UNIVERSITÉ DU QUÉBEC À MONTRÉAL

BIOINFORMATICS APPROACH FOR THE IDENTIFICATION OF FRAGILE REGIONS ON THE HUMAN GENOME

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

OF THE MASTERS OF BIOLOGY

BY

GOLROKH KIANI

SEPTEMBER 2015

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPROCHE BIOINFORMATIQUE POUR L'IDENTIFICATION DES RÉGIONS FRAGILES SUR LE GÉNOME HUMAIN

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN BIOLOGIE

PAR

GOLROKH KIANI

SEPTEMBRE 2015

DEDICATION AND ACKNOWLEDGEMENTS

No research could be performed without the assistance and intellectual comradeship of many individuals. At the outset my appreciation goes to my supervisor, Dr. Abdoulaye Baniré Diallo. I admire him the most as my mentor, friend and a very respectful member of the society. I am honored to work under his supervision. His confidence in me pushed me the most through my years of working with him. He helped me to grow and to challenge myself in ways I never thought I could in my field of study. By his constant availability and day-to-day support he taught me how to figure out my problems and solve them.

I would like also to express my gratefulness to my co-adviser, Dr. Emmanuel Mongin, by whom this project has been initially inspired and he continued to guide and encourage me to the end and hopefully in future. He helped me through the study design, read and corrected my scripts.

I would like to express my warmest gratitude to Bruno Daigle, as a good friend who I love and admire greatly. His patience and full-time technical assistance eased my way during this thesis.

I am profusely thankful to my good friend and colleague, Mohammed Amine Remita with whom I started this project in the first place. His love and support accompanied me all along to the end of this road.

I also take this opportunity to acknowledge Dr. Alix Boc as my helpful and patient teacher who also provided me with one of his programming script. So as Etienne Lord and Mickael Leclercq for their support and guidance since my first days at

UQAM.

I should as well mention all my classmates during my first year at UQAM and members of our laboratory who assisted me patiently with my French language problems, which accelerated my integration process in Canada. And of-course the friendly and hospitable environment of UQAM left me nothing but joyful memories from these years.

To my grandparents for their years of guidance and efforts during my whole life to whom I am greatly indebted for my personality and open-mindedness.

Finally, I would like to acknowledge my best friend in life, Rastin Azizbigloo, whose encouragements let me recognize and appreciate myself and my potentials like no one has ever done for me.

I dedicate this study to Mehrdad Aavani, Helen Lachini and Rastin Azizbigloo to whom I owe my life and my happiness. Without their presence the imagination of such achievement even seems to be impossible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ABSTRACT

Evolution is emergence and disappearance of species traits due to small-scale mutations and genome rearrangements in favour of species fitness to their dynamic environment. Genome rearrangements happen when DNA breaks in two or more positions (***breakpoint***) and reassembles in a different order. Comparative analyses of contemporary genomes have shown that several genomic regions have been resistant to any form of structural modification (synteny regions). This could indicate the functional implication of those regions in survival and/or reproduction of the species. Whereas the genomic regions that have been more subjected to rearrangements (fragile regions) could be for example, associated to traits that differentiate species. Our objective in this Master thesis was to design and develop an approach to identify these fragile regions on human genome. Hence, we selected 11 well-sequenced vertebrates, 10 from different major superorders of mammalian tree of life and chicken as an outgroup. We then extracted the multiple sequence alignment (MSA) of the selected genomes from the MSA of 45 species against human genome available on UCSC genome browser public site. Using comparative analysis and Lowest Common Ancestor method we have identified 33,424 human lineage specific breaks on chromosome 1. With a sliding window approach on the chromosome 1, we computed the enrichment breakpoints of the regions of chromosome 1. We identified 72 fragile regions of size 70 Kbp to 140 Kbp. These regions are associated to genes known to be associated to disease. Finally, the developed approach will constitute an ideal framework to study the whole genome and then exploit the predictions in the study of the correlation between fragile regions and cancer associated rearrangements.

Keywords. *breakpoint, synteny regions, rearrangements, comparative analysis, fragile regions.*

RÉSUMÉ

L'évolution est l'émergence et la disparition des caractéristiques des espèces dues à des mutations à petite échelle et réarrangements génomiques en faveur de l'adaptation des espèces à leur environnement dynamique. Les réarrangements génomiques se produisent lorsque l'ADN se casse en deux ou plusieurs positions (***points de rupture***) et se remonte dans un ordre différent. Des analyses comparatives de génomes contemporains ont montré que plusieurs régions génomiques ont été résistantes à toute forme de modification structurelle (régions de synténie). Ce qui pourrait indiquer l'implication fonctionnelle de ces régions en matière de survie et/ou de reproduction de l'espèce. Considérant que les régions génomiques qui ont été plus soumises aux réarrangements (régions fragiles) pourraient être, par exemple, associées à des traits qui distinguent les espèces. Notre objectif dans ce mémoire de maîtrise est de concevoir et de développer une approche pour identifier ces régions fragiles dans le génome humain. Par conséquent, nous avons sélectionné 11 vertébrés bien séquencés, 10 à partir de différents grands superordres de mammifères arbre de la vie et le poulet comme exogroupe. On a extrait ensuite l'alignement de séquences multiples (MSA) des génomes sélectionnés à partir de l'alignement multiple de séquences de 45 espèces contre le génome humain disponibles sur le "UCSC genome browser". En utilisant une approche basé sur l'analyse comparative et la méthode du plus proche ancêtre commun, nous avons identifié 33,424 cassures de synténies sur le chromosome 1. Avec une approche de fenêtre coulissante passée sur le chromosome 1, nous avons calculé l'enrichissement des cassures de synténies sur les différentes régions du chromosome 1. Cela a permis d'identifier 72 régions fragiles de taille variant de 70 Kbp à 140 Kbp. Ces régions sont associées à des gènes connus pour être associés à plusieurs maladies. Enfin, l'approche développée constitue un cadre idéal pour étudier le génome complet, puis exploiter les prévisions de l'étude dans la corrélation entre les régions fragiles et les réarrangements associés au cancer.

Mots clés. *points de rupture, régions de synténie, réarrangements, analyses comparatives, régions fragiles.*

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

Biological information needed for every living organism to survive and reproduce is encoded in its genome. The genomes of all organisms, except a group of viruses, consist of a double stranded deoxyribonucleic acid (DNA). DNA is a linear polymer of a combination of four types of a monomeric structure called nucleotide. Each nucleotide is made up of three components : 1) a pentose sugar (2-deoxyribose), 2) a nitrogenous base (adenine, cytosine, guanine and thymine. 3) a phosphate group. These four nucleotides are as follows : Adenosine triphosphate (dATP), Cytidine triphosphate (dCTP), Guanosine triphosphate (dGTP) and Thymidine triphosphate (dTTP), or when referring to a DNA sequence, A, C, G and T, respectively. See Figure 1.1. The double stranded DNA forms when two single-stranded DNA molecules coil around each other in opposite direction and hydrogen bonds-interactions pair the bases on the two strands (Base-pairing). These hydrogen bonds are specifically between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine. The double helix structure gives the DNA enough stability to protect genomic information. Figure 1.2 shows the commonest structural conformation of the DNA double helix in living cells (**B-conformation**). In this conformation the DNA double helix is

about 2 nm or 20 Å in diameter.

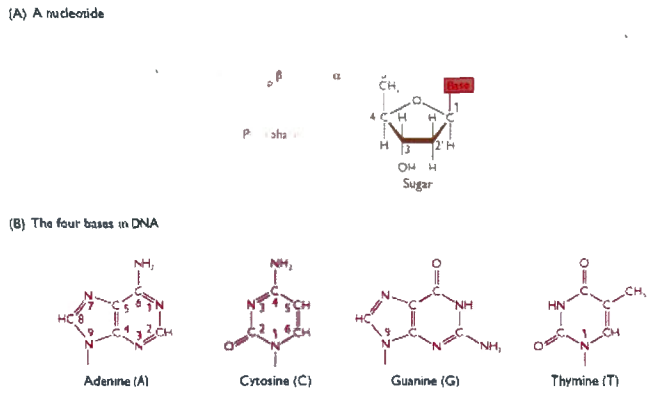


Figure 1.1: "(A) The general structure of a deoxyribonucleotide, the type of nucleotide found in DNA. (B) The four bases that occur in deoxyribonucleotides" (Brown, 2002).

1.1.1 Genome organization in vertebrates

Genome size in vertebrates varies from 100Mb (mega bases pairs or millions of base pairs) to several Gb (giga base pairs or billions of base pairs) while human haploid genome size is ~3.08 Gb(Consortium *et al.*, 2004b). The majority of genomic material in all vertebrates is enclosed in the nucleus by the nuclear membrane. This portion of the genome is called nuclear genome, and is divided in big segments, which are wrapped around octamers of histone proteins. This structure is known as *'beads-on-a-string'*. Each segment is coiled and compacted into 30 nm fibers called chromatin structure (DNA-histone complexes). See Figure 1.3. This 30 nm form is the most common form of the chromatin in the nucleus between each cell division cycle. Furthermore, during the cell division, the 30 nm structure coils and packs more and more (super coiling) to its most compact form, which would be visible under the light microscope in metaphase stage of cell division (Bernardi, 2005; Brown, 2002). This condensed structure is also known as ***chro-***

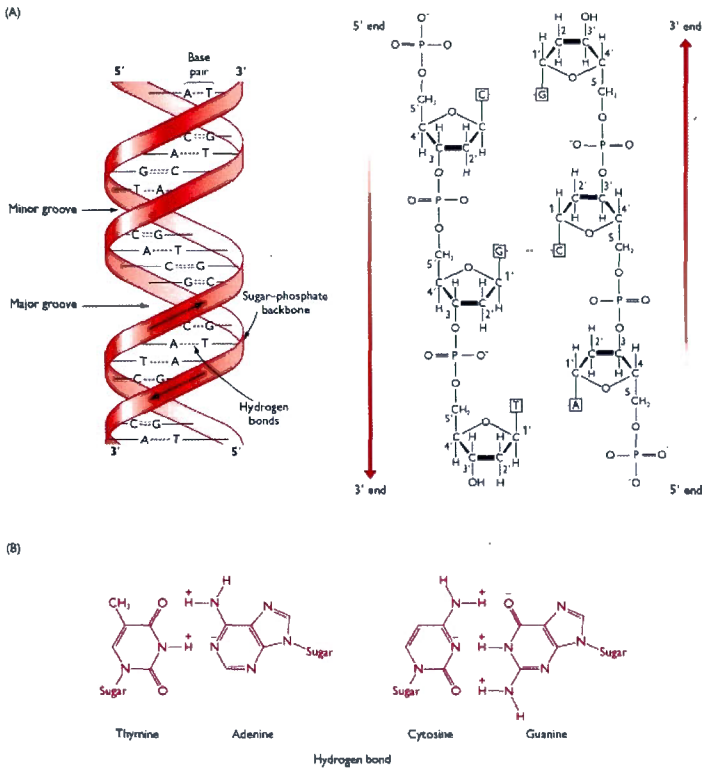


Figure 1.2: "(A) Two representations of a DNA double helix. On the left the structure is shown with the sugar-phosphate 'backbones' of each polynucleotide drawn as a red ribbon with the base pairs in black. On the right the chemical structure of the sugar backbone for three base pairs is given. (B) A base-pairs with T, and G base-pairs with C. The bases are drawn in outline, with the hydrogen bonding indicated by dotted lines. Note that each G-C base pair has three hydrogen bonds whereas an A-T base pair has just two. The structures in part (A) are redrawn from Turner et al. (1997) (left) and Strachan and Read (1999) (right)." (Brown, 2002). The figure is taken from `http://www.ncbi.nlm.nih.gov/books/NBK21134/`

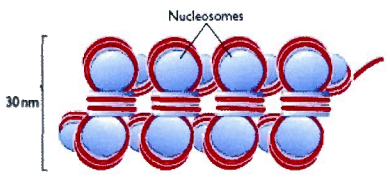


Figure 1.3: "The solenoid model for the 30 nm chromatin fiber : The *beads-on-a-string* structure of chromatin is condensed by winding the nucleosomes into a helix with six nucleosomes per turn." (Brown, 2002).

mosome. Metaphase is the stage after the termination of DNA replication. Each chromosome in this stage has two copies of a replicated DNA segment. The two copies are attached together at some place on the chromosome structure called *centromere* (See Figure 1.4). Centromere on each chromosome has specific location. Thus, chromosomes could be distinctly identified by virtue of their size and the location of their centromeres. Chromosomes are also distinguishable in terms

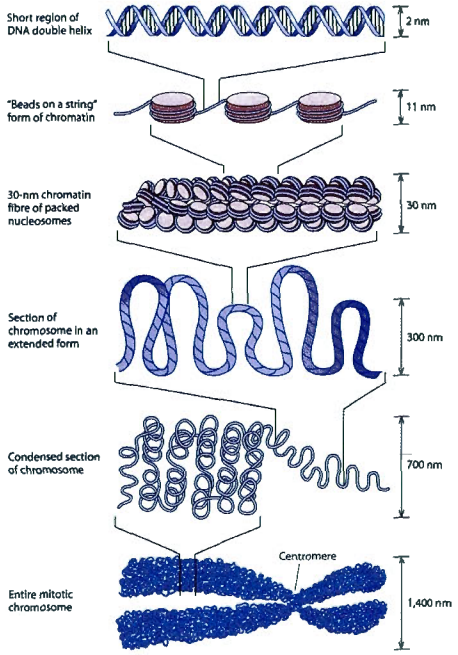


Figure 1.4: The figure shows the steps in DNA packaging from a double stranded DNA to chromosome structure during the cell division. The figure is taken from : `http://imgkid.com/nucleosome-structure.shtml`

of the patterns of reaction to different staining methods. Any staining method results in a banding pattern, which is specific to each chromosome. This is due to the non-homogeneous chemical nature of the genome. Figure 1.5 shows the human chromosome 1 stained with Giemsa stain. Dark bands represent regions with higher A=T pairings as light bands represent C≡G. Other than linear nuclear genome, mitochondria carry the rest of genome content in the form of a circular

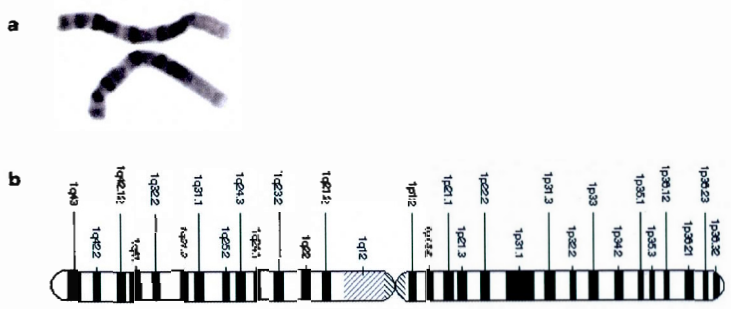


Figure 1.5: a : Human chromosome 1 stained by Giemsa stain. It is taken from : http://http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?TAXID=9606&CHR=1&MAPS=ideogr[0.00:15100.00]. b : Human chromosome 1 ideogram. It is taken from : http://ghr.nlm.nih.gov/chromosome/1

dsDNA. Mitochondria are the power house of the cells and their genome replicates independently from nuclear genome (Brown, 2002).

1.1.2 Regions of vertebrate genome

Other than the non-homogeneous chemical nature of the genome, as it is observable in their reactions to stains in karyograms, genomic regions are not homogeneous in terms of their functions either. Such unique differences could be used to categorize genomic regions. Genomic regions could be classified in the following groups :

- **Coding :** Codes for proteins and RNAs (Coding and noncoding genes)
- **Regulational :** Regulates those coding regions (Promoters)
- **Structural :** Responsible for genome structure (Centromeres and telomeres)
- **Non functional :** Regions with no evidence of functional activities

Genes

Genes may be the most important part of the human genome as they carry biological information that code for biological molecules (polypeptide/protein and RNA molecules). Most genes are expressed through an intermediate molecule, called messenger or mRNA, which is transported outside the nucleus and will be translated to specific protein in the cytoplasm. Another group of genes are not protein coding genes and they code for non-coding RNAs, which play various roles in the cell such as regulation (Brown, 2002).

Other genomic regions

- Pseudogenes : Genome is constantly subjected to changes and modifications. One of the products of such modifications could be genes that have lost their functions. These non-functional genes are called *Pseudogenes* (Brown, 2002).
- Repetitive DNA : Repetitive DNA seems to be originated from transposable element, which are DNA segments that jump from one place to another and leave a copy of themselves as they move (Brown, 2002). These repetitive segments are known to have a higher rate of mutation and participate in genome rearrangements driving the evolution. Such rearrangements could modify gene regulation and expression without any modification in coding regions (Shapiro and von Sternberg, 2005). Also, it has been shown that modifications in repetitive regions could affect chromatin formation which suggested their structural function in the genome (Shapiro and von Sternberg, 2005).

1.1.3 Human genome organization

The size of human haploid nuclear genome is estimated to be around ~3.08 Gb. This DNA is divided into 23 pairs of chromosomes. 22 pairs are autosomes, and

two sex chromosomes, X and Y. (Brown, 2002). See Figure 1.6. The rest of the human genome is stored in mitochondria. The mitochondrial is much smaller than the nuclear genome. It has only 16 569 bp and contains 37 genes (Brown, 2002).

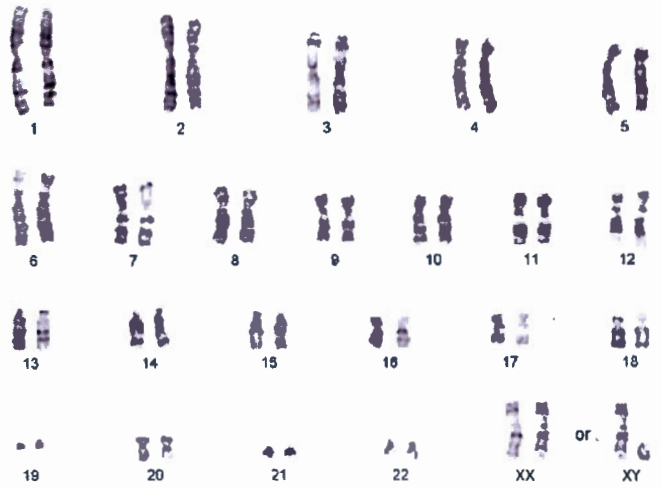


Figure 1.6: 23 pairs of human chromosomes. This image is taken from : http://education-portal.com/academy/lesson/karyotype-definition-disorders-analysis.html#lesson

Coding regions in human genome

The human genome contains about 22,000 genes (Rosenbloom *et al.*, 2013). However, only the functions of half of them are known or could be inferred. The majority of human known genes have protein-coding function. Almost 25% of these genes are responsible for expression, replication and maintenance of the genome. About 17.5% of the known genes are coding for enzymes responsible for general biochemical functions. Another 20% of these genes are in a way involved in pathways that regulate cellular activities in response to signals from outside of the cell (Brown, 2002).

Other genomic region in the human genome

It is estimated that about 25% of the human genome that lies between genes (intergenic regions) have no known function. These regions are previously called *junk DNA* or *gene deserts* (Venter *et al.*, 2001). But recent studies have shown that some of these regions are carrying regulatory elements. These regions could not be eliminated without any phenotypic effects. This could indicate that such regions harbor elements with critically important conserved biological roles. The review by Ovcharenko *et al.* (2005) shows that there are two categories of gene desert : stable and variable. Stable gene deserts have lower repeat density compared to the gene-rich regions. This could suggest that these gene deserts are under a considerable degree of selective pressure (Ovcharenko *et al.*, 2005). Furthermore, throughout vertebrate evolution, these non-coding stable regions maintain their position and orientations (synteny) with respect to their neighboring genes. This could suggest the existence of an important linkage between these regions and their coding neighbors that could not be disturbed (Mongin, 2009).

