC	44.47	-0.07	0.51	-0.13	0.89	0.99	1
GAATTCTACGGTGGAAAGCTC	64.11	0.04	0.49	0.083	0.93	0.99	1
GAATTCTACGGTGGAAAACCTC	110.89	-0.21	0.44	-0.48	0.63	0.99	1
TTTCCACCGTAGAATTCTGGT	63.63	-0.13	0.49	-0.27	0.79	0.99	1
ACCGTAGAATTCTGGTGCCAAA	79.25	0.12	0.47	0.26	0.79	0.99	1
AGACATATACAGTTTGGGTGTT	80.24	0.05	0.47	0.10	0.92	0.99	1
ATTCAAGTCAGACATATACAGT	77.02	-0.28	0.48	-0.60	0.55	0.99	1

Tableau 5.16 Résultats des 10 premières observations statistiques et expression différentielle temps T1-T5-6. Les résultats restants sont sur (fichier-test-expression-differentielle-T1-T5-6)

microARNs	baseMean	log2FC	lfcSE	stat	pvalue	padj	padjBonf
TATATTATCAGGTTTTAGGTCC	68.34	-0.26	0.38	-0.67	0.50	0.99	1
TGATAATCCGGCACGAAAGTAC	34.47	-0.01	0.41	-0.01	0.99	0.99	1
TTTTGAAATGATAATCCGGCAC	30.51	-0.035	0.40	-0.10	0.93	0.99	1
TTTCGTGCCGGATTATCATTTC	35.56	-0.08	0.41	-0.20	0.84	0.99	1
GAATTCTACGGTGGAAAGCTC	55.94	0.35	0.40	0.87	0.38	0.99	1
GAATTCTACGGTGGAAAACCTC	129.02	-0.23	0.37	-0.62	0.53	0.99	1
TTTCCACCGTAGAATTCTGGT	69.03	-0.06	0.38	-0.16	0.87	0.99	1
ACCGTAGAATTCTGGTGCCAAA	73.03	-0.06	0.38	-0.17	0.87	0.99	1
AGACATATACAGTTTGGGTGTT	81.38	0.02	0.37	0.05	0.96	0.99	1
ATTCAAGTCAGACATATACAGT	76.46	-0.02	0.38	-0.07	0.94	0.99	1

Remarque : Pour le reste des résultats globaux des dispersions et de l'expression différentielle, les fichiers Excel restants ont été déposés à la bibliothèque de l'UQAM.

5.3 Discussion et interprétation

La phase normalisation de nos résultats permet de s'assurer que nos données du blé sont exploitables et comparables. Cette partie sert à réduire les biais techniques expérimentaux de notre approche. Elle est aussi importante pour comparer les données des différentes microARNs entre elles (**figure 5.2**), (**figure 5.3**), (**figure 5.4**). Dans chaque figure ci-dessous on observe une différence entre les quatre variables du blé. Mais la phase de normalisation d'analyse d'expression différentielle permet de s'approcher des hypothèses favorables à une distribution Gaussienne ou normale des données dans notre approche proposée.

La phase d'estimation des dispersions dans (**Résultats d'estimation des dispersions**) permet d'analyser les microARNs du blé aux temps T1 ou T5. Nous