1.1.4 Genome variation in human body

Most cells in a multi-cellular organism, such as human, are product of multiple divisions of one single cell (zygote). However, we know that cells in each individual come in variety of shapes and sizes. This is the result of differences in gene expression and/or regulation so that each cell type could be specialized for certain functions. Due to such specialization, different cell types have distinct sensitivity and level of exposure to internal/external chemical or physical signals. For instance, alcohol consumption on cells in gastrointestinal (GI) tract and liver (Pelucchi *et al.*, 2006), nicotine consumption on cells in respiratory tract, GI tract, tongue, kidneys and liver (Gandini *et al.*, 2008), UV rays on skin cells (de Gruijl

et al., 2001), nitrate in food preservatives on GI tract (Van Loon *et al.*, 1998), urinary tract (Tazima *et al.*, 1975) and liver cells (Van Loon *et al.*, 1998) and narcotics could affect nervous system, liver and kidneys cells (Rivière *et al.*, 2000) much easier than other cells in other tissues. Hence different cell types in different body organs could react to specific signal distinctively as well as independently. This creates a diverse genome variety in the same individual. Diseases such as cancers, which are not inheritable, could manifest in just one single cell in an individual who has no other cell with such genome variation.

1.1.5 The Human Genome Project

The Human Genome Project (HGP) is an international collaboration, which begun in 1990 and was aimed to determine the nucleotide sequence of the entire ~3.08 Gb letters of the human haploid nuclear genome (Rosenbloom *et al.*, 2013). Their goal is to provide researchers with powerful tools to understand the genetic factors in human diseases, which could help them to develop new strategies in their diagnosis, treatments and preventions. All information produced by the HGP are available in public databases. The HGP has already identified over 1800 genes related to different diseases. These data enabled researchers to develop more than 2000 tests for diseases and conditions caused by those genes. Such information could be used by health-care professionals in diagnosing the condition in early stages as well as in designing more efficient treatments (Institute, 2013).

1.1.6 Representation and storage of genomic data

Since Human Genome Project started, human genomic data became more and more abundant publicly. Biological databases store and maintain the different types of biological data around the world. Most of these databases represent their

data through interactive websites so they could be easily browsed, analyzed and retrieved. Today biological information comes in different formats. Some well-known biological databases are presented in Table 1.1. Among biological databases, UCSC

Database	Biological data	Size
GenBank[1]	DNA/RNA/Protein sequence	~189 Gbp ~182.2 M sequences
UniProt[2]	Protein sequence	59,744,893 entries
PDB browser[3]	Protein structure	108,957 structure
KEGG[4]	Genomic information	~101,33 M genomic info.
	Health information	14,249 health info.
UCSC Database[5]	Genomic information	3.2 Gig human nucleotides
	Multiple Sequence Alignment	~250 Gig
	Genome annotation	> 200

1. Benson *et al.* (2012)
2. Consortium *et al.* (2008)
3. Berman *et al.* (2002)
4. Kanehisa *et al.* (2014)
5. Karolchik *et al.* (2003)

Table 1.1: Example of biological databases.

Genome Browser Database is one of the most popular biological databases. It provides genomic sequence data, comparative data (Multiple Sequence Alignment), as well as graphical interface. UCSC genome browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UC Santa Cruz Genomics Institute and the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). The UCSC Genome Browser provides genomic information on a variety of organisms from yeast to higher mammalians. The information includes complete genome of 100 species, pairwise alignment of 78 species against human genome, full annotation data on 67 vertebrates, multiple alignment of 99 species against human genome, and more. The interactive site empowers users with a powerful visualization tool that allows them to visualize personalized information tracks on the human ge-

nome. It also provides users with tools such as *liftOver* (Hinrichs *et al.*, 2006) that converts genome coordinates and genome annotation files between assemblies, and *phyloGif*, which creates a gif image from the phylogenetic tree specification given. Moreover it provides users with the source codes for some tools, freely and downloadable for academic, noncommercial, and personal uses (Karolchik *et al.*, 2003).

1.2 Evolution

Evolution involves in the emergence and disappearance of traits and behaviors, in favor of species fitness to their dynamic environments. Evolution is the result of gradual processes occurring at the genome level, which modifies the genomic materials. Consequently, after a certain time, in two groups of the same species, different traits and behaviours would emerge. Eventually, they could be classified as two distinct species (Blanchette, 2001). Not all genome modifications could participate in the evolutionary process. Only those, that occur in the genome of germline cells could be passed to the next generation; and most importantly, are in the favour of the species survival and fitness. Only in such case the given modifications would be fixed into the genome.

1.2.1 Mechanisms that drive evolution

Genomic modifications could be induced by different exogenous as well as endogenous factors. Exogenous factors are those that cells receive from their environments such as environmental toxins, radiations and toxic chemicals. Endogenous factors, however, are factors that have no external source such as flaws in replication machinery of the cells and recombination. These modifications accompanied by environmental factors such as species migrations, competition over resources, climate change and diseases (natural selection) are the major forces driving evo-

lution (Brown, 2002). These genomic modifications could be in three forms :

1- Point mutations : correspond to single nucleotide insertions, deletions or substitutions by another nucleotide from one genomic sequence into another (Brown, 2002). Figure 1.7a illustrates the deletion of a $T{=}A$ base-pair in the DNA sequence on the left or an insertion of the base-pair to the sequence on the left. Also, the conversion of the sequence on the left to the one beneath by replacement of $T{=}A$ base-pair with a $C{\equiv}G$ is an example of substitution point mutation.

2- Small-scale mutations : are modifications that affect a small number of nucleotides such as deletion or insertion of a small DNA segment (Brown, 2002). Figure 1.7b shows the insertion or deletion of $TCACA$ between the two DNA sequences.

3- Large-scale genome rearrangements : are type of modifications that engage a large region on the chromosome and changes the genomic landscape. That could include translocation, inversion of DNA segment and fusion or fission of two DNA segments. Figure 1.7c shows one type of such modifications, which is the translocation of the region in *red* to somewhere along the *green chromosome*.

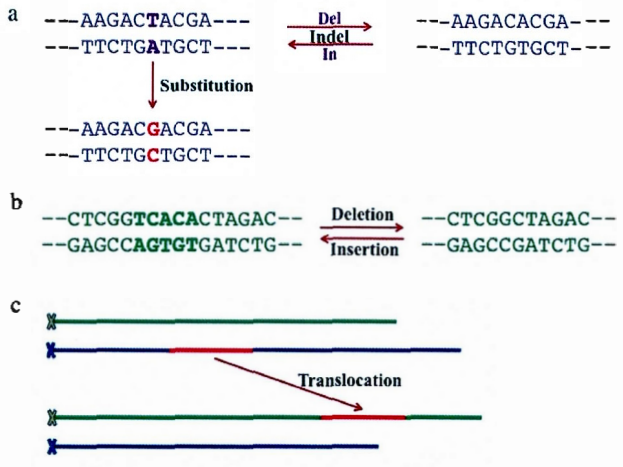


Figure 1.7: Mutation types : a) Point mutation b) Small-scale mutation c) Large-scale mutation or genome rearrangement

1.2.2 Example of evolutionary scenarios

The constant changing environment of all species along with the unstable nature of their genomes are the main forces that drive species diversity and evolution. For instance, as schematized in Figure 1.8, geographical separation of two groups of the same species (white population) in generation 0 could happen as a result of an earthquake. This event has also separated two individuals with two new variations (*red* and *green*). The *red* variant gives the individual the ability to produce and survive much easier than the others. If this population have enough resources, in just few generations they could overpopulate the down side of the valley. On the other side of the valley, the *green* variant just slightly boosted the reproduction ability. Hence, after about almost the same time as the *red* population, we could observe that the *green* and white populations are both occupying the region almost equally. By accumulating such different modifications in a group of the same species during a long time (high number of generation), the members of that species could not be classified as a single species.

1.2.3 Prediction of evolutionary scenarios

Evolutionary relationship between species usually could be revealed by digging their relationship at the genomic level which is called ***phylogenetics*** (Brown, 2002). Using comparative analysis of the genomes or ***comparative genomics***, phylogenetic study could unveil the evolutionary scenarios that the genomes of contemporary species have been subjected to since a common ancestor (Hardison, 2003). One of the basic assumptions of comparative genomics is the fact that a common phenotype in two given species are often encoded within the region that is conserved between those species since their common ancestor. Such regions mainly code or regulate functions which have positive effects on survival and re-

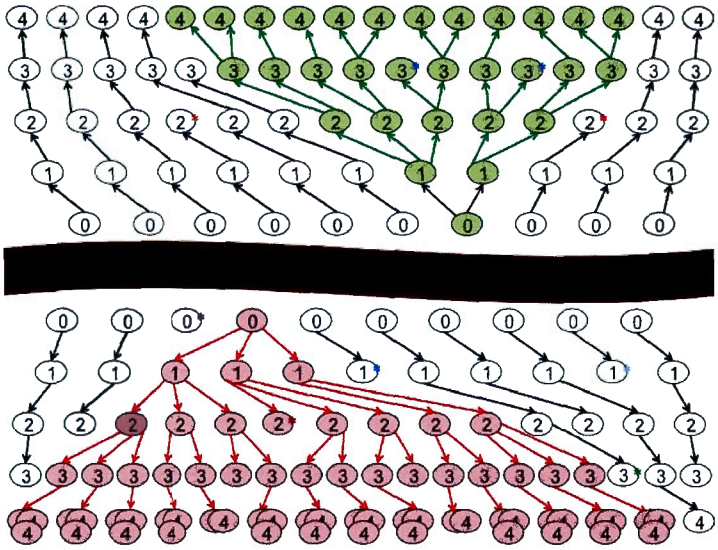


Figure 1.8: Example of an evolutionary scenario. Due to a geographical separation of two groups of the same species, two new variants (the *green* and the *red* variants) have been separated as well. As displayed in this figure, the *red* variant boosts the reproduction potential of the species therefore in just few generations they can overpopulate the down-side of the valley. On the other side the *green* variant affect slightly the reproduction of the species. At about the same time that species carrying the *red* caused their cosines go to the verge of extinction, species with the *green* variant have co-occupied the upside of the valley, almost equally with the original population. Therefore, studying individuals in the fourth generation, three different variation would be observed, which emanated from the original white variant.

production of those species (Hardison, 2003). Comparative genomics is a powerful approach for understanding the ancestral genome architecture and genomic rearrangements scenarios during evolution by examining three main characteristics of contemporary genomes (Horvath *et al.*, 2011) : a) DNA sequence conservation b) Genome function conservation c) Synteny conservation.

(a) DNA sequence conservation : One of the fundamental assumptions of comparative genomics is the fact that contemporary regions carrying highly similar (or identical) sequences might derive from a common ancestor irrespective of their evolutionary processes. To find the similarity between sequences, se-

quences have to be aligned together. Sequence alignments come in two formats : ***pairwise alignment***, which aligns two sequences together and finds the corresponding characters on those sequences, ***multiple alignment*** (Figure 1.9), which is the extension of pairwise alignment to more than two sequences. Aligning the genomic sequences is one of the core steps in phylogenetic analyses (Diallo, 2009). Several methods have been developed to identify highly conserved genomic regions in a given Multiple Sequence Alignment (MSA) such as Mauve (Darling *et al.*, 2004), PhyloP (Pollard *et al.*, 2010) and PhastCons (Siepel *et al.*, 2005). These methods assign a conservation score to each genomic region according to the number of nucleotide identities, synonymous and non-synonymous substitutions, etc...

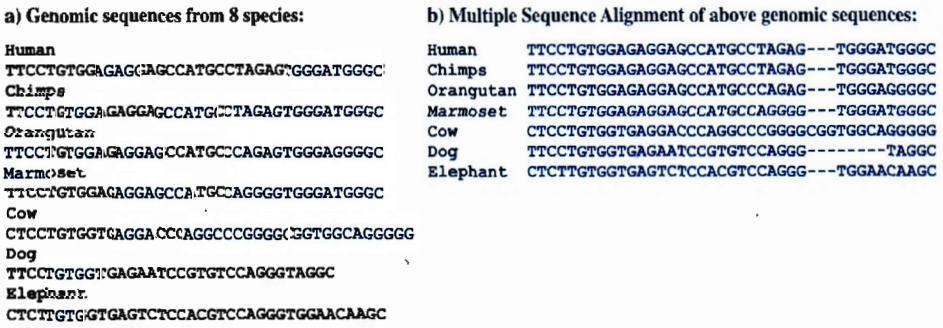


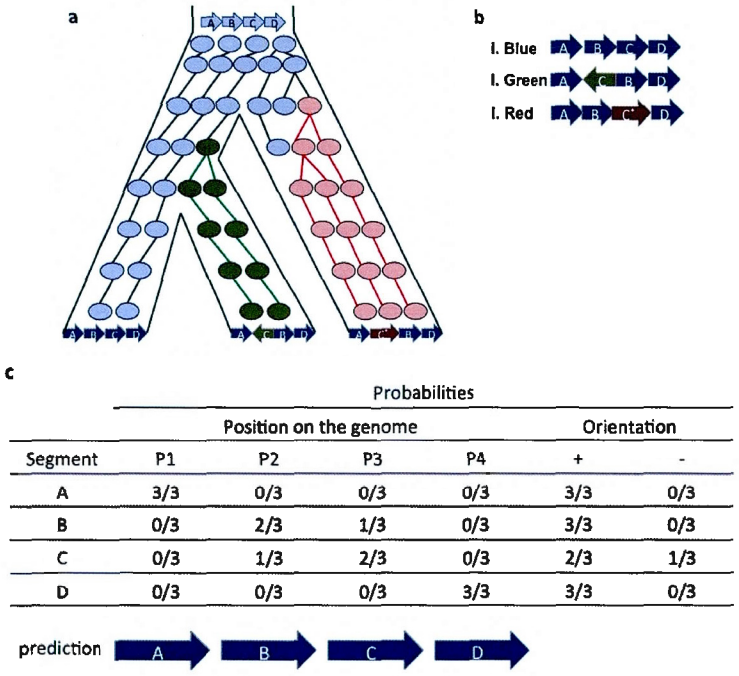
Figure 1.9: Multiple Sequence Alignment. a) Genomic sequences from 7 species.) Multiple Sequence Alignment of those sequences in *a*.

(b) Functional conservation of genomic regions : During evolution, it is likely that genes coding for functions that are essential for the survival of species conserve from their last common ancestor. Moreover, to maintain the integrity of the underlying functions of these genes, their regulating and controlling regions should be conserved in the same manner. In contrast, regions that encode or regulate proteins and RNAs responsible for species-specific traits might be different (Hardison, 2003). Lack of information on functional elements and non-coding conserved genomic regions is an obstacle to identify these regions.

However, selective pressure and fitness maintain the sequence conservation of functional regions. Thus, they undergo a slower rate of sequence change through time (Ganley and Kobayashi, 2008). Yet, predicting the exact function of those region remains a major challenge in computational biology (Hardison, 2003).

(c) Synteny conservation : Genomes of distinct species do not share the same architecture (gene or genomic segment organization). However, species sharing more evolutionary history tend to share several regions in common. Hence, despite many modifications in the genome sequence and conformation during evolution, there are highly conserved regions in terms of their order integrity and their positions across the genome. As defined by Nadeau and Taylor (1984), any uninterrupted chromosomal region that is occupied by two or more gene (genomic region) in two (or more species) are called ***"conserved segment"*** (Nadeau and Taylor, 1984) or ***"synteny block"*** (Pevzner and Tesler, 2003a). This rigidity to rearrangement during evolution has been often associated to functional constraints of genomic region. Studies on the genome synteny have shown that conserved regions are not only significantly enriched in putative regulatory regions (Mongin *et al.*, 2011; Kikuta *et al.*, 2007) but also are associated with transcriptional regulations and developmental processes (Mongin *et al.*, 2011; Sandelin *et al.*, 2004; Woolfe *et al.*, 2004). For instance, if we go back to the example of the evolutionary scenario explained in Figure 1.8 at the time of generation 4, the only data available is the three groups of species in both sides of the valley. To infer the evolutionary scenario that could explain the origin of the three similar species, we have to run a comparative analysis on their genome sequences. The multiple alignment of their genomes would highlight most of their genomic regions as conserved among the three species. The only region that shows variation among the three groups is a region with four homologous segments as shown

in Figure 1.10b. It shows that in all the species segments A and D have the same position, orientation and sequence homology. On the other hand, block C shows a sequence conservation in the first two species (*blue* and *green*) with two different orientations (loss of synteny conservation). Also the same block, C, in the species in *red* has the same orientation as species in white with less sequence similarity with the other two species. This shows that in $2/3$ of species block C is located between B and D and in $2/3$ the orientation of C is positive. The suit help us infer their ancestral genome architecture as the simplified demonstration in Figure 1.10.c.



	Probabilities					
	Position on the genome				Orientation	
Segment	P1	P2	P3	P4	+	-
A	3/3	0/3	0/3	0/3	3/3	0/3
B	0/3	2/3	1/3	0/3	3/3	0/3
C	0/3	1/3	2/3	0/3	2/3	1/3
D	0/3	0/3	0/3	3/3	3/3	0/3

Figure 1.10: Prediction of an evolutionary scenario : This simplified example shows the prediction process of an evolutionary scenario from contemporary genomes. a) Shows the real evolutionary scenario of a segment of the genome presented in three contemporary species. b) Shows the original conformation of the genome of these three species with respect to 4 homologous segments/regions/blocks/genes presented in all species. In all species A and D, are located on the same position. In 2/3 of species C is located between B and D. In 2/3 of species, segment C has a positive orientation (*blue and red species*). Segment C has the exact same sequence in *blue* and *green* species but shows less homology in the *red* species. The suit helps to infer the ancestral genome architecture as demonstrated in c.

Studies in comparative genomics are often based on a direct analysis of multiple sequence alignments and their underlying phylogenies (Siepel *et al.*, 2005; Darai-Ramqvist *et al.*, 2008; Ma, 2011). From a MSA, phylogenetic analysis would enable us to construct the evolutionary relationships, or genealogies among compared organisms. It also presents the historical course of their speciation through an arborescent format so-called phylogenetic tree (Wiley and Lieberman, 2011) as shown in Figure 1.11.

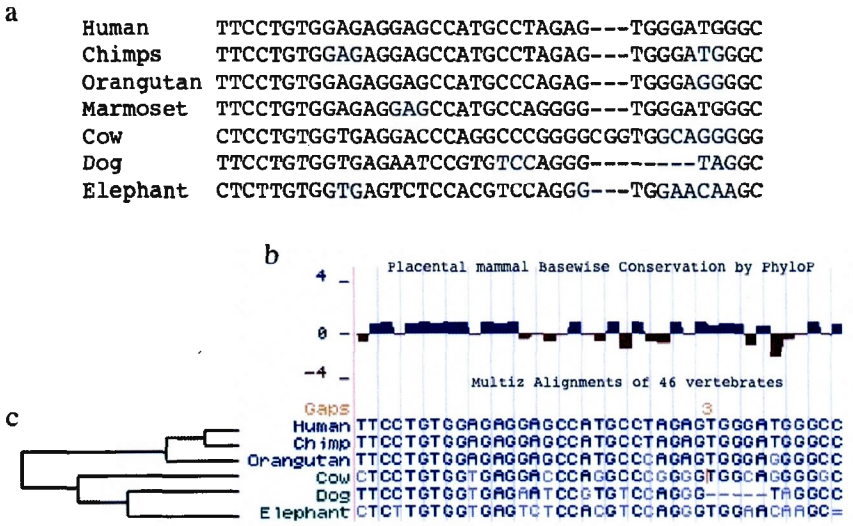


Figure 1.11: Comparative genomics. a) Multiple sequence alignment. b) MSA of a conserved region of 5 mammals genome sequences against the human genome and the corresponding conservation track from UCSC genome browser. c) Inferred phylogenetic tree based on the above alignment block.

Having phylogenetic data, we could trace back a specific phenotype and/or genotype to the point in time that it originated. This would be possible by using a mathematical algorithm in graph theory called the ***Lowest Common Ancestor (LCA)***. Given a rooted tree T, node $x \in T$ is an ancestor of node $y \in T$ if the path from x to the *root*, passes through x. Also, a node $v \in T$ is called to be a *common ancestor* of x and y if it is an ancestor of both. The Lowest Common Ancestor (LCA) of two nodes of x and y is a node whose distance to the x and y is shorter than any of their common ancestors in that tree. In any tree, the *root* is the common ancestor of all nodes (Moufatich, 2008). For example, in the tree in Figure 1.12, the LCA of the two nodes, *4* and *6*, is the node *1*.

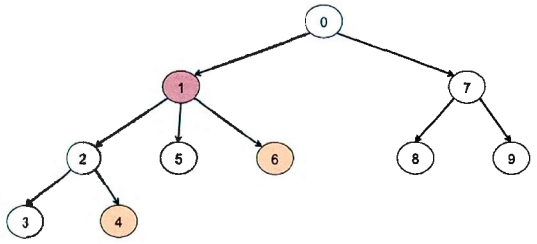


Figure 1.12: The Lowest Common Ancestor (LCA) of two nodes of **4** and **6** is a node whose distance to the **4** and **6** is shorter than any of their common ancestors in that tree. The ancestors of the nodes **4** and **6**, sorted by distance are 2, 1, 0 and 1, 0 respectively. Therefore the least common ancestor of these two nodes is the node ***1***.

1.2.4 Species tree vs. gene tree

Species tree is a phylogenetic tree constructed based on comparative analyses of species and their evolutionary relationship. The species tree represents the evolutionary pathways and processes that those species have gone through in general. In contrast, the gene tree is a phylogenetic tree that shows the evolutionary history of a gene or a genomic sequence in different genomes or within a single genome. It is now well-known that the gene tree does not always agree with the species tree (Brown, 2002). Processes such as duplication of a genomic region, Horizontal Gene Transfer, and gene loss could cause such a discordance between the two trees. For instance in Figure 1.13 a gene duplication has happened in time *T1*. Some time after the speciation, in each two new groups of species, one copy has been lost. And finally, around the time *T3* another duplication has occurred in the species group carrying the original gene copy. The phylogeny of the species highlights the evolutionary processes of their speciation due to accumulation of genomic modifications in time accompanied by environmental changes and natural selection. For example, the phylogeny of species in Figure 1.13 would be represented as in 1.14.a. Also, the evolutionary scenarios belonging to only one of

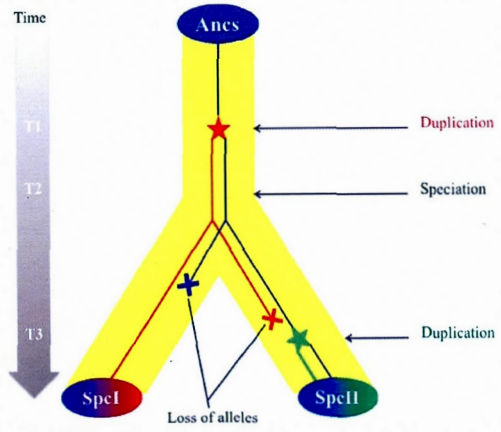


Figure 1.13: Example of an evolutionary scenario of a gene during the speciation of species carrying that gene. This figure shows that somewhere before the speciation a duplication has happened. Each new species received both copies of that gene. But farther away after the speciation, each species lost one of their copies. Going down in time, another duplication has happened in an ancestor of *SpcII*. The result of these processes was three homologous copies of this gene in contemporary species, one copy in *SpcI* and two copies in *SpcII*.

those modifications is depicted by the gene tree in Figure 1.14.a. One can easily recognize the disagreement between these two trees. This is a simple example of discordances between species tree and gene tree.

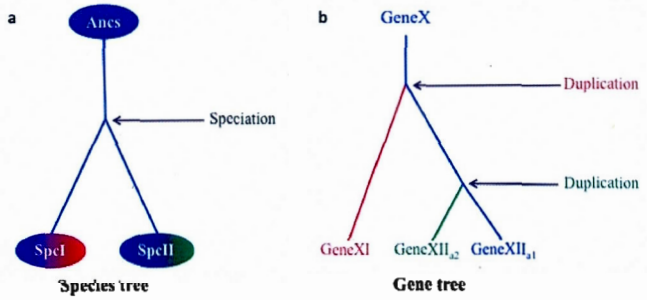


Figure 1.14: a and b, represent the species tree and gene tree, inferred by the comparative analysis of the contemporary species in figure 1.13, respectively. As demonstrated in this figure, the duplication of the ***GeneX*** happened before the speciation event. The second duplication of the same allele happened within the ***SpcII***, which is the source of the topological difference between the species tree and gene tree in this example.

1.3 Genome Rearrangements

Genome rearrangements are considered as evolutionary earthquakes and tend to dramatically change the genomic landscape (Peng *et al.*, 2006). Over time, new traits appear in favor of species fitness to their environments (Mongin, 2009). Genome rearrangements, such as deletion, duplication or translocation of a DNA segment, happen when double-stranded DNA breaks at two or more locations (breakpoints) and merge at different locations during DNA replication. This results in a distinct genome conformation from the original one (Blanchette, 2001) as shown in Figure 1.15. Previous studies have shown that such rearrangements do

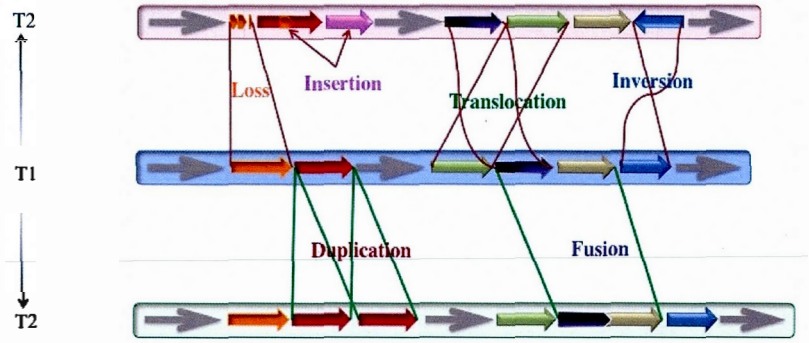


Figure 1.15: Common genome rearrangements : Rearrangements occurred passing from time T1 to T2 in fragile regions and have changed completely the original size and conformation of those regions. Gray arrows demonstrate regions resistant to rearrangements (refractory regions), which keep their synteny (order and content) during evolution.

not occur randomly across the genome. During genome evolution, certain regions have kept their synteny (***refractory regions*** (Mongin, 2009)), whereas others have been more subjected to rearrangement (***fragile regions***). These regions are not distributed randomly across genomes (Peng *et al.*, 2006; Lemaitre *et al.*, 2009; Pevzner and Tesler, 2003c). Any rearrangements in fragile regions could be just damaging without any effect on the survival or breading of the species, such as some subtelomeric rearrangements (Hengstschläger *et al.*, 2005). They could be

even in favor of fitness, such as rearrangements in immunoglobulin genes (Maizels, 2005). Genomic regions could also be very rigid (so-called breakpoint-refractory regions). Rearrangements in those regions are expected to be deleterious and could not be fixed in the genome (Mongin *et al.*, 2009). Such regions are mostly carrying developmental genes or regulatory elements responsible for their regulation (Mongin *et al.*, 2011; Sandelin *et al.*, 2004; Woolfe *et al.*, 2004). The heterogeneity in the distribution of synteny breaks across genomes could be due to less functional pressure on the different regions and thus less resistant to rearrangement (Mongin, 2009; Ciccarelli *et al.*, 2005).

1.3.1 Mechanisms of genomic rearrangements

Large-scale genomic rearrangements could be due to three major mechanisms :

— **Non-Allelic Homologous Recombination (NAHR) :** is a genomic recombination between products of a segmental duplication also known as Low Copy Repeat (LCR). It appears in cell division and is due to a misalignment in crossover between two non-allelic homologous regions, instead of two allelic regions. NAHR could lead to a deletion, inversion, duplication or translocation of DNA segments (Gu *et al.*, 2008). See Figure 1.16, a. This mechanism plays a major role in DNA repair and genome evolution by producing allelic variations (Gu *et al.*, 2008; Barış *et al.*, 2013a).

— **Non-homologous end joining (NHEJ) :** is a repair mechanism in DNA double strand breakage. This mechanism also known as nonhomologous recombination that necessitates little or no homology to join two free DNA ends together (Moore and Haber, 1996). See Figure 1.16, b. It generally causes variation of genetic materials (Barış *et al.*, 2013a). A study has shown that the majority of identified rearrangements in tumor genomes

were consistent with faulty NHEJ repairs (Raphael *et al.*, 2008).

— **Retrotranspositions or mobile element insertions (MEIs) :** are mobile genetic elements spread throughout genome by 'copy-and-paste' mechanism. In this process, DNA segments jump from one place to another and leave a copy of themselves as they move (Brown, 2002). It could alter the number of copy of that segment in the genome or Copy Number Variation (CNV). See Figure 1.16, c. The mechanism involved in MEI, is mediated by LINEs (long interspersed nucleotide elements), SINEs (short interspersed nucleotide elements) and retrovirus infections (Kazazian, 2004).

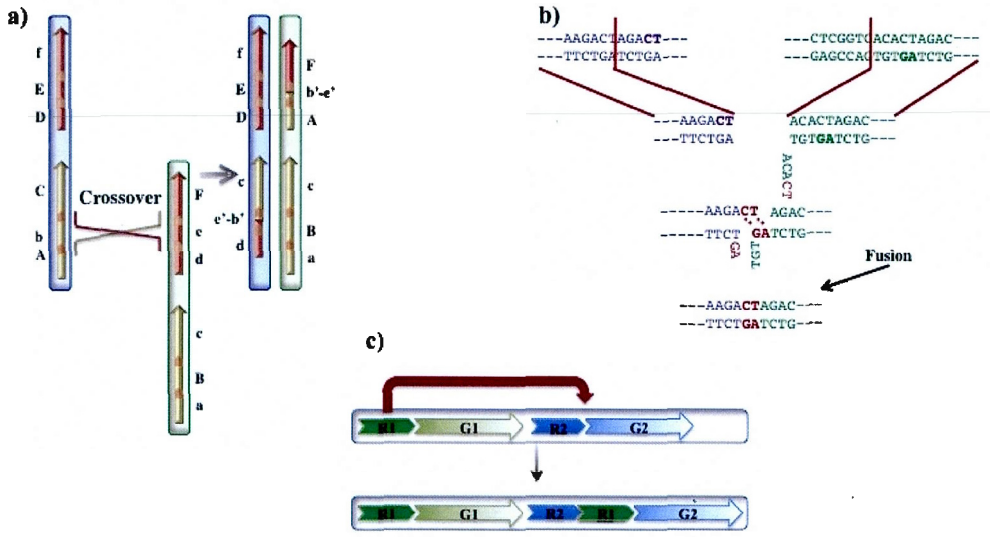


Figure 1.16: Genome rearrangement mechanisms. a) Non-allelic homologous recombination (NAHR) : A misalignment in crossover between two non-allelic homologous regions, instead of two allelic regions. b) Non-homologous end joining (NHEJ) : Double strand DNA breakage in two different chromosome regions and reassemble in new fashion. c) Retrotransposition or mobile element insertions (MEIs) : A transposable element, R1 that transposed between R2 and G2 and left a copy of itself as well.

1.3.2 Characteristic of genomic rearrangements

Studies have shown that synteny breaks are more frequent in regions carrying specific markers such as GC-rich (Lemaitre *et al.*, 2009; Darai-Ramqvist *et al.*, 2008), segmental duplication (Darai-Ramqvist *et al.*, 2008; Carbone *et al.*, 2009; Bailey *et al.*, 2003) and simple repeats (De and Michor, 2011; Carbone *et al.*, 2009). CpG islands are regions with lengths greater than 200 bp containing over 50% GC. They are more presented in or proximate to regulatory regions and are involved in gene regulation by obstructing the transcriptional factors (Larsen *et al.*, 1992; Wang and Leung, 2004). CpG methylations has shown to be more susceptible to rearrangements (Lemaitre *et al.*, 2009; De and Michor, 2011; Carbone *et al.*, 2009). DNA secondary structure such as G-quadruplex has also been shown to be mutagenic by obstructing the replication machinery (Kruisselbrink *et al.*, 2008; Pontier *et al.*, 2009; De and Michor, 2011). Regions enriched in arrangement sites essentially contain adaptive genes such as genes associated with inflammatory response and muscle contractility (Larkin *et al.*, 2009).

1.3.3 Genome rearrangements and evolution

Mutations and large-scale rearrangements empower genome with a particular dynamism by virtue of which species could cope with their constantly changing environment, survive natural selection and maintain their fitness. Genome rearrangements have mostly deleterious effects as they could cause loss or modification of traits that are crucial to the species survival and reproduction capacity. However, these modifications occasionally cause modifications that are in favor of species fitness and would undergo a positive selection and be fixed in the genome. This is the major force that drives species diversity and evolution (Brown, 2002; Watson, 2003; Mongin, 2009; Barış *et al.*, 2013b). Figure 1.17 shows a simplified

example of such rearrangements in genomic region in *red* in few generation. In this example, the red genomic region (the *red block*) is duplicated in one of the descendants of *A* (*B1*). Then in the next generation the duplicated segment is lost partially in the next generation (*C1*). At the same time the very same segment moves to another position (translocation) in *C2*. On the other hand, the other descendant of the *A* (*B2*), the *red block* stays conserved. Furthermore, in the next generation, this segment goes through an inversion in just one descendant of *B2* (*C3*). Such genome rearrangements in germ-line cell that pass to next generations would change the original genome conformation slowly such that two groups of the same species would diverge enough to be recognized as two close but different species (speciation processes) (Blanchette, 2001). Murphy *et al.* categorized these rearrangements into four different categories : 1) lineage-specific : are those rearrangements that found only in one species. 2) order-specific : are those that overlap between species of the same order. 3) super-ordinal : are those that happen in all representatives of a super-ordinal clade. 4) reuse : are the rearrangements that occur in the same region in species on different branches of species tree (Murphy *et al.*, 2005). Genome rearrangements could be recognized only by comparison of at least two different genomes (Sankoff, 2009). Regions that are more susceptible to synteny breaks (rearrangements) during evolution are so called "***rearrangement hotspots***" or "***fragile regions***" (Peng *et al.*; 2006). Figure 1.18 shows two different genomic regions on the human chromosome 1 from UCSC genome browser (Karolchik *et al.*, 2003) conservation track. Both of these regions are shared among 6 other mammals and chicken. As can be seen in this figure, these two regions show different levels of conservation. In the figure below the region, ***synteny region*** shows a very high conservation between all species even in chicken. Whereas the region above, a ***fragile region***, has a very weak conservation in all non-primate species.

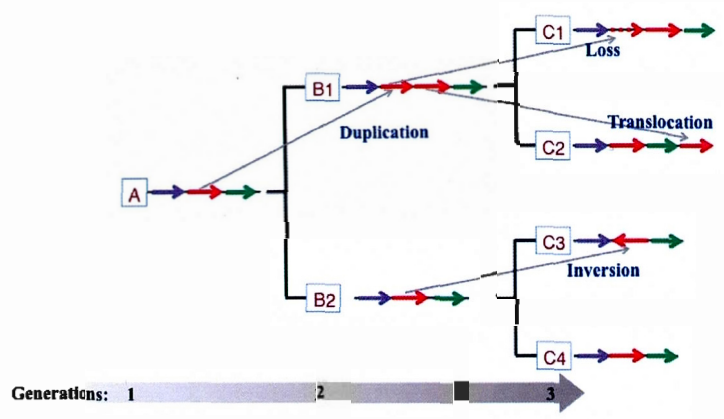


Figure 1.17: Genome rearrangements in the course of evolution. The *red block* rearranged in few generations and modified the original conformation of that specific region in different contemporary genomes.

1.3.4 Association of evolutionary rearrangements with genome functions

Genome rearrangements could have three kinds of effects on genome in terms of functionality :

i. Rearrangements which are not in favor of the species fitness to the environment are deleterious and will be lost in time (Blanchette, 2001). One good example could be the *1p36 deletion syndrome*. The syndrome is caused by a deletion in the short arm of chromosome 1. Some symptoms of this particular deletion include intellectual disability, distinctive facial features, and structural abnormalities in several body systems. This means that individuals carrying this rearrangement have much lesser chance to survive and reproduce.

ii. Rearrangements that could create new functions and lessen the selective constraints will be fixed in genome (Ciccarelli *et al.*, 2005). Such as DRD4 7-repeat allele, originated about 40,000 years ago showed a higher proportion in migratory populations (Chen *et al.*, 1999). This gene is also known as *novelty seeking gene*. Children diagnosed with Attention deficit hyperactivity disorder (ADHD) was shown to have a higher rate of this variation (Gören, 2014). This variation may

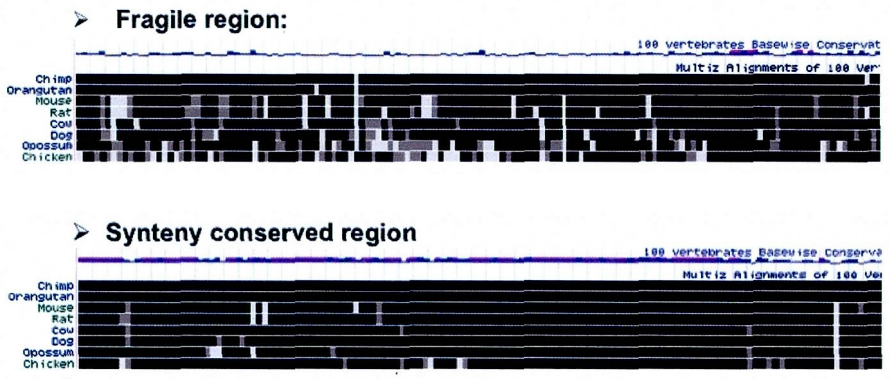


Figure 1.18: Example of conservation track on UCSC genome browser for two genomic regions of equal sizes on the human chromosome 1 with different conservation levels among 7 other vertebrates. Fragile region above shows week conservation between compared species whereas the synteny region below is almost conserved between all the species and shows a complete conservation with respect to primates.

not be favorable in today's complex societies but it was an advantage for those of our ancestors who left Africa 50,000 years ago.

iii. Rearrangements happening in non-functional regions could stay, change or be lost in time and could have no effect on individuals or on species fitness (Blanchette, 2001). Rearrangements in intronic regions are generally of those kinds.

1.3.5 Genome rearrangement and human diseases

Unlike evolutionary rearrangements that occur in germline cells, genomic rearrangements in somatic cells have an immediate effect on the very same individual and could not be passed to offspring. The effect of such rearrangements could vary from a complete loss of a DNA region to sometimes hundreds of copies of a DNA fragment (Stratton *et al.*, 2009). Somatic rearrangements could alter genes and gene regulation causing a variety of diseases and disorders in human. For instance, apposing a gene to the regulatory elements of another gene (Santoro *et al.*, 1996; Dalla-Favera *et al.*, 1982), or altering a protein-coding gene (result

into loss, gain or modification of a protein function) involved in cell growth and proliferation (Rowley, 2001) could cause an uncontrollable cell division and growth leading to cancers (Futreal *et al.*, 2004; Gollin, 2005). For example, a deletion involving the HBA1 and HBA2 genes located on chromosome 11 (11p15.5) causes *α-thalassemia.* Fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer is another example of such rearrangements (Tomlins *et al.*, 2005). Other than exogenous factors (e.g. nicotine, chemical hair dyes and UV exposure), sometimes DNA adopts non-B conformations. This destabilizes and damages the DNA. For example it has been shown previously that Guanine-rich genomic regions can adopt a four-stranded DNA structure (G-quadruplex/G4). See Figure 1.19. This structure plays a key role in genomic alterations observed in cancer genomes (De and Michor, 2011).