constatons que, les points noirs sont les estimations de dispersion pour chaque microARN obtenu en considérant les informations de chaque microARN séparément. Ces valeurs fluctuent fortement autour de leurs vraies valeurs. Par conséquent, nous ajustons la ligne de tendance rouge, qui montre la dépendance des dispersions par rapport à la moyenne, puis réduisons l'estimation de chaque microARN vers la ligne rouge pour obtenir les estimations finales (points bleus) utilisées dans le test d'hypothèse. Les cercles bleus au-dessus du principal "nuage de points" sont des microARNs qui ont des estimations de dispersion génétiquement élevées qui sont marquées comme des valeurs aberrantes de dispersion. Ces estimations ne sont donc pas rétrécies vers la ligne de tendance ajustée entre la vraie valeur et la dispersion des ces quatre catégories microARNs du blé. Comparant les résultats du temps T1 et T5, nous voyons que au temps T1 les estimations des dispersions (**figure 5.5**) sont plus proches de la droite comparées aux valeurs trouvées au temps T5. Car au temps T1 les valeurs sont centrées autour des vraies valeurs et de la droite d'ajustement. Les valeurs aberrantes sont nombreuses au temps T5 (**figure 5.6**). Les estimations des dispersions aux temps T1 et T5 montrent que les valeurs finales (**figure 5.7**) des microARNs sont presque positionnées sur la droite d'ajustement en rouge. Les tableaux des valeurs de dispersion (**tableau 5.1**), (**tableau 5.2**), (**tableau 5.3**) permettent de d'étayer notre interprétation, en se basant sur les valeurs numériques obtenues des dispersions des microARNs du blé. En résumé, les résultats des estimations des dispersions aux temps T1 et T5 sont plus nettes par rapport au temps T1 ou T5. Les valeurs aberrantes sont moins nombreuses aux temps T1 et T5, les valeurs des dispersions sont plus proches de la droite d'ajustement.

La partie **Résultats statistiques et test d'expression différentielle**, les résultats des tableaux des paramètres (**tableau 5.4**) jusqu'au (**tableau 5.15**) sont presque expliqués dans la section 5.2.4. Pour chaque tableau, nous avons le paramètre **Log2FC** qui contrôle l'expression du microARN. Au niveau des ta-

bleaux, nous calculons pour chaque microARN une valeur de p-value permettant d'observer, si le microARN est significativement exprimé par rapport à seuil afin de détecter les FDR (Faux Discovery Rate). Les résultats obtenus nous ont permis de faire une analyse d'expression différentielle des microARNs étudiés, mais dans la phase de normalisation, il n'est facile parfois de se rapprocher d'une distribution normale si les données sont dans des formats différents. Les données initiales nécessitent des pré-traitements avant d'entamer la normalisation proprement dite. Le problème de sur-dispersion peut être noté dans l'analyse de l'expression différentielle des microARNs mais notre approche a permis de minimiser cette sur-dispersion. Les approches d'ajustement comme celle de Bonferroni sont parfois conservatrices et restrictives dans les statistiques des expressions. Ces modèles d'ajustement incorporés qui ne fournissent pas parfois des valeurs satisfaisantes font partis des limites dans notre approche proposée.

5.4 Limites de notre approche

Les étapes de contrôle de la qualité des données tout au long de ce processus sont recommandées mais non obligatoires, et le fait de ne pas vérifier les caractéristiques du jeu de données peut entraîner des résultats erronés. Nous pouvons examiner différentes courbes pour vérifier que les données sont de bonne qualité et que les échantillons sont conformes aux attentes afin de mieux tenir en compte des gènes faiblement exprimés et que nous comptons stocker dans un objet `DGEList`. Les données de comptage ne sont normalement pas distribuées. Par conséquent, si nous voulons examiner les distributions des comptages bruts, nous devons enregistrer les comptages. Nous devrions utiliser ensuite des diagrammes en boîte pour vérifier la distribution des comptes de lecture sur l'échelle $\log_2$. Nous pourrions utiliser la fonction `vst` (variance-stabilising transformation) de `DESeq2` pour appliquer une transformation stabilisante de la variance. L'effet est de supprimer la dépendance de la variance sur la moyenne, en particulier la variance élevée du logarithme des

données de comptage lorsque la moyenne est basse. Les comptes résultants ont également été normalisés en fonction de la taille de la bibliothèque ou d'autres facteurs de normalisation. Une autre limite importante de notre rapport réside sur une analyse en composante principale (PCA) qui est un exemple d'analyse non supervisée, dans laquelle il n'est pas nécessaire de spécifier les groupes. Si notre approche était bien contrôlée et avait bien fonctionné, nous espérons que les principales sources de variation dans les données sont les groupes qui nous intéressent le plus. Il s'agit également d'un outil extrêmement utile pour le contrôle de la qualité et le contrôle des valeurs extrêmes pour faire une analyse d'expression différentielle.

5.5 Conclusion

Les résultats présentés dans ce chapitre, montrent les trois étapes fondamentales de toute analyse d'expression différentielle. Les quatre catégories de microARNs du blé simulés ont permis d'identifier les profils d'expression différentielle de façon plus précise. Car les résultats des histogrammes, des estimations des dispersions et des tests statistiques d'expression différentielle sont importants pour connaître la variabilité des microARNs étudiés dans ce rapport. Beaucoup de mutations se sont déroulées durant chaque étape de notre analyse des données. Nos résultats corroborent avec les principaux objectifs visés dans le chapitre précédent.

CHAPITRE VI

CONCLUSION ET PERSPECTIVES

Ce rapport présente plusieurs approches bioinformatiques adaptées aux caractéristiques particulières des données d'expression de microARNs. Dans ce mémoire, nous utilisons plusieurs familles de méthodes, telles que des distributions normales, de poisson et binomiales négatives, afin de déterminer des statistiques d'expression différentielle, basées sur des moments d'ordre inférieur observés tels que la moyenne et la variance. Le partage des informations, par exemple, sur les paramètres d'une distribution donnée, sur la base du principe que les données provenant de microARNs différents suivent des schémas de variabilité similaires. Nous avons proposé une approche basée sur DESeq2 qui permet de mieux tenir en compte des variabilités entre les différentes conditions et cultivars. Pour y arriver, la diminution de certains estimateurs dans notre approche a contribué grandement à améliorer la fiabilité et la cohérence des résultats obtenus par rapport aux autres types de méthodes d'expression différentielle proposées dans la littérature. Les trois grandes étapes de notre approche qui sont la normalisation, l'estimation des dispersions et les tests d'expression différentielle sont primordiales pour toute étude d'analyse d'expression de données génomiques. Nos résultats obtenus et comparés à ceux de (Varet *et al.*, 2016) et (Anders et Huber, 2010) permettent de confirmer la pertinence de notre approche sur l'analyse des profils d'expression des microARNs du blé. Car dans la partie des tests d'expression différentielle notre

approche fournit des résultats statistiquement plus significatifs par rapport aux autres propositions développées. A cela s'ajoute que la définition et l'estimation des paramètres des modèles d'ajustement sont bien pris en compte dans notre approche.

En perspective, comme toute approche, nous avons noté des limites pour la mise en place de méthode d'expression différentielle basée sur DESeq2. Dans ce mémoire, nous constatons que les gènes ne sont pas très bien différencier si la taille de notre base données est petite. Les différentes conditions ou mutations des microARNs affectent une faible proportion des gènes qui sont sur-exprimés et sous-exprimés durant la phase d'estimation des gènes. On peut noter aussi des limites sur les gènes différentiellement exprimés en cherchant des valeurs de $\log_2$ de ratios significativement supérieures à 1 ou inférieures à -1 et en utilisant le test de **Wald**. L'approximation de la méthode de Bonferroni est en général conservatrice et restrictive dans les calculs statistiques des expressions des microARNs. Cette approche d'ajustement incorporé ne donne pas des valeurs normales si la séquence du microARNs est très petite. Nous continuerons d'explorer d'autres approches d'analyse d'expression pour mieux différencier les gènes qui sont sur-exprimés et sous-exprimés. Car on ne peut trouver un modèle fiable et performant en analyse d'expression différentielle du fait des conditions biologiques, techniques et des réplicats de certaines données. Les étapes de l'analyse en composante principale (ACP) permet d'identifier le degré d'expression différentielle de chaque microARN de notre étude. La classification hiérarchique ascendante des microARNs détermine les clusters des gènes exprimés suivant les types de classes identifiées et l'établissement du heatmap pour représenter le niveau d'expression de nombreux gènes sur un certain nombre d'échantillons comparables, est une partie intégrante dans le processus d'analyse différentielle pour compléter notre étude sur les microARNs du blé.