Figure 1.19: G-quadruplex structure on dsDNA. The image is taken and modified from : `http://biologicalexceptions.blogspot.ca/2013/10/dna-is-as-easy-as-b-z.html`

Genomic rearrangements in cancer

Cancers may be triggered by accumulation of mutations and genome rearrangements in somatic cells which alter cell division and growth. Somatic Copy Number Alterations (SCNA) are extremely common in cancer (Baudis, 2007; Stephens *et al.*, 2009; De and Michor, 2011; Zack *et al.*, 2013). SCNA are genome alterations that cause an abnormal number of copies of one or more DNA segments in somatic cells. These variations are frequent in cancer genomes. Many geno-

mic markers are known to be associated with cancer genome alterations, such as G4 structures by obstructing the movements of DNA polymerase (Sun and Hurley, 2010), CpG methylation (Behe and Felsenfeld, 1981; Vargason and Ho, 2002), and repeat elements (Hanahan and Weinberg, 2000). Moreover, a recent study has shown a significant presence of G4 structure proximate to translocation breakpoints in lymphoid genome (Katapadi *et al.*, 2012). Epigenetic factors such as modifications in DNA methylation and histone acetylation are other key role-players in human carcinogenesis (Kanai, 2010; Archer and Hodint, 1999; Feinberg and Tycko, 2004). The other phenomenon involved in rearrangements of cancer cells is injection of an alien DNA in genome through some viral infections such as Human Papilloma Virus (HPV), Epstein–Barr Virus (EBV) and Hepatitis B Virus (HBV) (Stratton *et al.*, 2009; Talbot and Crawford, 2004).

Cancer vs. evolutionary rearrangements

Genome instability and rearrangement mechanisms in both cancer associated and evolutionary rearrangements are driven by the same mechanisms. Somatic rearrangements have immediate effects in the individual whereas rearrangements in germline cells could pass to the next generations and participate in evolutionary processes. In a comparative analysis of the human genome with 6 non-primate species, performed by Murphy *et al.* (2005), evolutionary scenarios of rearrangements between all species and their ancestors have been reconstructed. 367 evolutionary breakpoints have been identified. Comparing these data with cancer-associated breakpoints has shown that distribution of the cancer-common chromosomal rearrangements are three times more frequent than those of the less common, proximal or within evolutionary breakpoints. Furthermore, the results showed a complete absence of cancer-associated breakpoints within the three longest synteny blocks in all species (Murphy *et al.*, 2005). Other studies reported

the colocalization of evolutionary fragile regions with tumor-associated deletions in human chromosome 3 (Kost-Alimova *et al.*, 2003; Darai *et al.*, 2005). In 2008, Darai-Ramqvist and his team conducted a comparative genomic analysis on three mammals, a primate and a lower vertebrate genome against human chromosome 3. They found out that tumor break-prone segmental duplications share sequence features with some genomic fragile regions. Other than physical proximity, they share CG content, presence of gene clusters associated with diversity and speciation, satellite repeats, transposable elements, and evolutionary history. They identified two tumor-related breakpoints on chromosome 3, presented distinguished tumor break-prone segmental duplications (TBSDs), which have also been involved in recent evolution of primates. It has been also noted that regions carrying TBSDs were broken more frequently during mammalian evolution than a random region on the same or other chromosomes (Darai-Ramqvist *et al.*, 2008).

1.3.6 Previously identified chromosomal fragility in human

Fragile regions (fragile sites) are cytogenetically defined as genomic regions that are more prone to break during the cell division (metaphase) causing partial inhibition of DNA synthesis. The chromosomal fragility is visible in metaphase chromosomes as gaps, breaks or poor staining in cell cultures under certain chemical stress (Durkin and Glover, 2007; Lukusa and Fryns, 2008; Mrasek *et al.*, 2010; Savelyeva and Brueckner, 2014). So far, over 200 fragile sites are identified on human genome (Mrasek *et al.*, 2010). Majority of these sites are common in all normal chromosomes in every individual. Such regions are called Common Fragile Sites (CFS). CFS are mostly induced by aphidicolin (DNA-polymerases α and δ specific inhibitor). The other CFS induce by synthetic nucleotides analogues, bromodeoxyuridine (thymidine analogue) and 5-azacytidin (cytidine analogue) (Lukusa and Fryns, 2008; Mrasek *et al.*, 2010). On the other hand rare fragile

sites are present in less than 5% of the populations. This characteristic of rare fragile sites make it easier to be identified using comparative analysis between individuals as they are less presented in the population (Savelyeva and Brueckner, 2014). Rare fragile sites are divided into two sub-groups : folate sensitive (induced by deficiency in folic acid and non-folate sensitive (rare fragile sites). About 20 fragile sites are identified on chromosome 1 that only one is classified as a rare fragile site (FRA1E/M at 1p21.3) (Lukusa and Fryns, 2008; Mrasek *et al.*, 2010).

CHAPTER II

HYPOTHESIS AND OBJECTIVES

2.1 Hypothesis

As previously mentioned in the last chapter, genome rearrangements do not occur randomly along the genome. Different genomic regions have distinct propensity to such rearrangements. This non-random distinct propensity of such rearrangements to different genomic regions could be explained by different impacts of such rearrangements on survival or reproduction of species due to their functional differences. The aforesaid affinity has been also observed in breakpoints associated with diseases, specifically cancers (Abeysinghe *et al.*, 2003; Moore *et al.*, 2006). For example, in one study, a complete absence of cancer-associated breakpoints have been observed in the three largest evolutionary conserved blocks (Murphy *et al.*, 2005). This could suggest that cancer-associated breakpoints may have a higher tendency to be localized in evolutionary fragile region. Also, in three other consecutive studies of the same group some co-localizations of evolutionary fragile regions with tumor-associated rearrangements in human chromosome 3 have been observed (Kost-Alimova *et al.*, 2003; Darai *et al.*, 2005; Darai-Ramqvist *et al.*, 2008). Furthermore, it has been previously hypothesized that evolutionary fragile region could be more prone to cancer-associated rearrangements (Mongin, 2009). Based on these observations as well as the fact that both rearrangement catego-

ries (evolutionary and disease-associated rearrangements) operate via similar mechanisms, we hypothesize that disease-related rearrangements, cancer-associated rearrangements in particular, have more affinity to the evolutionary fragile regions.

2.2 Goal

To test this hypothesis, the following three major steps are necessary :

1. Identification of genomic regions that are more susceptible to rearrangements in the course of the human evolution.
2. Identification of diseases that are associated with genomic rearrangements as well as the position of such rearrangements along the human genome (evolutionary fragile region).
3. Mapping these two genomic region categories together to see whether there is any correlation between genomic regions that carry these two types of genome rearrangements.

By performing these steps the following question could be answered : Are the evolutionary fragile regions more prone to disease-related rearrangements ? This would permit us to identify genomic regions that are more susceptible to disease-related, specially cancer, genome rearrangements.

2.3 Objective

Due to the fact that each previously mentioned steps constitute a challenge and a major problematic, during this thesis we have focused on the first step. This would provide consistent evolutionary fragile region, so that the next two steps could be performed with high accuracy in my following PhD thesis to achieve the final goal of this project to highlight genomic regions that are more susceptible to disease-related rearrangements.

As explained before, ***evolutionary fragile region*** or ***rearrangement hotspots*** are regions that are more likely to be rearranged. These regions should be enriched in breakpoints (regions bounded by two consecutive synteny regions (Lemaitre *et al.*, 2008) or rearrangement positions)). To pinpoint these regions, breakpoints should be identified along the genome initially. Having identified the breakpoints along the genome, we could identify the fragile regions in which those breakpoints are significantly more accumulated. In order to reach this objective, the following steps are necessary :

1. Identification of evolutionary synteny breaks on the human genome.
2. Identification of evolutionary fragile regions specific to human lineage.

2.3.1 Identification of evolutionary synteny breaks on the human genome

To identify synteny breaks, one should conduct a comparative analysis to identify the synteny along the genome first. Consequently, by excluding these highly conserved synteny regions, the genomic regions that could not maintain their synteny and have been subjected to rearrangements all along would be disclosed (Lemaitre *et al.*, 2008). To perform a comparative analysis on the human genome, the most appropriate candidates are of two kinds : 1) neighboring species that share really common features, and 2) more or less distant species that could have a broad divergence with the human. This helps to capture more evolutionary patterns in the analysis.

2.3.2 Identification of evolutionary fragile regions specific to the human lineage

Having identified synteny breaks on the human genome, we want to identify the genome "rearrangement hotspots" or "fragile regions". Thus, we will scan the ge-

nome for synteny breaks and identify genomic regions that represent significantly more rearrangements using statistical methods. At the same time, we are going to study the association of these regions with markers that have been previously identified to be associated with genome conservation and fragility.

CHAPTER III

METHODOLOGY

This chapter presents the required steps to achieve the objective explained in 2.3. First, a comparative analysis (with several vertebrates) is performed to identify the synteny region on the human genome and where those regions have been broken. In each comparative analysis, the first step is to sample species and collect their genomic information. The genome sequences should be aligned to reconstruct their evolutionary history and identify the genomic regions that kept their synteny in the course of evolution and the regions bounded by those as synteny breaks (breakpoints). Next, with statistical methods we will identify the evolutionary fragile regions on the human genome where human-specific breakpoints are more frequent. Figure 3.1 shows the complete pipeline to achieve this thesis objective. It is important to mention here that the developed procedure has been limited to human chromosome 1 for this thesis. This is due to the time consuming computation to cover the whole genome. However, when the pipeline and results have been assessed for this benchmark, it will be easily extended to the genome.

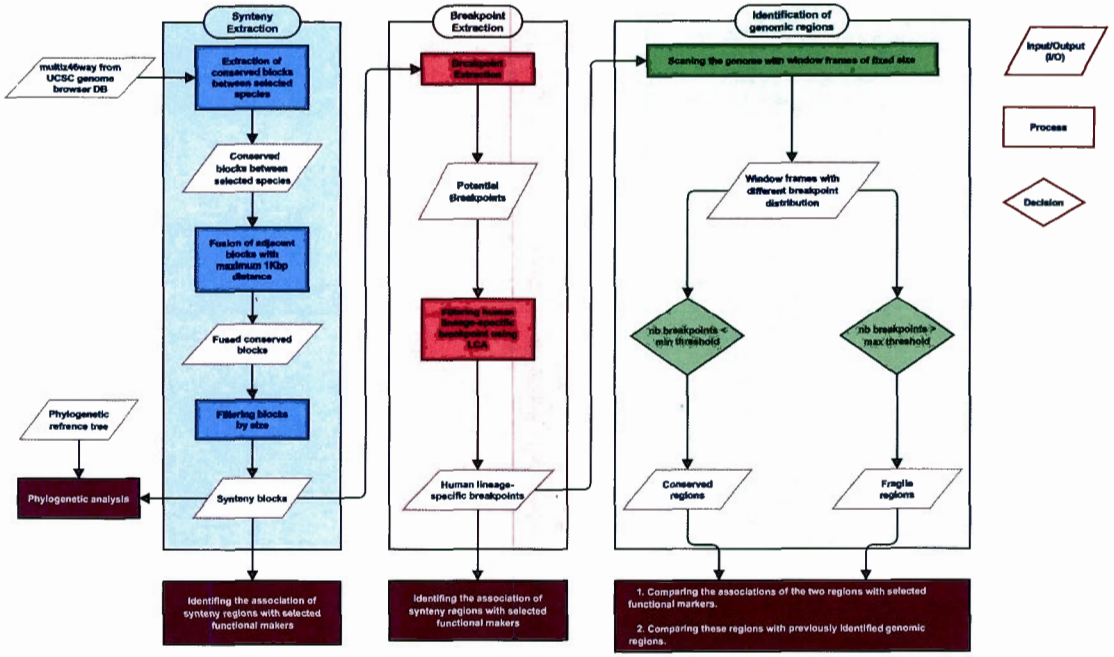


Figure 3.1: This diagram explain major steps of our pipeline. As separated in this image, this pipeline is divided into three major following groups : 1) Extraction of synteny blocks : First the conserved blocks between selected species have been extracted from multiple sequence alignment provided by UCSC genome browser ; Adjacent blocks have been fused together ; and synteny blocks were identified by filtering the fused block by their size. A phylogenetic analysis have been conducted on these blocks and their associations with selected markers have been identified as well. 2) Extraction of breakpoints : Region between each two consecutive synteny blocks (the results from the previous step) were identified as potential breaks ; Human lineage-specific breakpoints were extracted from these breaks using LCA. The association of these breakpoints with the selected markers were identified. 3) Identification of fragile region : The chromosome have been divided into window-frames of fixed size and the number of breaks in each window has been counted. Using a threshold on calculated z-score for the number of breaks, we identified fragile regions as windows with a z-score higher than the defined threshold and conserved regions as those with a z-score lower than the defined threshold. The association of identified regions with functional markers were identified and these regions were compared with previously identified fragile or conserved regions to evaluate our method.

3.1 Identification of evolutionary synteny breaks on human genome

3.1.1 Species sampling

As explained in 2.3.1 there are two kinds of appropriate species to conduct a comparative analysis. As a member of the primate family, the best candidates neighboring species are mammalians including primates. It should be noted that due to the high similarities between genomes of primates and lack of enough evolutionary signals, it is not easy to detect genomic rearrangements specific to human. However, the high similarity between the genomes of primates could lower the uncertainty caused by missing information. Thus, three well-studied primates have been chosen for this study. Also from each major branches of mammalian tree of life, two well-sequenced and well-studied species have been selected. Having selected two species from each major branch of mammalian tree of life, will help to handle the lack of information in each branch. If, for example, one species misses some genomic information, the information of its close neighbor could compensate that. Moreover, for those of more distant species, one placental mammal as well as one non-mammalian vertebrate have been chosen to perform a comparative analysis between those species and human. These species are as follows : Chimpanzee, Orangutan and Marmoset (Primates), Rat and Mouse (Rodents), Dog and Cow (Laurasiatheria), Elephant and Armadillo (non-Boreotheria), Opossum (Marsupial) and Chicken (non-mammalian vertebrate). See Figure 3.2.

3.1.2 Genome sequences information

The genomic information on the selected species have been already produced and aligned by ENCODE project (Consortium *et al.*, 2004a). This multiple alignment is an alignment of 45 vertebrate genomes with human stored in MAF file

format and is available for download at : `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/maf`. An example of MAF file format is shown in Figure 3.3. Alignments are organized in blocks. Each block includes information of the name of the source for the sequence, the chromosome, the start position of each sequence in the source sequence, the size and the orientation of the sequence, the size of the source of the entire sequence, and the aligned sequences. To each alignment block, a conservation score is attributed. See table 3.1 for assemblies information describing the details of the sequenced genomes.

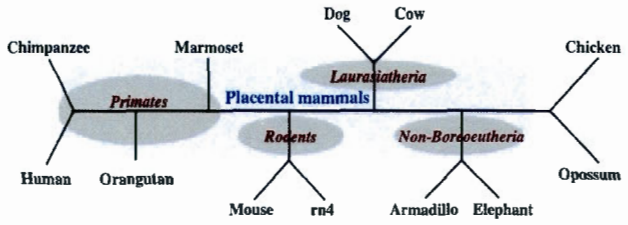


Figure 3.2: Extracted species tree for the 12 selected species

```
a score=56851.000000
s hg19.chr1              12072  28 + 249250621 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s calJac1.Contig8673     79080  28 +    105741 TTCCTGTGGAGAGGAGCCATGCCAGGGG
s canFam2.chr27       45129600  28 +  48908698 TTCCTGTGGTGAGAATCCGTGTCCAGGG
s equCab2.Contig2343     51245  28 +     71245 CTCCTGTGGTGACGACCCAGGCCCGGGG
s bosTau4.chr5       113865000  28 + 125847759 CTCCTGTGGTGAGGACCCAGGCCCGGGG
s loxAfr3.scaffold_15  8810173  27 -  55688157 CTCTTGTGGTGAGT-TCCACGTCCAGGG
s panTro2.chr15          14681  28 - 100063422 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s ponAbe2.chr2b       21132696  28 + 135000294 TTCCTGTGGAGAGGAGCCATGCCCAGAG
```

Figure 3.3: Each alignment block begins with a line that starts with 'a' which stores information for the entire block such as conservation. Lines starting with 's' store information on sequence within an alignment. Each 's' line has the following information : src) the name of the source sequences for the alignment. For sequences that are resident in a browser assembly, the form 'database.chromosome' allows automatic creation of links to other assemblies. start) The start of the aligning region in the source sequence. This is a zero-based number. size) The size of the aligning region in the source sequence. This number is equal to the number of non-dash characters in the alignment text field. strand) If '-', then the alignment is to the reverse-complemented source. srcSize) The size of the entire source sequence, not just the parts involved in the alignment. text) The nucleotides in the alignment. All information in this caption is taken from `http://genome.ucsc.edu/FAQ/FAQformat.html`

vertical-alignment

Name	Scientific Name	Date	Size (bp)	Total assembly gaps/
Human	Homo sapiens	Feb. 2009 hg19/GRCh37	3,137,144,693	239,845,127
Marmoset	Callithrix jacchus	June 2007 calJac1/Callithrix jacchus-3.2	2,914,958,544	162,452,744
Dog	Canis lupus familiaris	May 2005 canFam2/CanFam2.0	2,528,446,953	143,450,410
Armadillo	Dasypus novemcinctus	Sep. 2008 dasNav2/Dasnov2.0	4,813,823,562	2,442,329,690
Chicken	Gallus gallus	May 2006 galGal3/Gallus_gallus-2.1	1,098,770,941	56,179,590
Cow	Bos taurus	Oct. 2007 bosTau4/Btau_4.0	2,917,945,987	186,142,499
Elephant	Loxodonta africana	Jul. 2009 loxAfr3/Loxafr3.0	3,196,738,102	78,195,493
Mouse	Mus musculus	Jul. 2007 mm9/MGSCv37	2,745,142,291	96,619,540
Opossum	Monodelphis domestica	Oct. 2006 monDom5/MonDom5	3,598,443,077	98,827,259
Chimpanzee	Pan troglodytes	Mar. 2006 panTro2/Pan_troglodytes-2.1	3,349,648,539	440,199,864
Orangutan	Pongo abelii	Jul. 2007 ponAbe2/P_pygmaeus_2.0.2	3,441,227,734	347,678,322
Rat	Rattus norvegicus	Nov. 2004 rn4/RGSC_v3.4	2,826,224,306	254,484,453

Table 3.1: Information on genome assembly of the chosen species

3.1.3 Extraction of Multiple Sequence Alignment

From `hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/` we can download the multiz46way vertebrate alignment of human chromosomes. To avoid the effect of missing data and ambiguities mostly found in Marmoset, Elephant and Armadillo genomic data, and phylogenetic distance between chicken and other mammals we decided to accept all alignment blocks within which at least 7 selected species were presented as a conserved block among the 12 species (See Algorithm 3.1). Information on all other species has been removed from the alignments blocks. All the columns in the alignments that only consisted of gaps have been removed as well. An example is shown in Figure 3.4. In this Figure the above original alignment MAF file format block carries 7 species out of 12 selected species. This block also includes sequence information of horse (*equCab2.Contig2343*) that we didn't choose for our study. So when we extracted this block as an accepted block for our analysis, we removed the information on species other than the selected 12, in this case the horse. Then we added the missing species in the block. These added information are colored in blue in this Figure. All the extraction steps have carried out using *Perl scripts* programming language.