APPENDICE A

CODE RSTUDIO IMPLÉMENTÉ

```
#source('http://Bioconductor.org/biocLite.R')
#biocLite('DESeq2')
#####importation des libraires de travail#####
library(DESeq2)
library( Biobase)
library (gplots)
library('RColorBrewer')
library(limma)
mirna <- read.table("C:/data4.txt",header=T, row.names=1, sep="\t")
#View(mirna)
dim(mirna)
mirna= mirna[,-1]
mirna[is.na(mirna)] <- 0
#####
# Boucle sur la combinaison des colonnes
#####
combinations=combn(c(1:ncol(mirna)),2) # Toutes les combinaison de colonnes
for (i in 1:ncol(combinations)) {
  col1=combinations[1,i] #Index des colonnes
  col2=combinations[2,i]
  datas=mirna[,c(col1,col2)]
  #nom = paste(col1,"_",col2)
  colname1=colnames(datas)[1] #Noms des colonnes
  colname2=colnames(datas)[2]
  nom = paste(colname1,"_",colname2)
  condition= factor(colnames(datas))
  type= factor(rep ("single",2))
  colData= data.frame(condition,type,row.names=colnames(datas)) #
  ## creation de l'objet dds
```

```

dds <- DESeqDataSetFromMatrix(datas, colData, formula(~ condition))
colData(dds)$condition= relevel(colData(dds)$condition, ref=colnames(datas)[1])
# creation de objet dds
class(dds)
colData(dds)
design(dds)
# manipulation des données de comptage
dim(counts(dds))
head(counts(dds))
summary(counts(dds))
#visualisation des données
## combien il y a-t-il de comptages nuls par intervalle de temps ?
apply(counts(dds), 2, FUN= function(x) sum(x ==0))
## quelles sont les profondeurs de gènes de chaque intervalle de temps ?
colSums(counts(dds))
par(mfrow = c(1,2))
barplot(colSums(counts(dds)))
## distribution des gènes suivant leur log-fréquence pour chaque intervalle de temps
par(mfrow = c(1,2))
for(i in 1:2){
hist(log(counts(dds)[,i] +1) ,
     breaks = seq(0,14,1), col="grey", cex.lab =0.8,
     main = colnames(counts(dds))[i] ,
     xlab="Valeur du comptage", ylab="Frequence du comptage")
dev.print(device = png, file = paste("C:/histo-log-frequence", nom, ".png"), width = 1000)
}
#filtrage arbitraire : on prend en compte que les gènes avec un comptage moyen superieur à 5
count_mean5= datas[rowMeans(datas)>=5,]
dds= DESeqDataSetFromMatrix(count_mean5,colData,design=~condition)
dim(counts(dds))
par(mfrow = c(1,2))
for(i in 1:2){
hist(log(counts(dds)[,i] +1) ,
     breaks = seq (0,14,1), col="grey", cex.lab =0.8,
     main = colnames(counts(dds))[i],
     xlab="Valeur du comptage", ylab="Frequence du comptage")
dev.print(device = png, file = paste("filtrage-arbitraire", nom, ".png"), width = 1000)
}
## la fonction DESeq( ) réalise la normalisation et l'analyse diff en une seule étape
## la normalisation des donnees de comptage se fait par la méthode " effective library size" .
dds= DESeq(dds, quiet=TRUE, fitType="local")
colData(dds)

```

```

###Visualisation de la normalisation ###
#profondeur de gènes après normalisation
par(mfrow = c(1, 2))
colSums(counts(dds,normalized = TRUE))
barplot(colSums(counts(dds,normalized = TRUE)),main="Profondeur de gènes apres normalisation")
dev.print(device = png, file =paste("histogramme-normalise", nom, ".png"), width = 1000)
#données de comptage normalisées
head(counts(dds,normalized = TRUE),n=13)
summary(counts (dds,normalized = TRUE))
# distribution des gènes suivant leur log-fréquence après normalisation
par(mfrow = c( 1, 2))
for( i in 1:2) {
  hist(log(counts(dds)[,i] +1) ,
       breaks = seq(0,14,1),col="grey",main = colnames(counts(dds))[i],
       xlab="Valeur du comptage",ylab="Frequence du comptage",cex.lab=0.8)
  dev.print(device = png, file = paste(histo-log-freq-normalise", nom, ".png"), width = 1000)
}
# DESeq a effectué une estimation de la dispersion des gènes et Valeurs de dispersion
dispersion= as.data.frame(dispersions(dds))
head(dispersion,n=13)
print(summary(dispersion))
write.table(file=paste("disp", nom, ".xls"),row.names=TRUE, col.names=NA, quote=FALSE)
# Boxplot des valeurs de dispersion
par(mfrow=c( 1,1))
boxplot(sqrt(dispersion),horizontal=T,las =1,cex =0.5)
title("racine carrée de la dispersion calculee par DESeq2")
dev.print(device = png, file = paste("/dossier5/boxplots", nom, ".png"), width = 700)
par(mfrow = c( 1 , 1))
DESeq2::plotDispEsts(dds)
title("Relation entre la dispersion et la moyenne des comptages")
dev.print(device = png, file = paste(relation-disper-moyenne", nom, ".png"), width = 700)
## statistiques et pvalues ###
details= mcols(dds,use.names=T)
head(details,n=13)
mcols(mcols(dds,use.names=T))
#relation entre le comptage moyen d'un gène et son log2-ratio entre les variables
par(mfrow = c( 1,1))
DESeq2::plotMA(dds)
dev.print(device = png, file = paste("dossier7/Maplot des donn", nom, ".png"), width = 700)
# les gènes différenciellement exprimés au seuil de 10% après ajustement
# des tests multiples par procédure de Benjamini-Hochberg(BH) sont indiqués en rouge
#resultat contient les résultats de l'analyse différentielle

```



```

resultat= results(dds)
write.table(file=paste("anl-dif", nom, ".xls"), row.names=TRUE, col.names=NA, quote=FALSE)
head(resultat)
resultsNames(dds)
mcols(resultat, use.names=T)
# distribution des p-valeurs brutes des modeles d'ajustement
par(mfrow = c( 1 , 1))
hist(resultat$pvalue, breaks =100, border=" slateblue" ,
      main="Histogramme de la distribution des p-values brutes")
dev.print(device = png, file = paste("histo des pvalues brutes", nom, ".png"), width = 700)
#Combien de gènes sont déclarés DE à 5% ?
table(resultat$padj < 0.05, useNA="always")
# Selection des gènes déclarés DE à 5% dans une table ordonnée par p-value croissante
resSig= na.omit(resultat)
resSig= resSig [resSig$padj < 0.05,]
resSig= resSig[order(resSig$pval),]
nrow (resSig)
head(resSig)
# calcul des p-values ajustées selon la méthode de bonferroni
resultat$padjBonf= p.adjust(resultat$pval, meth="bonferroni")
head(resultat)
#histogramme des distributions des p-values
par(mfrow=c(1,1))
hist(resultat$padjBonf, border = "blue", xlab = "pvalue",
      main="distribution des p-values")
hist( resultat$padj, add = TRUE, border = "red")
hist( resultat$pval, add = TRUE)
legend(x="center", legend = c ( "pVal", "padjBonf", "padjBH"), fill=c( "black ", "blue", "red"))
dev.print(device = png, file = paste("dossier9/histo-des-3-modeles", nom, ".png"), width = 1000)
write.table(file=paste("pv-BF-BH", nom, ".xls"), row.names=TRUE, col.names=NA, quote=FALSE)
# p-values cumulatives des trois modeles d'ajustement
resordered= resultat[order(resultat$pval),]
plot(resordered$pval, ylab="Pvalues", cex=0.5, xlab="rang du gene")
points(resordered$padjBonf, col="blue", cex =0.5)
points(resordered$padj, col="red", cex =0.5)
abline(h=0.05, lty =2)
legend(x = "bottomright", legend = c("pVal", "padjBonf", "padjBH"),
      col= c("black", "blue", "red"), pch=20)
title("Les P-values cumulatives")
dev.print(device = png, file = paste("dossier10/courbes-cumulatives", nom, ".png"), width = 1000)
}