Algorithm 3.1 Algorithms to extract alignment blocks having at least 7 selected species presented from multiz46way.

```
for each block do
  if nbSelectedSpc(block) >= 7 then
    block ←removeUnSelectedSpc(block)
    block ←addMissingSpecies(block , listMissingSpecies)
    writeInMaf(block)
```

3.1.4 Identification of synteny blocks

To identify synteny blocks on human chromosome from the MSA of the 12 selected species, we first fused all the neighboring contiguous blocks. Two neighboring

blocks are considered to be contiguous if for all species in those blocks, both sequences are contiguous (Pevzner and Tesler, 2003b; Murphy *et al.*, 2005). See Algorithm 3.2.

a) MAF block presented by UCSC multiz46way

```
a score=56851.000000
s hg19.chr1              12072  28 + 249250621 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s calJac1.Contig8673     79080  28 +    105741 TTCCTGTGGAGAGGAGCCATGCCAGGGG
s canFam2.chr27       45129600  28 +  48908698 TTCCTGTGGTGAGAATCCGTGTCCAGGG
s equCab2.Contig2343     51245  28 +     71245 CTCCTGTGGTGACGACCCAGGCCCGGGG
s bosTau4.chr5       113865000  28 + 125847759 CTCCTGTGGTGAGGACCCAGGCCCGGGG
s loxAfr3.scaffold_15  8810173  27 -  55688157 CTCTTGTGGTGAGT-TCCACGTCCAGGG
s panTro2.chr15          14681  28 - 100063422 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s ponAbe2.chr2b       21132696  28 + 135000294 TTCCTGTGGAGAGGAGCCATGCCCAGAG
```

b) Modified MAF block

```
a
s hg19.chr1              12072  28 + 249250621 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s calJac1.Contig8673     79080  28 +    105741 TTCCTGTGGAGAGGAGCCATGCCAGGGG
s canFam2.chr27       45129600  28 +  48908698 TTCCTGTGGTGAGAATCCGTGTCCAGGG
s dasNov2.Un                0   0 +         0 ----------------------------
s galGal3.Un                0   0 +         0 ----------------------------
s bosTau4.chr5       113865000  28 + 125847759 CTCCTGTGGTGAGGACCCAGGCCCGGGG
s loxAfr3.scaffold_15  8810173  27 -  55688157 CTCTTGTGGTGAGT-TCCACGTCCAGGG
s mm9.Un                    0   0 +         0 ----------------------------
s monDom5.Un                0   0 +         0 ----------------------------
s panTro2.chr15          14681  28 - 100063422 TTCCTGTGGAGAGGAGCCATGCCTAGAG
s ponAbe2.chr2b       21132696  28 + 135000294 TTCCTGTGGAGAGGAGCCATGCCCAGAG
s rn4.Un                    0   0 +         0 ----------------------------
```

Figure 3.4: a) Example of an alignment block in maf format from UCSC multiz46way that includes >= 7 selected species. The line in *red* carries information on a species not selected for our analysis (horse). b) We modified the eligible block by removing all unselected species and adding missing species with default information. Added species are presented in *blue*.

Algorithm 3.2 Algorithms to fuse all adjacent blocks to identify synteny blocks.

```
block_I ←removeFirstBlock(listOfBlock)
for each block do
  if listOfBlock.size == 1 then
    listOfFusedBlocks.push(block_I)
  else
    block_II ←removeFirstBlock(listOfBlock)
    if (!areContiguous(block_I, block_II) then
      listOfFusedBlocks.push(block_I)
      block_I ← block_II
    else
      for each species in (block_I) do
        species.size← species.size + block_II.species.size
        species.seq←concatenate(species.seq,block_II.species.seq)
```

Then, we applied the same algorithm with a maximum distance of 1 Kbp for all species in the block. These blocks will then be filtered by length. This means that all blocks having less than 1,000 nucleotides, with respect to the human sequence, would be excluded from the study. The remaining blocks would be considered as synteny blocks. With the 1,000 nucleotides size rule we are able to observe genomic regions among their neighboring regions and to see if they are conserved with respect to their locations among other regions between selected species. This process in schematized in the example of Figure 3.5. In this example, fusion procedure has been applied to the four alignment blocks on the left. Moreover, other than for the last two blocks, *Blc_III* and *Blc_IV*, we can not fuse the rest of neighboring blocks. *Blc_I* and *Blc_II* couldn't been fused due to the separation of those blocks by more than 1,000 nucleotides in both genomes of *Cow* and *Dog*. The same force inhibited the fusion of *Blc_II* and *Blc_III* but this time on the mouse genome. Applying the same method explained above, we fused all the conserved blocks to extract and identify our synteny blocks for our analysis. As we mentioned before, all blocks having a size less than 1,000 nucleotides (1 Kbp), with respect to human sequence, have been filtered as well.

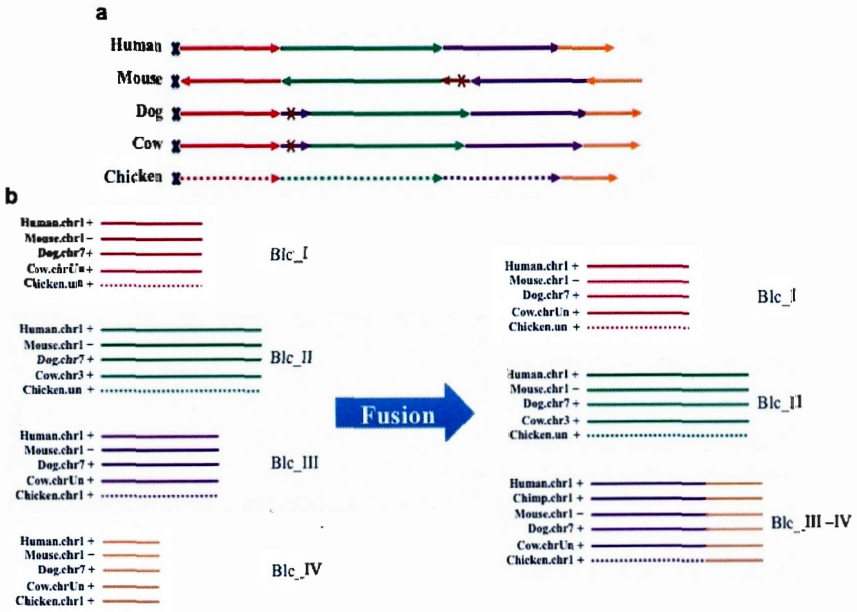
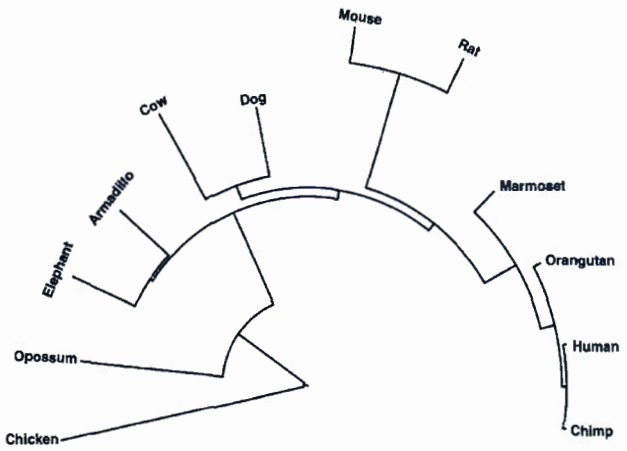


Figure 3.5: Synteny block extraction : a) This figure shows the original architecture of each species genome. Each color represents a conserved genomic region presented in most of the compared genomes. $\boldsymbol{X}$s are used to label the genomic regions that are not common between those species with a size $>$ 1 Kbp. Each arrow represent the orientation of the blocks with respect to the reference genome, which is, in this case, the human genome. Figure b, on the left represents each conserved region among these species in the form of an alignment block. In the fusion steps, a gap of $>$ 1 Kbp between *Blc_I* and *Blc_II* in both *Cow* and *Dog* genome, prevents us from fusing these two alignment blocks. The same phenomenon in the genome of the *Mouse* between *Blc_2* and *Blc_III* restricts the fusion between those two blocks. On the contrary the continuity between purple and yellow blocks (the last two blocks) is not disturbed in any of the species so that we could fuse these to blocks to a single conserved block shown in Figure b, on the right.

3.1.5 Phylogenetic analysis

To study the evolutionary history of genomic conserved regions, we need to know the evolutionary relationship between those species (their phylogenetic history). It is important to have accurate phylogenies, since the breakpoint prediction algo-

rithm will rely on those phylogenies. As we explained in 1.2.3, species phylogenetic tree were constructed based on these evolutionary relationships. The MSA that we used for this study is also constructed based on the same phylogeny. Thus, to understand the evolutionary nature of extracted genomic regions from this MSA we need the phylogenetic tree of the selected species. To obtain this tree, we have downloaded the species phylogenetic tree that our multiple sequence alignment of 46 species was constructed on, from `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/multiz46way/46way.nh` in Newick format. We then altered this tree by removing all the species that are not included in this study using *retree* program of PHYLIP package (Felsenstein, 2005). This phylogenetic tree (see Figure 3.6) for the 12 selected species is used as our reference tree for our phylogenetic analysis. To infer the evolutionary history of the extracted synteny blocks and their associated potential breakpoint regions, we used ***MrBayes*** program (Huelsenbeck *et al.*, 2001), based on a Bayesian approach, which is one of the main robust approaches in phylogenetic analyses. It is important to have accurate phylogenies, since the breakpoint prediction algorithm will rely on those trees. Robinson and Foulds (RF) topological distance algorithm (Robinson and Foulds, 1981) was used to compare inferred phylogenetic trees with the species tree. Robinson and Foulds (RF) topological distance is a metric to compute the number of splitting and merging edges needed to convert one tree topology into another. This distance metric is well-known and has been used in several studies (Swofford, 1998; Kumar *et al.*, 2004) (see Figure 3.7). In this example the Robinson and Foulds (RF) topological distance between these two trees is equal to 2 as the tree on the left needs two modifications to be converted to the one on the right. With the same method, we calculated the topological distances between all inferred phylogenetic tree of extracted synteny blocks with the tree of species. These topological distances were calculated using a *C++ script* produced in our laboratory.



Phylogenetic tree for 12 selected species in Newick format:

```
(((((((Human:0.00642,Chimpanzee:0.00636):0.01154,Orangutan:0.0185):
0.03537,Marmoset:0.06828):0.08797,(Mouse:0.08571,Rat:0.09147):0.28743):
0.02117,(Cow:0.23598,Dog:0.17093):0.03546):0.02249,(Elephant:
0.16964,Armadillo:0.16054):0.0073):0.24458,Opossum:0.34891):
0.20202,Chicken:0.60993);
```

Figure 3.6: Phylogenetic tree for 12 selected species (Karolchik *et al.*, 2003) in Newick format. Next we excluded all the species from the well-accepted phylogenetic tree of species from UCSC genome browser, using *retree* application of **PHYLIP** package (Felsenstein, 2005). Figure above shows the final tree for the 12 species in Newick format. This tree was then converted to this circular tree format using **iTol** (Letunic and Bork, 2007).

3.1.6 Identification of synteny breaks (breakpoints)

To identify synteny breaks or breakpoints on the human genome, we consider regions between each two neighboring synteny blocks on the human genome as a potential breakpoint, such that the size of the region is lower than 1 Mb. This constraint is necessary to avoid ambiguous regions that could not be associated correctly to breakpoints (e.g. sequences of heterochromatin). Heterochromatic sequences have high levels of repetitive elements, which make it difficult to assemble (Hoskins *et al.*, 2002) such as centromeres and telomeres. In the next step, the

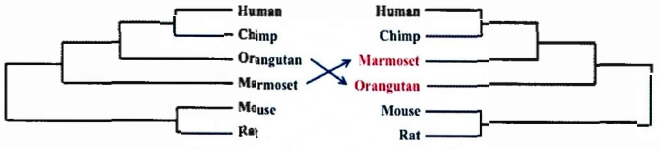


Figure 3.7: The RF topological distance : The number of modifications needed to convert the tree topology on the left to the one on the right two operations are necessary. This means that the RF topological distance between the two topologies is equal to 2.

fusion procedure is applied one more time to the extracted blocks to document all the forces that prevent us to fuse each two neighboring blocks. Simultaneously, all the forces that prevented to fuse two neighboring blocks have been documented as a potential break. The process has been presented in Figure 3.8. In this example, each accolade represents a synteny block and distances between each two synteny blocks. The potential breakpoints are marked in red. The two ends of the chromosome are considered as breaks in all species (*Brk_1* and *Brk_2*). *Brk_2* has occurred due to a distance between the first two synteny blocks with respect to the chicken genome. *Brk_3* is detected by a change in orientation in the purple (fifth) block with respect to the genome of *Armadillo*. And so on and so forth, all the breakpoints would be identified.

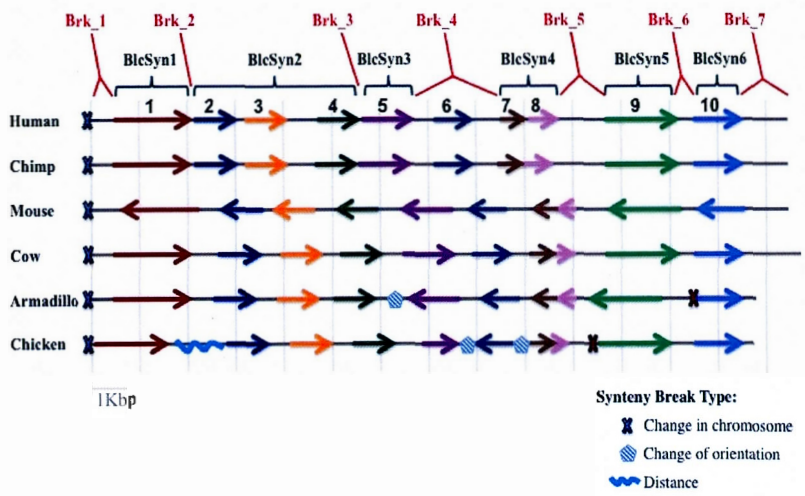


Figure 3.8: Extraction of synteny blocks and their corresponding breaks : Each arrow represents a conserved region on each genome. Each color specify same genomic conserved region in all species. Directions in each arrow represent the orientation of that region in each species. The distance between each two vertical lines represents 1 Kbp. Each X shows the start of a chromosome. As shown in this example, blocks 2, 3, and 4 as well as block 7 and 8 have been fused together in the fusion step as there was a distance less than 1 Kbp between each two blocks in all species. After the fusion step, block 6 has been eliminated from our study due to its size ($<$ 1 Kbp). Forces that prevented the fusion between each two synteny blocks have been documented as a potential breakpoint.

3.1.7 Identification of breakpoints specific to human lineage

Having documented all the breaks in different selected species between each two neighboring blocks permitted us to track down the origin of genome rearrangements of contemporary genomes on the species tree. We have applied the *get_LCA* method of **BioPerl** (Stajich *et al.*, 2002) to find the Lowest Common Ancestor (LCA) for each set of rearrangements shared among two or more species, on the species tree. This has enabled us to identify and eliminate the breakpoints, which are probably not present in any of human ancestors. As presented in Figure 3.9.a, a breakpoint observed in rat and mouse could be traced back to the common

ancestor of the two species carrying the same break. In this case, the common ancestor of the break is probably a rodent living some where after the separation of rodents from their common ancestors with primates, as no such synteny break is observed neither in human nor in chimpanzee. So the break in this Figure would not be considered as a human lineage-specific breakpoint. On the other hand the common ancestor of the species carrying the break in 3.9b, could be traced back to a human ancestor before the separation of the primates and rodents. Since it has originated from a human ancestor, this break will be considered as human lineage specific.

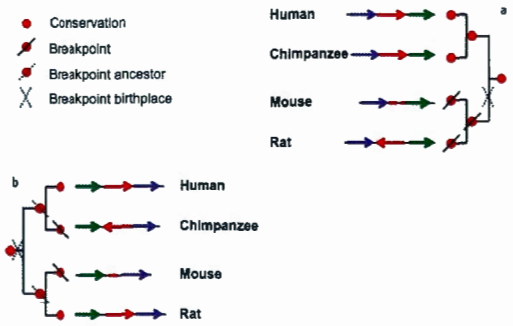
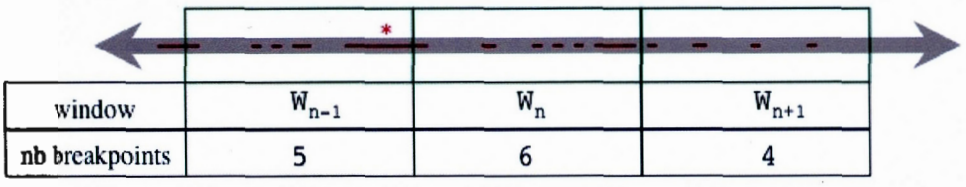


Figure 3.9: The figure shows how LCA algorithm can predict the common ancestor of the rearranged red block. a) The block in red have been rearranged in all rodent so the possible birth place of this synteny break is some time after separation of the rodent from primates and it does not concern the human lineage. b) The block in red is rearranged in rat and chimp. So it should originate from one of the common ancestors of primates and rodents. So the Lowest Common Ancestor of this break is the human ancestor *Euarchontoglires*.

3.2 Identification of fragile regions

To identify genomic fragile region based on the comparative analysis carried out before, the first step is to understand the distribution of synteny breaks in different genomic region. The chromosome was scanned using a sliding window approach. Different frame sizes were used (from 20 Kbp to 70 Kbp) and the number of

breaks in each window have been counted as shown in Figure 3.10. Breakpoints were considered to fall into a window if they have at least one position overlap with that window. Then, under the distribution curve, all windows falling in the left most 5% have been considered as highly conserved windows and those of the right most 5% have been considered as highly fragile. Then these two groups of windows were put aside and a z-score was calculated for each window, based on the remaining population ($z\text{-}score = \frac{breakpoints - \bar{x}}{\sigma}$). In the next step windows were classified in three groups; ***fragile (F)*** : those with a z-score greater than 2, *conserved (C)* : those with a z-score lower than -2, and the rest as *others (O).*



window	W_{n-1}	W_n	W_{n+1}
nb breakpoints	5	6	4

Figure 3.10: In this Figure, red lines represent breakpoints that distribute along the chromosome (two-headed arrow). We scanned the chromosome with a sliding window and counted the number of breakpoints that have at least one position overlap with that window.

3.2.1 Association of genomic markers

As we explained in 1.3.2 and 1.3.5, several markers are known to be associated with genomic instability in the course of evolution as well as disease development. Based on this fact, we have decided to choose 4 of such genomic markers to study their associations with our identified genome regions in this study. These markers are as follows : Known genes, CpG-islands, repeats and G-quadruplex. Annotation tracks of known genes, CpG islands and repeats have been obtained from ENCODE project available on UCSC Genome Browser public site for the Feb. 2009 assembly of the human genome (hg19/Genome Reference Consortium Human Reference

37) (Rosenbloom *et al.*, 2013). Annotation tracks for G-quadruplex have been obtained from G4 database (Wong *et al.*, 2010) and converted from NCBI Build 36 to GRCh37 coordinates using liftOver (Hinrichs *et al.*, 2006) utility that is available on UCSC Genome Browser website. We assigned a unique ID to each annotation. Annotation with different names but exact positions and orientations were combined together using *Perl scripts*. Association of those markers with each region is defined in such a way that a marker is associated with a region if it has at least one nucleotide that overlaps with that region. As schematized in the Figure 3.11, all purple annotations are considered to be associated with the *Blc_I*. Annotations in blue are associated with the *Blc_II* and the green annotations is associated with both. And lastly, the red marker is considered not to be associated with any of the two regions.

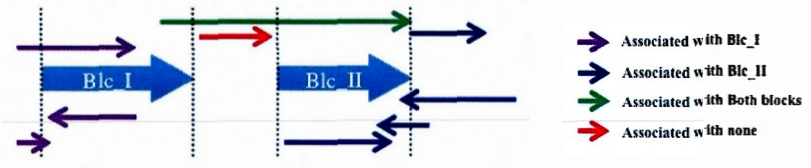


Figure 3.11: Each thin arrow represents an annotation. Purple and blue annotations are considered to be associated with Blc_I and Blc_II respectively with a minimum 1-nucleotide overlap with those blocks. Green annotation is considered to be associated with both blocks and the red annotation has not been considered as associated to any block.

CHAPTER IV

RESULTS AND DISCUSSION

4.1 Extraction of synteny region on human chromosome 1

From MSA of 45 vertebrate genomes against human chromosome 1 (multiz46way), available from UCSC genome browser website (Kent *et al.*, 2002), 1833468 alignment blocks with at least 7 selected species have been extracted. All adjacent blocks with a distance less than 1 Kbp with respect to all species fused which resulted in 141,035 conserved blocks as explained in 3.1.4. The size distribution of these blocks are presented in Figure 4.1, which varies from 6 to 16,869 nucleotides (nt) with regard to human sequence. Although the contribution of human genomic sequence would diverge from 1,000 to 16,869 nucleotides but they extend from 1,001 nt to 17,694 nt on the chromosome. This is due to the fusion of neighboring blocks with a distance < 1 Kbp following the absence of small regions in the multiple sequence alignment. This is another reason that we tolerated up to 999 nucleotide-distance to fuse the neighboring blocks. As we described our method in the last chapter, alignment blocks with size less than 1 Kbp will not be included in our analysis. Excluding all the blocks with a size less than 1 Kbp (100,510 blocks), with respect to human, 40,525 conserved alignment blocks was identified as synteny blocks. We have extracted 40525 synteny blocks on human chromosome 1 with sizes varying from 1000 to 17694 base pairs. More than 80%

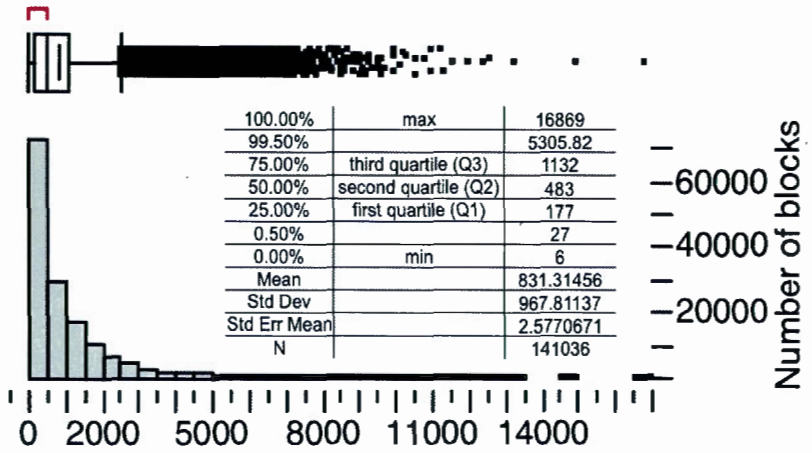


Figure 4.1: Distribution of size of conserved blocks on chromosome 1 resulted from the last fusion step.

of these blocks have a size less than 5 Kbp. The size distribution of these blocks are presented in Figure 4.2. These fragmentary blocks are mostly due to either missing data or micro-rearrangement affecting the genome architecture of several species.

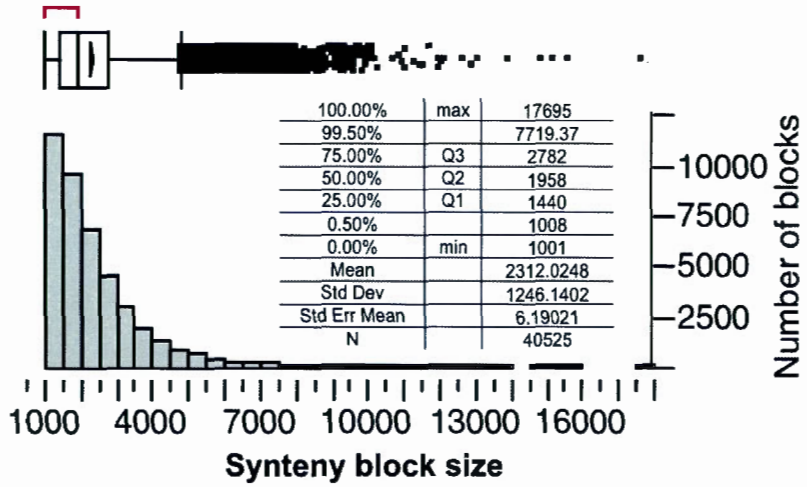


Figure 4.2: Distribution of size of human synteny blocks on chromosome 1.

The proportion of each genome in our blocks represented by Figure 4.3, which

shows that the extracted blocks have covered less than half of chromosome 1 (~47%). As can be seen in this Figure, the contribution of each species in our data shows almost a direct correlation with the evolutionary distance between that species and human. Other than opossum, all mammals contributed in over 60% of our extracted blocks. This should be as a result of the low-coverage genome assembly of the opossum. The low contribution of chicken (~6%) is due to the evolutionary distance between human as a primate and chicken as a non-mammalian vertebrate and was expected as well. The only incoherence between contributions and the evolutionary distances between species is with regards to the rodent branch. Even though they are Human closest neighbors, their contributions are less than those of the more distantly related species. But, this is mainly due to the absence of their sequence in a fraction of about 30% of the initial alignment blocks. Table 4.1 shows the contribution of each species in more details. As mentioned in this table, all placental mammals shared at least 85% of human synteny blocks, except the rodents. This could be due to the short life spam of the rodents compare to other placental mammals which means more generations lead to higher genomic modifications with compare to their evolutionary cousins. Moreover this coverage goes over 90% in primates.

4.1.1 Association of genomic markers with synteny blocks

To gain more insights on functional and structural characteristics of extracted synteny blocks we studied their associations with four genomic markers known to have a role in genome stability as described in 3.2.1. These markers are as follows : known genes, repeats, CpG-islands, and G-quadroplex. The summary of the association of these markers with each synteny block is presented in Figure 4.4. The result shows that about 60% of extracted synteny blocks were associated with coding genes. 20% of the blocks are not associated with any of the selected

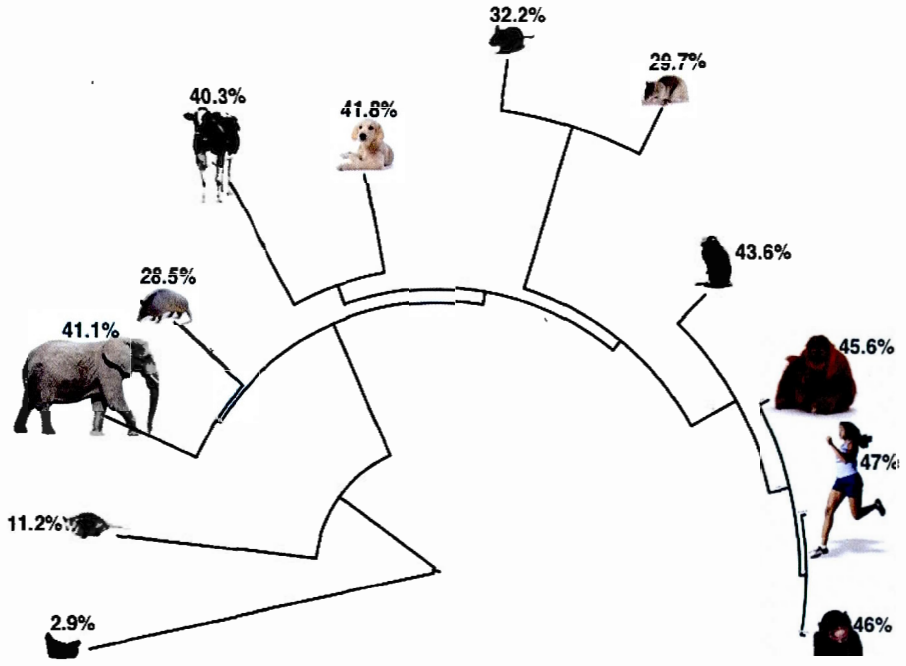


Figure 4.3: Percentage contribution of each genome in extracted blocks. The percentage values describe the contribution of each genome in MSA of 45 vertebrates against human chromosome 1. The values are coherent with evolutionary distances between each genome and the human genome. The lower values for rodents are due to their absence on 30% of initial alignment blocks. Furthermore the level of contribution of armadillo as a placental mammal on human chromosome is probably as a result of its low-coverage genome assembly.

markers. These non-coding conserved blocks that carry no repeat elements regions should be of those stable gene deserts explained in 1.1.3. The high number of synteny blocks that are associated with repeats (40%) was expected, as repeats are the most abundant elements in the genome.

4.1.2 Inference of the evolutionary history of extracted synteny blocks

For each synteny block, a phylogenetic tree has been constructed using a bayesian approach. The bayesian inferred trees were compared with the species tree using

Species name	chrCoverage	blockCoverage	nbNucleotide
Chicken	2.88%	6.12%	7183585
Opossum	11.23%	23.87%	27992836
Armadillo	28.54%	60.68%	71151661
Elephant	41.15%	87.48%	102569352
Cow	40.25%	85.58%	100345213
Dog	41.78%	88.83%	104151862
Rat	29.66%	63.06%	73939684
Mouse	32.20%	68.47%	80279859
Marmoset	43.57%	92.63%	108608627
Orangutan	45.63%	97.00%	113733562
Chimpanzee	46.31%	98.46%	115446892
Human	47.03%	100%	117245231

Table 4.1: Contribution of each species in extracted blocks. The second column displays the contribution of each genome in initial multiple sequence alignment of 45 species against human. The third column carries information on the presence of each genome in our extracted conserved blocks. The last column shows the contribution of each genome in the initial MSA by nucleotide.

Robinson and Foulds (RF) topological distance algorithm. As one can see in Figure 4.5, the distribution of the RF distances shows that inferred topologies have strong discordance with species tree topology. About 50% of synteny blocks show an RF distance over 4. The question that raises is : if these blocks are conserved, and they kept their synteny through evolution of mammals, why should such incongruity with the well-accepted phylogenetic tree of species be observed? To understand the reason, we looked at the quality of alignment blocks. We observed that most of the blocks having disagreement with the species tree are harboring missing information, and we have replaced them with gaps in the extraction step. Therefore, we verified all alignment blocks and for each block we identified every species with complete missing information and excluded them from the inferred tree. Then, for each such block we recalculated the RF distance of the altered tree with its corresponding species tree. After this cleaning, the distribution of RF distance showed much more agreement with the species tree (see Figure 4.6).

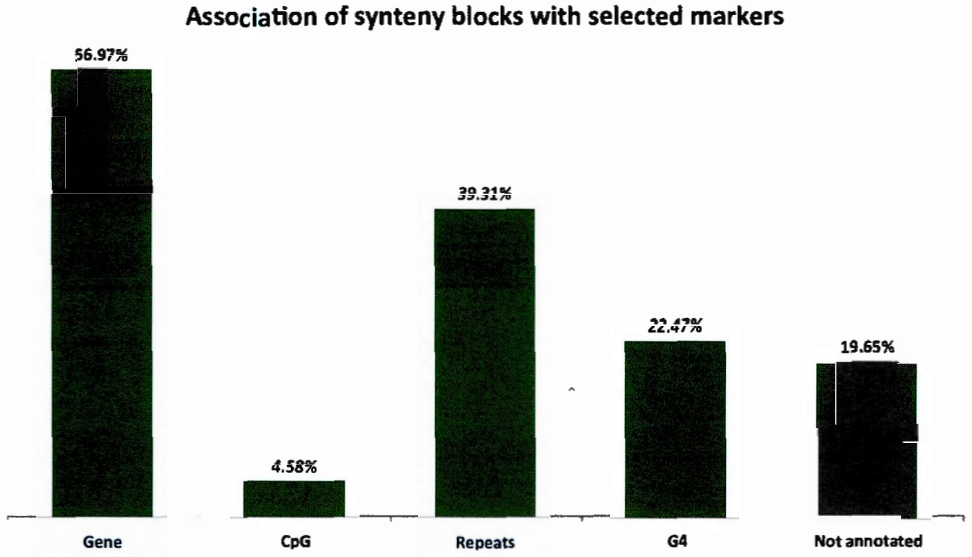


Figure 4.4: Association of synteny blocks with the four selected genomic markers. This shows the number of synteny blocks that overlaps with at least one nucleotide position with each marker.

Over 75% of synteny blocks showed an RF distance smaller than 4 from the species tree. The remaining disagreements could be still due to partial missing information or alignment quality. These results supported the quality of the synteny extraction method.

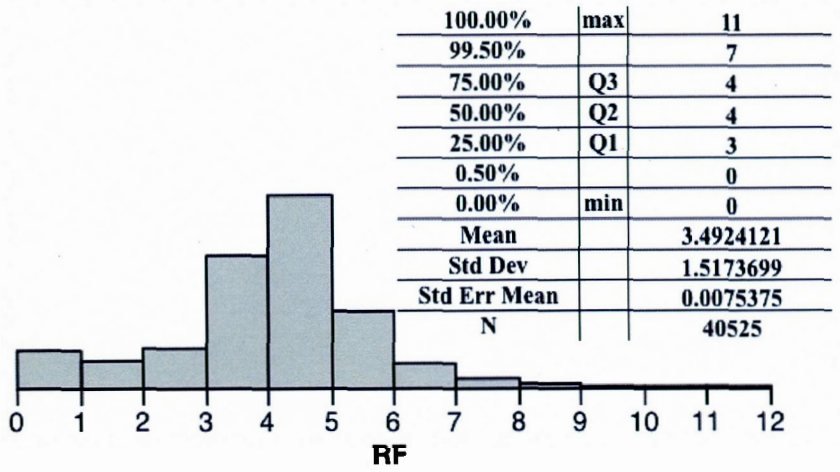


Figure 4.5: Distribution of RF distances between species tree and inferred tree

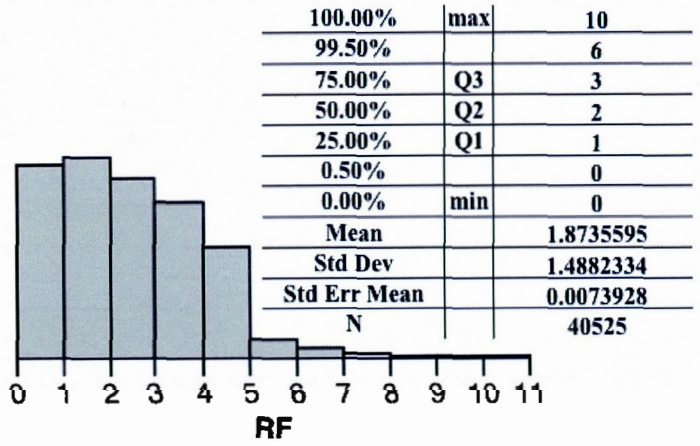


Figure 4.6: Distribution of RF distances between corrected inferred trees and their corresponding species tree.

4.2 Extraction of synteny breaks

To pinpoint genomic regions that are enriched in breakpoints (***fragile region*** or ***rearrangement hotspots***), one should first identify breakpoints along the genome and then identify the regions where those breakpoints are clustered. In this project, the breakpoints are defined as regions bounded by two consecutive synteny regions (Lemaitre *et al.*, 2008), such that the size of the region is lower than 1 Mbp. This constraint is necessary to avoid ambiguous regions that could not be associated correctly to breakpoints (e.g. centromeres). It should be reiterated that, we fused all adjacent conserved blocks with less than 1 Kbp gap between sequences of all species, to delineate synteny blocks. Also, we defined a synteny block, as a conserved block between the selected species with a minimum size of 1 Kbp. Hence, we excluded all the mini blocks in the previous step. We identified 40,525 synteny blocks. So any factor that prevented the fusion of two contiguous blocks could indicate the force that broke the synteny between those two blocks. To identify synteny breaks on human chromosome 1, we examined all regions between each two synteny blocks and documented all the forces against fusion of those blocks. It was observed that 7101 regions between synteny blocks were not surrounded by blocks with a broken synteny. This means that those regions don't have the characteristics of breakpoints. Looking back at all enclosing blocks it was observed that these blocks were not fused in the previous step due to the small micro-rearrangements that happened within those excluded short conserved regions between them. Figure 4.7 demonstrates such episodes. As one can see, the blue block, between the first two synteny blocks, is located on a different chromosome with respect to the genome of the chicken. Therefore, at the fusion step we were not able to fuse it with neither *BlcSyn1* nor with *BlcSyn2*. But as the blue block has been excluded due to its length, one could observe that *BlcSyn1* and *BlcSyn2* are in the same synteny. Also the small red block in chicken is located

after the *BlcSyn3* with a distance greater than 1 Kbp from the *BlcSyn2* which prevented the fusion process between these two blocks. Furthermore, when this small block would be removed due to its size, *BlcSyn2* and *BlcSyn3* no break of synteny between them could be observed. Due to similar phenomena, 7,101 out of

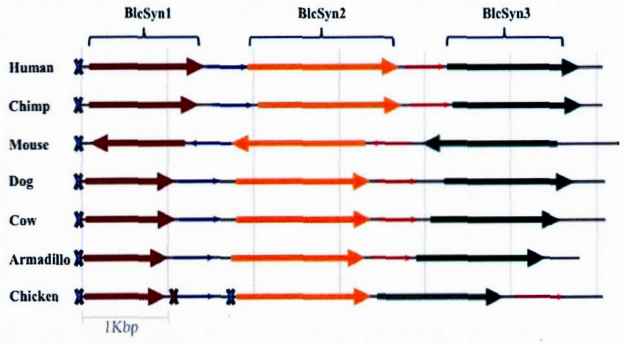
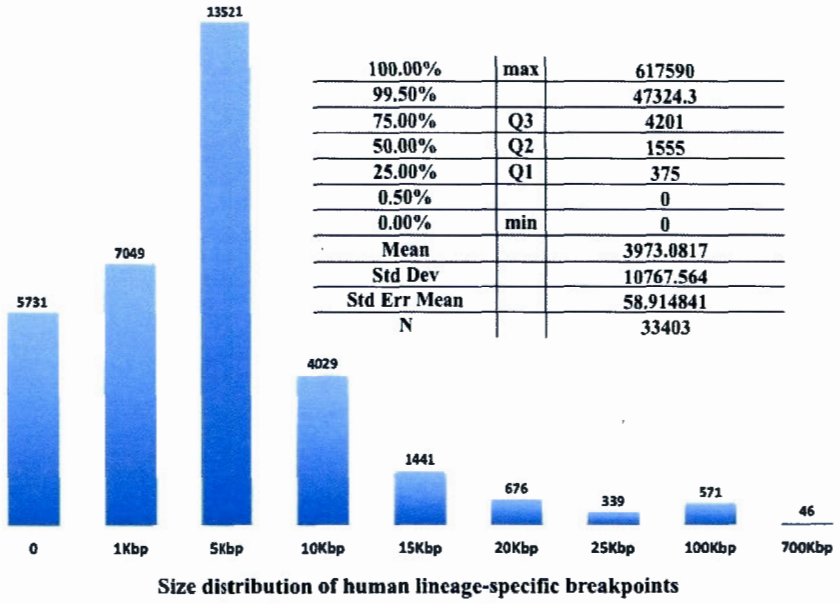


Figure 4.7: Micro-rearrangements phenomenons : In this example, the BlcSyn1 and BlcSyn2 will be associated to the same syntenic region when the small conserved region is removed. This small region has a size less than 1,000 nt (blue region) and is removed, which is a result of an insertion in the genome of the chicken rearrangement.

40,524 regions between synteny blocks have been excluded from the collection of potential breakpoints for further analysis. Next, we added the two chromosome extremities to these breaks. So we are left with 33425 potential breakpoints. Within these breaks only one had a length greater then 1 Mbp, which was overlapping with the centromere region of the chromosome, as expected, and was excluded from the potential breakpoints as well. Other than centromere regions, the rest of discontinuations should be most probably due to the missing data and alignment quality as well as our extraction protocol, which accepts all conserved alignment blocks having at least 7 species from our selected species presented. The size distribution of the extracted breaks that is presented in Figure 4.8 ranges from 0 to 617,590 bp. About 17% of those breaks have a size equal to 0. This was predicted as the multiple alignment has been produced based on the human genome. Therefore, all human blocks are almost adjacent.