```

BIBLIOGRAPHIE

- Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V. et Sundaresan, V. (2005). Computational prediction of mirnas in arabidopsis thaliana. *Genome research*, 15(1), 78–91.
- 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.
- 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.
- Anders, S. et Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10), R106.
- Bartel, D. P. (2004). MicroRNAs : genomics, biogenesis, mechanism, and function. *cell*, 116(2), 281–297.
- Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. *FEBS letters*, 579(26), 5904–5910.
- Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V. et Hannon, G. J. (2003). Dicer is essential for mouse development. *Nature genetics*, 35(3), 215.
- Bray, N. L., Pimentel, H., Melsted, P. et Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5), 525.
- Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. et Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1), 25–36.
- Brennecke, J., Stark, A., Russell, R. B. et Cohen, S. M. (2005). Principles of microRNA–target recognition. *PLoS biology*, 3(3), e85.
- BRET, C. et SCHVED, J. (2009). Le contrôle de l'expression des gènes par les microarn : Implications au cours de l'hématopoïèse et des hémopathies malignes. *Correspondances en onco-hématologie*, 4(1), 15–19.
- Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L. et Voinnet, O. (2008). Widespread translational inhibition by plant mirnas and sirnas. *Science*, 320(5880), 1185–1190.

- Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. *et al.* (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences*, 101(9), 2999–3004.
- Cavaillé, J. (2004). Des microarn comme s' il en pleuvait. . . . *médecine/sciences*, 20(4), 399–401.
- Chen, X. (2004). A microRNA as a translational repressor of *apetala2* in arabidopsis flower development. *Science*, 303(5666), 2022–2025.
- Chen, X. (2005). MicroRNA biogenesis and function in plants. *FEBS letters*, 579(26), 5923–5931.
- Costa-Silva, J., Domingues, D. et Lopes, F. M. (2017). Rna-seq differential expression analysis : An extended review and a software tool. *PloS one*, 12(12), e0190152.
- Cuperus, J. T., Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Burke, R. T., Takeda, A., Sullivan, C. M., Gilbert, S. D., Montgomery, T. A. et Carrington, J. C. (2010). Unique functionality of 22-nt mirnas in triggering rdr6-dependent siRNA biogenesis from target transcripts in arabidopsis. *Nature structural & molecular biology*, 17(8), 997.
- Dardel, F. et Képès, F. (2002). *Bioinformatique : Génomique et post-génomique*. Editions Ecole Polytechnique.
- Dharmasiri, N., Dharmasiri, S. et Estelle, M. (2005). The f-box protein tir1 is an auxin receptor. *Nature*, 435(7041), 441.
- Esquela-Kerscher, A. et Slack, F. J. (2006). Oncomirs—microRNAs with a role in cancer. *Nature Reviews Cancer*, 6(4), 259.
- Fleming, A. J. (2005). Formation of primordia and phyllotaxy. *Current opinion in plant biology*, 8(1), 53–58.
- Hardcastle, T. J. et Kelly, K. A. (2010). bayseq : empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1), 422.
- Hartmann, C., Corre-Menguy, F., Boualem, A., Jovanovic, M. et Lelandais-Brière, C. (2004). Les microarn-une nouvelle classe de régulateurs de l'expression génique. *médecine/sciences*, 20(10), 894–898.
- Houbaviy, H. B., Murray, M. F. et Sharp, P. A. (2003). Embryonic stem cell-specific microRNAs. *Developmental cell*, 5(2), 351–358.
- Ibrahim, F., Rohr, J., Jeong, W.-J., Hesson, J. et Cerutti, H. (2006). Untemplated oligoadenylation promotes degradation of risc-cleaved transcripts. *Science*, 314(5807), 1893–1893.
- Jack, T. (2004). Molecular and genetic mechanisms of floral control. *The Plant Cell*, 16(suppl 1), S1–S17.
- Jones-Rhoades, M. W. et Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced mirna. *Molecular cell*, 14(6), 787–799.

- Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D. M. et Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes & development*, 19(4), 489–501.
- Kong, Y. W., Ferland-McCollough, D., Jackson, T. J. et Bushell, M. (2012). microRNAs in cancer management. *The lancet oncology*, 13(6), e249–e258.
- Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., Da Piedade, I., Gunsalus, K. C., Stoffel, M. et al. (2005). Combinatorial microRNA target predictions. *Nature genetics*, 37(5), 495.
- Lai, E. C. (2004). Predicting and validating microRNA targets. *Genome biology*, 5(9), 115.
- Law, C. W., Chen, Y., Shi, W. et Smyth, G. K. (2014). voom : Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2), R29.
- 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.
- Lee, Y. S., Kim, H. K., Chung, S., Kim, K.-S. et Dutta, A. (2005). Depletion of human microRNA mir-125b reveals that it is critical for the proliferation of differentiated cells but not for the down-regulation of putative targets during differentiation. *Journal of Biological Chemistry*.
- Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. et Kendziora, C. (2013). Ebsseq : an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8), 1035–1043.
- Li, J. et Tibshirani, R. (2013). Finding consistent patterns : a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 22(5), 519–536.
- 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.
- 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, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A. et al. (2005). MicroRNA expression profiles classify human cancers. *nature*, 435(7043), 834.
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. et Gilad, Y. (2008). RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*.

- Nath, U., Crawford, B. C., Carpenter, R. et Coen, E. (2003). Genetic control of surface curvature. *Science*, 299(5611), 1404–1407.
- Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C. et Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature*, 425(6955), 257.
- Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. et Ruvkun, G. (2000). The 21-nucleotide let-7 rna regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403(6772), 901.
- 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.
- Robinson, M. D., McCarthy, D. J. et Smyth, G. K. (2010). edgeR : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.
- Robinson, M. D. et Smyth, G. K. (2007a). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21), 2881–2887.
- Robinson, M. D. et Smyth, G. K. (2007b). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2), 321–332.
- Rogers, K. et Chen, X. (2013). Biogenesis, turnover, and mode of action of plant microRNAs. *The Plant Cell*, tpc-113.
- Smyth, G. K., Ritchie, M., Thorne, N., Wettenhall, J., Shi, W. et Hu, Y. (2002). limma : Linear models for microarray and rna-seq data user's guide.
- Smyth, G. K. et Verbyla, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 565–572.
- Soneson, C. et Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1), 91.
- Suh, M.-R., Lee, Y., Kim, J. Y., Kim, S.-K., Moon, S.-H., Lee, J. Y., Cha, K.-Y., Chung, H. M., Yoon, H. S., Moon, S. Y. et al. (2004). Human embryonic stem cells express a unique set of microRNAs. *Developmental biology*, 270(2), 488–498.
- Sunkar, R. et Zhu, J.-K. (2004). Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *The Plant Cell*, 16(8), 2001–2019.
- Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A. et Conesa, A. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic acids research*, 43(21), e140–e140.

- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. et Conesa, A. (2011). Differential expression in rna-seq : a matter of depth. *Genome research*, gr-124321.
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. et Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1), 46.
- van Rooij, E. (2011). The art of microRNA research. *Circulation research*, 108(2), 219-234.
- Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. et Dillies, M.-A. (2016). Sartools : a deseq2-and edger-based r pipeline for comprehensive differential analysis of rna-seq data. *PLoS One*, 11(6), e0157022.
- Watson, J. et Crick, F. (1953). A structure for adn. *Nature*, 171(4356), 737-8.
- Wienholds, E. et Plasterk, R. H. (2005). MicroRNA function in animal development. *FEBS letters*, 579(26), 5911-5922.
- Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C. et Qi, Y. (2010). Dna methylation mediated by a microRNA pathway. *Molecular cell*, 38(3), 465-475.
- Xu, J., Yang, J.-Y., Niu, Q.-W. et Chua, N.-H. (2006). Arabidopsis dcp2, dcp1, and varicose form a decapping complex required for postembryonic development. *The Plant Cell*, 18(12), 3386-3398.