100.00%	max	617590
99.50%		47324.3
75.00%	Q3	4201
50.00%	Q2	1555
25.00%	Q1	375
0.50%		0
0.00%	min	0
Mean		3973.0817
Std Dev		10767.564
Std Err Mean		58.914841
N		33403

Figure 4.8: Distribution of size of human synteny breaks on chromosome 1

4.2.1 Identification of breakpoints specific to the human lineage

The main objective of this master thesis was to identify the fragile regions on the human genome. Synteny breaks were identified using comparative analysis of the twelve selected species in the previous step. In the next step, human lineage specific synteny breaks have been identified. The rest of the breaks have been excluded for further analyses. As explained above, during the synteny break identification step, for each break, all the forces that have broken the synteny (change in chromosome location or orientation as well as a distance >= 1 Kbp) between each two adjacent synteny blocks have been documented. Going back again to the Figure 3.5, we noted, for example, that the first break is with respect to the mouse genome. The second break was due to the separation of two blocks on cow and dog genomes and the third was again with regards to the mouse genome. Thus based on the assumption that similar elements in two or more contemporary genomes are driven from the common ancestor of those who carry that element, we now want to find

out the common ancestor of breaks shared among species. Using LCA algorithm, the common ancestor of all extracted breaks have been identified. 33,404 breaks out of 33,425 extracted breaks, have been identified as human lineage-specific synteny breaks, as they were located on one of the common ancestors of human. The LCA result is presented in Figure 4.9. As can be seen in this Figure, the number of identified non-human lineage-specific breaks is negligible (0.06%) as almost 100% of the breaks are identified as human-lineage specific.

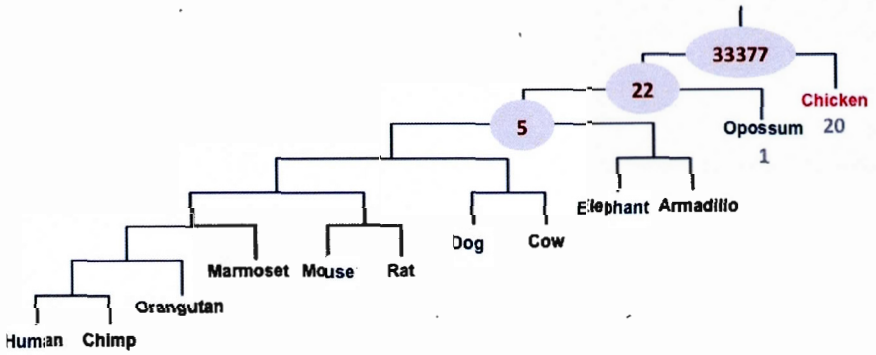


Figure 4.9: This shows the number of breaks identified for each leaf or internal node on the species tree. Dark red indicates the breaks that are traced back on human lineage. Almost all the breaks were identified as human-lineage specific breaks. And only 21 breaks were identified as other species-specific breaks.

4.2.2 Association of genomic markers with breakpoints

Certain genomic markers are known to be associated with genome rearrangements. For instance, markers such as repeats, G-quadruplex and some genes (see 3.2.1). Thus, we studied the association of our extracted breaks with those markers. The result presented in Figure 4.10 shows that 60% of breaks are associated with genes. This result was not expected but it could be associated to the fact that these regions constitute to most sequenced ones in the different organisms. The obtained preliminary results should be investigated further in the future since here

we only provide it as an overview.

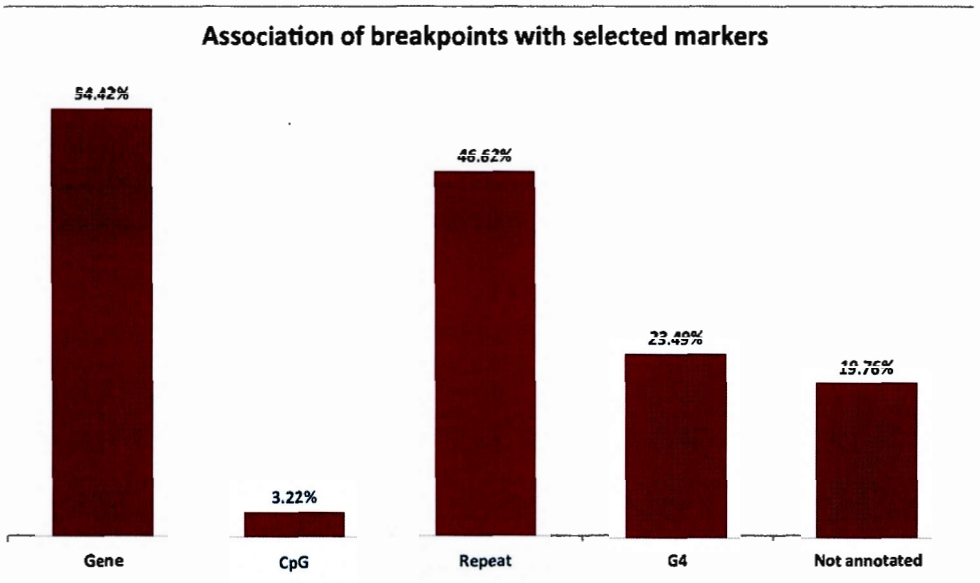


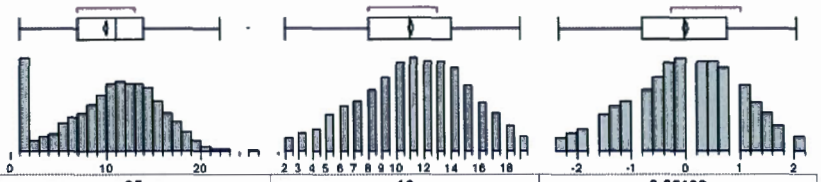
Figure 4.10: Association of synteny breaks with the four selected genomic markers. 60% of synteny breaks were associated with the genes. This was not as expected and it should be verified.

4.3 Identification of fragile region on human chromosome 1

We scanned the chromosome using a sliding window approach with frames from 20 Kbp to over 100 Kbp. In each window, the number of synteny breaks (breakpoints) were counted. Then, we studied the distribution of those counts. We observed that windows with sizes below 30 Kbp show very week distributions. The same thing was true for window size over 80 Kbp. The distributions in window over 80 Kbp were highly dispersed. Comparing the distribution of breaks in different window size, we observed that the breakpoint distribution in 70 Kbp windows is more consensual. Hence we fixed the window size at 70 Kbp for the remaining of the study.

As explained in 3.2, the two extremities (outliers) of the distribution were put

aside as highly fragile and highly conserved on the right and left side of the distribution respectively. Next, the z-score for each window was computed. Then the distribution was divided in three parts. The left and right tails falling among respectively Z-score <-2 and Z-score > 2 were classified as *conserved (C)* and *fragile (fragile)* regions respectively and the rest of the windows are categorised as non-fragile non-conserved (*others (O)*) regions. The summary of these distributions are presented in Figure 4.11. The overall view on the chromosome confirmed



100.00%	max	25	19	2.05123
99.50%		20	19	2.05123
75.00%	Q3	14	14	0.75856
50.00%	Q2	11	11	-0.017
25.00%	Q1	7	8	-0.7926
0.50%		1	2	-2.3439
0.00%	min	1	2	-2.3439
Mean		10.010671	11.06592	-8.91E-17
Std Dev		4.9356987	3.8679558	1
Std Err Mean		0.0827109	0.0691921	0.0178885
N		3561	3125	3125
Distribution		number of breaks (nbBrk)	Excluding two 5% extremities	z-score

Figure 4.11: Distribution of synteny breaks in sliding window of size 70 Kbp on chromosome 1.

a non-homogeneous distribution of breakpoint accumulation as expected. Figure 4.12 shows the distribution of breakpoints along human chromosome 1. As explained in 4.2.1, only one identified breakpoint were filtered due to its size (> 1 Kbp) which also overlapped with the chromosome centromere. This region is

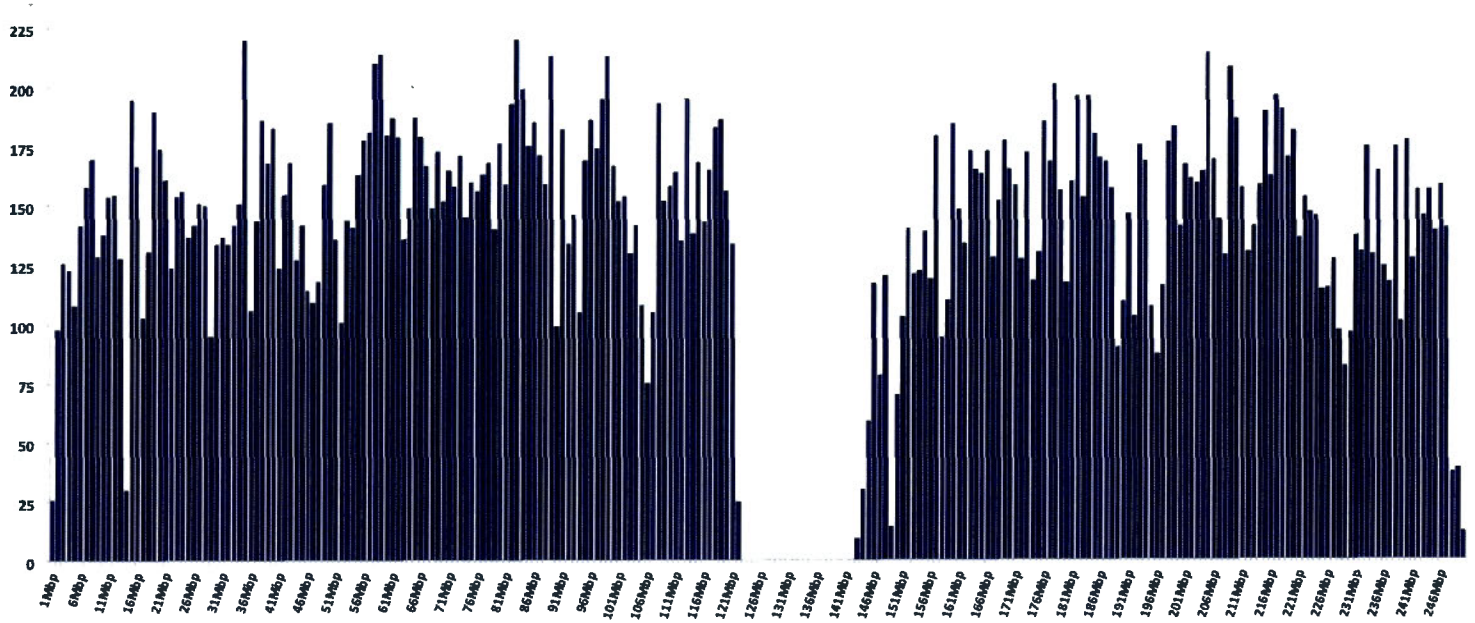


Figure 4.12: The Figure shows the number of human lineage-specific synteny breaks in genomic regions of 1 Mbp along human chromosome 1. As one can see in this Figure, extracted synteny breaks don't have the same affinity to different genomic regions. The region presenting no breakpoint overlaps with the chromosome centromere.

With the previous procedure, 75 windows were identified as fragile region. The contiguous windows were fused together to identify the fragile regions. 72 fragile regions were identified having sizes from 70 Kbp to 140 Kbp.

4.3.1 Association of genomic markers with each window frame

To have more information on the characteristics of these regions, we decided to analyze the functional and structural characteristics of these regions. As explained in 1.3.2 and 1.3.5, some genomic markers are known to be associated with the level of fragility in the genome. Therefore, we decided to study the distribution of the four selected markers since their involvement in genome fragility are well accepted such as known genes, CpG-islands, repeats and G-quadruplex. The association was defined as at least one nucleotide overlap with each window for all sizes. The number of total genomic markers associated with these regions are presented in Figure 4.13. Surprisingly, the results of the association shows a presence of G4 markers and repeats in all the 72 identified blocks. One can highlight that the corresponding markers are overrepresented in the 72 selected window frames.

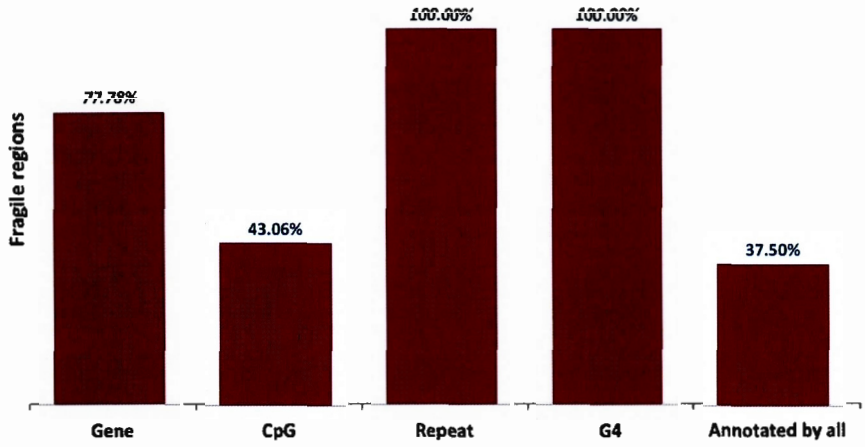


Figure 4.13: Association of fragile regions with selected markers.

4.3.2 Robustness of the identified fragile regions

The computational extracted fragile regions derive from several different computational step containing their own biases. Here, we decided to take a conservative approach by taking in each step the most stringent criteria. We are aware of the necessity of further analyses to better assess the robustness of our extracted fragile regions. However, the preliminary results, presented in this master thesis, highlight several interesting facts. For instance, other than 16 fragile regions that were not annotated with any gene, the rest of the regions overlap with 195 genes including genes associated with human diseases. The summary of those genes are represented in Table 4.2.

Gene name	Associated condition
PAX7	Alveolar rhabdomyosarcoma
LAMB3	Epidermolysis bullosa
ALDH4A1	Hyperprolinemia (HP)
GJA5	Atrial fibrillation
USH2A	Retinitis pigmentosa (RP)
USH2A	Usher syndrome (US)
NRAS	Malignant melanoma
NRAS	Autoimmune lymphoproliferative syndromes
NRAS	Noonan syndrome and related disorders
HMCN1	Macular degeneration
TNFRSF1B	Graft-versus-host disease
DPYD	Dihydropyrimidine dehydrogenase deficiency

Table 4.2: List of genes associated with identified fragile regions and diseases. The associated conditions are extracted from KEGG DISEASE Database (Kanehisa *et al.*, 2014). Maglott *et al.* (2005); Becker *et al.* (2004); Pletscher-Frankild *et al.* (2014); Huret and Senon (2006)

Identified fragile regions vs. chromosomal fragile sites

As mentioned in 1.3.6, certain sites on human chromosomes are cytogenetically identified as fragile sites. These sites are corresponding to chromosome bands and their exact positions in molecular level are not known yet (Savelyeva and Brueckner, 2014). However, we compared these regions with our results. 19 of these sites are identified on the chromosome 1, out of which, only one is known as a rare fragile site (Lukusa and Fryns, 2008; Mrasek *et al.*, 2010). This site (FRA1E) is located on 1p21.3 (chr1 :97,749,961-98,119,925) (Savelyeva and Brueckner, 2014). Figure 4.14 shows the FRA1E (rare fragile site) along with our identified regions uploaded as custom tracks on UCSC genome browser. As one can see, there is no identified conserved region presents in this site. Moreover, out of 69 regions identified as "others", 9 have a very elevated number of breakpoints (highlighted as pink and light red), means that these windows have 1 or 2 breaks less than the defined threshold (19) to be identified as fragile regions. Comparison of these regions with somatic mutation in cancer (COSMIC) (Bamford *et al.*, 2004) track shows that these non-fragile but on the edge of fragility, mostly overlaps with cancer mutations. In addition, out of 4 fragile regions presented in this site, two overlap with DPYD gene. DPYD mutations results in dihydropyrimidine dehydrogenase. Although the contribution of alterations in DPYD in cancer development is unknown, genomic alteration in this gene is not rare in cancer cells (Gross *et al.*, 2013; Savelyeva and Brueckner, 2014).

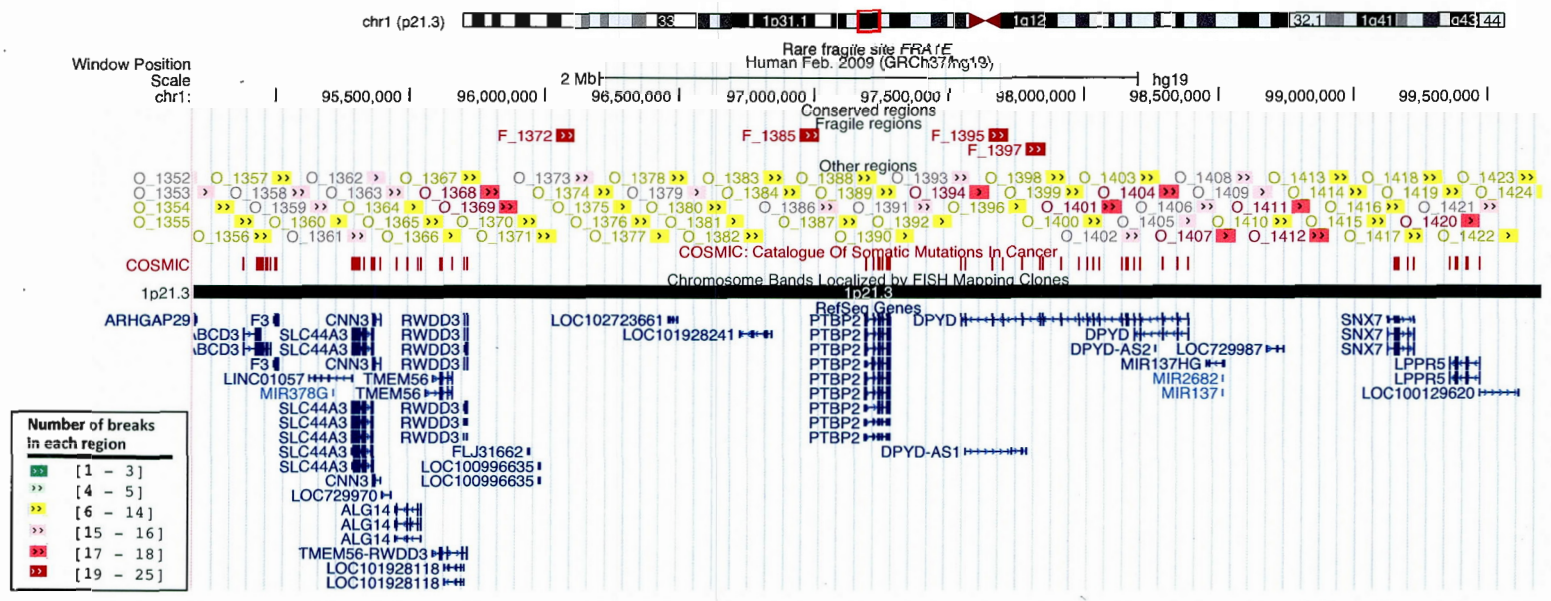


Figure 4.14: Visualization of the rare fragile site of chromosome 1. Identified regions are uploaded to UCSC genome browser as custom tracks. There is no identified conserved region presents in this site. In addition to those 4 fragile regions, we could observe that those regions categorised as non-fragile non-conserved (*others (O)*) are mostly overlap with somatic mutations in cancer, according to COSMIC (Bamford *et al.*, 2004)

Another interesting result shows that the most fragile region (F_2520 with 25 breakpoints) overlaps with PAPPA2 gene, According to ClinVar (Landrum *et al.*, 2013), this gene has been reported to be associated with lung cancer and malignant melanoma. The association of this gene with two identified fragile regions is shown in Figure 4.15. Looking at previously identified conserved synteny on chromosome

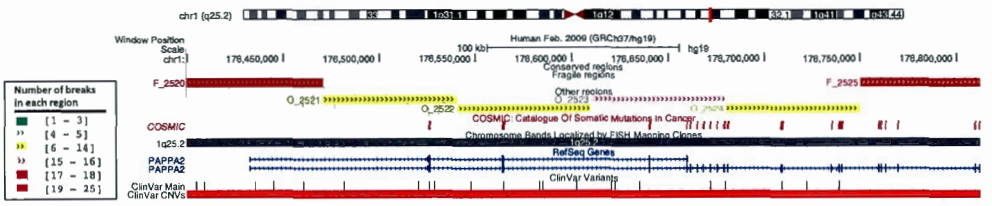


Figure 4.15: Visualization of overlaps of the most fragile regions with PAPPA2 gene. With 25 breakpoints, F_2520 region is the most fragile region that starts 32 Kbp upstream of this gene.

1 also showed some promising results. For instance, a region of ~3.7 Kbp on the q arm of the chromosome has been identified as a synteny block by a comparative analysis (Larkin *et al.*, 2009). Half of this region has also been identified as a non fragile site by Fungtammasan *et al.* (2012). However, our result identified only two specific regions as conserved regions as well as three as not a conserved but regions with very low number of breaks. See Figure 4.16. As this figure shows, these 5 regions have no overlaps with any known somatic mutations in cancer, COSMIC (Bamford *et al.*, 2004). Moreover there are 7 regions in this window were identified as non-fragile but with a high number of breaks. These regions show overlaps with somatic mutations in cancers. These results need to be confirmed with more profound analysis. However it seems that our method could identify fragile region with a higher specificity comparing with other methods.

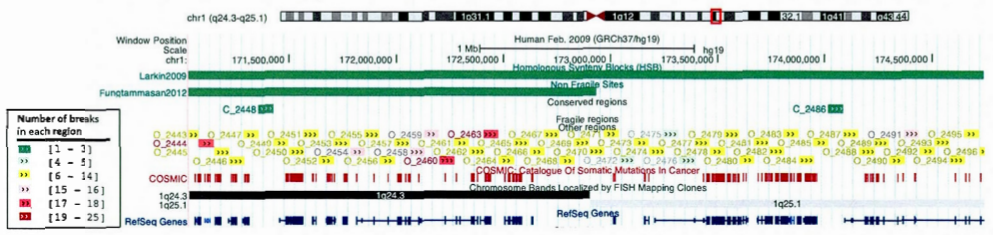


Figure 4.16: This figure presents a region that have previously identified as a conserved region by two studies along with our results as custom tracks on UCSC genome browser. Two conserved regions with three other non-conserved but with a very low number of breaks are located in this region that have no overlap with somatic mutation in cancer.

Gene Ontology (GO)

For the corresponding 159 protein coding genes, a Gene Ontology (GO) analysis was performed based on biological process, cell components and molecular functions. GO analysis revealed that, based on biological process, fragile regions are highly enriched in genes involved with anatomical structure development (1.398448e-014), circulatory system processes (p-value 1.77e-008), cell differentiation (p-value 4.59e-004), and morphogenesis (p-value 1.68e-001). Based on cellular component those regions are enriched in plasma-membrane (p-value 3.36e-04), proteinaceous extracellular matrix (p-value 3.00e-5), and cellular component (p-value 3.45e-03). Finally, based on molecular functions, they were enriched only in nucleic acid binding transcription factor activity (p-value 8.06e-04). These results are represented in Figures 4.17, 4.18, and 4.19, respectively.

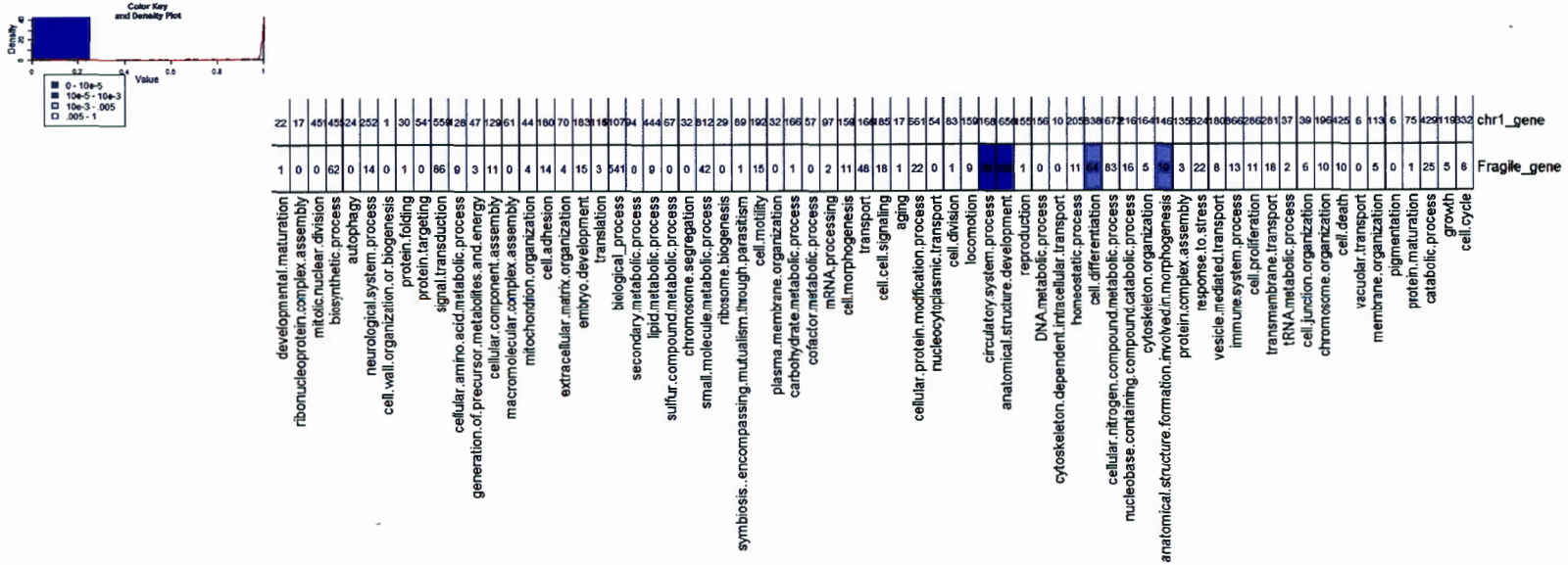


Figure 4.17: Gene Ontology (GO) based on biological process

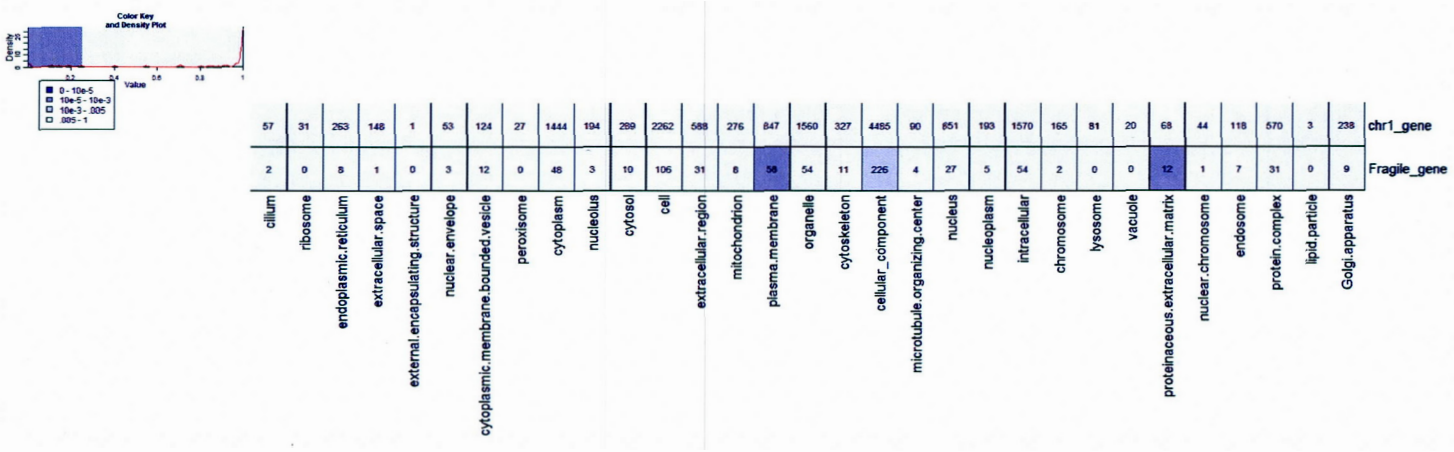


Figure 4.18: Gene Ontology (GO) based on cellular component

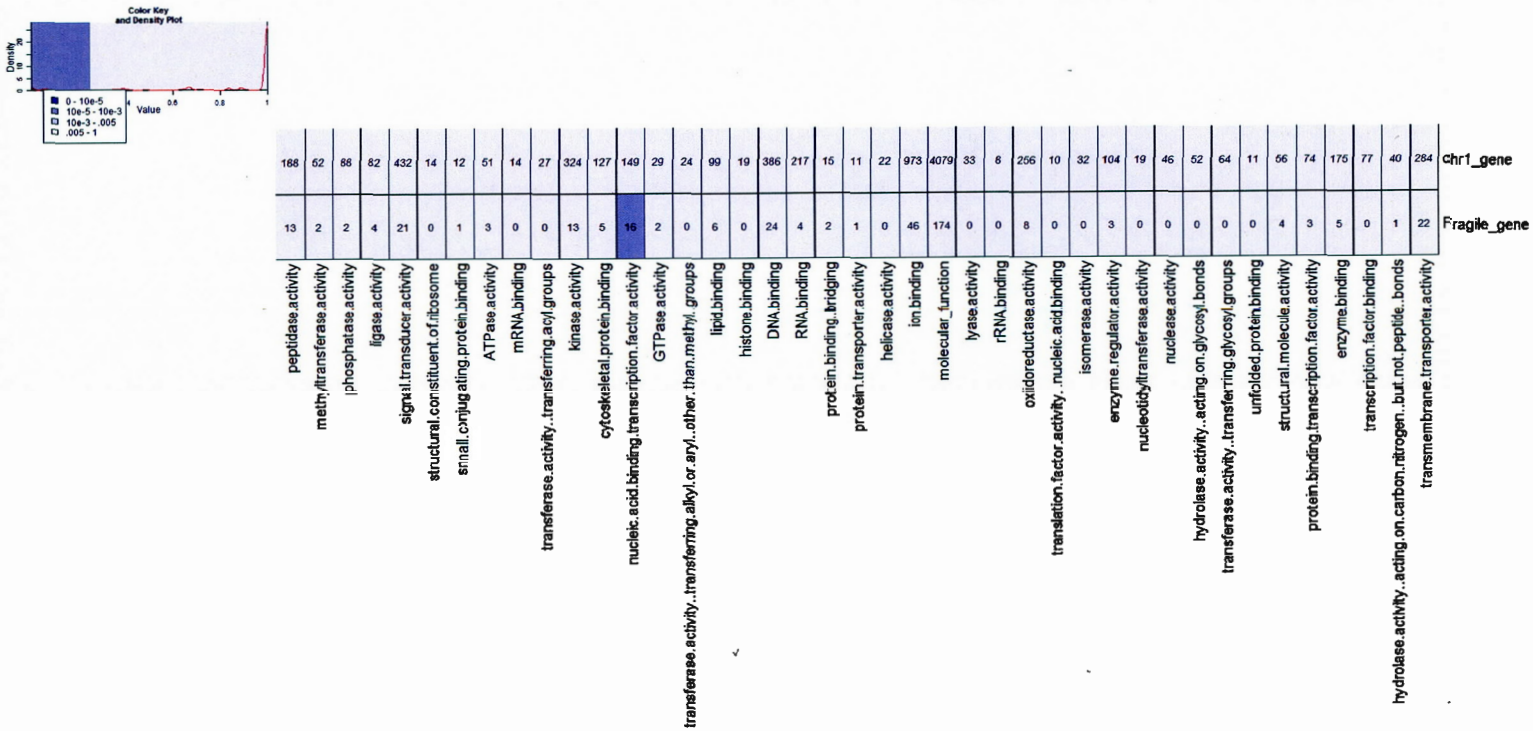


Figure 4.19: Gene Ontology (GO) based on molecular function

[Cette page a été laissée intentionnellement blanche]

CONCLUSION

The dynamic nature of genomes and genome conservation among diverse species have interested biologists for several years. Now, we know that genomic regions are not uniformly conserved throughout genomes, and this level of conservation is not randomly distributed throughout the genome either. Three main characteristics conduct to look for possible overlaps between genomic regions contributing to evolutionary processes and regions that have higher probability to be cancer-associated. These characteristics are (1) non-random distribution of genomic rearrangements that drive human diseases (such as cancer), (2) several genomic markers are known to be associated with genome instability, (3) and mechanisms of genome rearrangements in both evolutionary and cancer-developing processes.Within the limits of this thesis, we focused on the development of a method to extract evolutionary rearrangements (breakpoints), reconstruct evolutionary history of breakpoints and statistically identify genomic regions that are enriched for breakpoints (fragile regions).

Large genomic data for various species, improved sequencing and alignment methods, as well as annotation on genomic functional elements paved the way to a better understanding of genome structure and function. One way of exploiting such technology is through comparative genome analyses. Comparative genomics is a powerful approach to extract evolutionary conserved and/or highlight fragile regions. In this study, we conducted a comparative analysis of human with 11 other vertebrates (10 mammals) to identify evolutionary breakpoints. Species selection was performed based on their evolutionary relation with human as well as

the quality of their genome sequencing and assembly. Due to the quantity of genomic data and the required computational time to survey the whole genome, we limited this preliminary analysis to human chromosome 1. Chromosome 1 is one of the most ancient and longest human chromosome. This allowed us to develop, tune, test and improve our method on a simple human chromosome.

With chromosome 1, we identified 40,525 syntenic regions that cover 47.03% of the chromosome. Phylogenetic analysis showed that about 70% of these regions have the same evolutionary history as the species evolution. The discordance is mainly due to missing data and the quality of the alignment. These syntenic regions covered about 40% of the chromosome. Over 90% of these regions are conserved in all primates. Other placental mammals share also from 60% to 88% of these regions with human as well. It is important to notice, the identified breakpoints were not distributed randomly across the chromosome. This is coherent with previous studies (Pevzner and Tesler, 2003c; Peng *et al.*, 2006; Lemaitre *et al.*, 2009). Applying the LCA approach, we observed that almost all of those breakpoints are of reused type. To identify enrichment of breakpoint within genomic regions, we applied a sliding window approach. To decide on the length of the windows we scanned the chromosome with different sizes, ranges from 20 Kbp to 100 Kbp. The distribution of breakpoints using 70 Kbp window frames, was the closest to a normal distribution according to the goodness of fit. Hence, we continued our analysis with this frame size. The distribution of the breakpoints was normalized and z-score computed. Then using a threshold for z-score > 2, fragile windows were identified. This means that any window that has a z-score greater than two would be categorized as a fragile window. Using this approach, 72 fragile regions were identified having a size ranging from 70 Kbp to 140 Kbp. These regions overlap with previously identified fragile sites and gene markers associated with human diseases. In addition these To have a better understanding of the functional signi-

ficance of other associated genes, GO analysis was performed based on biological processes, cellular component, as well as molecular function. The result illustrated the enrichment of those regions with genes associated mostly with anatomical structure. Based on cellular component, identified fragile regions are enriched in plasma-membrane and extracellular matrix. Finally, GO analysis highlighted the enrichment of regions with genes associated with nucleic acid binding transcription factor activity. Further tests and analyses will be needed to investigate the role of identified fragile regions.

In the future, we will add other algorithms for breakpoint identification of fragile regions. Since the design of the pipeline is completed, after several assessment, we will apply it to the whole human genome to better identify and correlate fragile regions. This will be the starting point of our long term objective (in my PhD thesis) consisting of the study of the correlation between evolutionary fragile regions and cancer-associated rearrangements. Hence, we will attempt to answer the following question : Are evolutionary fragile regions more susceptible to cancer rearrangements ? To achieve this goal and answer to this question, we will continue this project by collecting cancer rearrangement data from previous published studies and databases (Kost-Alimova *et al.*, 2003; Darai *et al.*, 2005; Darai-Ramqvist *et al.*, 2008). Then we will compare the affinity of these breakpoints to different identified genomic regions from previous steps using statistical analyses. The results of such study would highlight the genomic regions that have potentials to be rearranged in cancer development and suggest therapeutic targets.

[Cette page a été laissée intentionnellement blanche]

ACRONYMS

CFS Common Fragile Sites. 31

CNV Copy Number Variation. 24

DNA deoxyribonucleic acid. ix, 1, 3, 6, 12, 14, 22, 23, 28–30

EBV Epstein–Barr Virus. 30

GRCh37 Genome Reference Consortium Human Reference 37. 51

HBV Hepatitis B Virus. 30

HGT Horizontal Gene Transfer. 20

HPV Human Papilloma Virus. 30

LCA Lowest Common Ancestor. ix, x, 19, 20, 49, 63

LCR Low Copy Repeat. 23

MSA Multiple Sequence Alignment. vi, 10, 15, 18, 19, 42, 46, 53, 56, 57

NAHR Non-Allelic Homologous Recombination. 23

RNA ribonucleic acid. 6

SCNA Somatic Copy Number Alterations. 29

[Cette page a été laissée intentionnellement blanche]

GLOSSARY

centromere "The constricted region of a chromosome that is the position at which the pair of chromatids are held together." (Brown, 2002). 4

Giemsa stain It is a stain that have more affinity to regions of chromosome that are enriched in A$\bar{\text{T}}$ bonding. This creates dark and light bands, which is specific to each chromosome. Giemsa-banding (G-banding) is schematized with ideogram. These bands are used as map to genomic locations (Brown, 2002; Library, 2013). ix, 4, 5

MAF file format "The multiple alignment format stores a series of multiple alignments in a format that is easy to parse and relatively easy to read. This format stores multiple alignments at the DNA level between entire genomes" (Rhead *et al.*, 2009). 39, 40, 42

Newick format "The tree file it is represented by the following sequence of printable characters: (B,(A,C,E),D); The tree ends with a semicolon. The bottommost node in this tree is an interior node, not a tip. Interior nodes are represented by a pair of matched parentheses. Between them are representations of the nodes that are immediately descended from that node, separated by commas. In the above tree, the immediate descendants are B, another interior node, and D. The other interior node is represented by a pair of parentheses, enclosing representations of its immediate descendants, A, C, and E. In our example these happen to be tips, but in general they could also be interior nodes and the result would be further nestings of parentheses, to any level.Tips are represented by their names. A name can be

any string of printable characters except blanks, colons, semicolons, parentheses, and square brackets. Because you may want to include a blank in a name, it is assumed that an underscore character ("_") stands for a blank; any of these in a name will be converted to a blank when it is read in. Any name may also be empty: a tree like (,(,),); is allowed. Trees can be multifurcating at any level. Branch lengths can be incorporated into a tree by putting a real number, with or without decimal point, after a node and preceded by a colon. This represents the length of the branch immediately below that node. Thus the above tree might have lengths represented as: (B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);" (PHYLIP, 2014).. 46, 47

nitrogenous base One of the purines (two-carbon nitrogen ring bases), A and G, or pyrimidines (one-carbon nitrogen ring bases) C and T (U in RNA), that form part of the molecular structure of a nucleotide. 1

pentose A sugar comprising five carbon atoms. 1

Robinson and Foulds (RF) topological distance It is a metric system to calculate the number of modifications needed to convert one topology to another.. 46, 57

REFERENCES

Abeysinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V. and Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer i : Nucleotide composition and recombination-associated motifs. *Human mutation*, *22*(3), 229–244.

Archer, S. Y. and Hodint, R. A. (1999). Histone acetylation and cancer. *Current opinion in genetics & development*, *9*(2), 171–174.

Bailey, J. A., Liu, G. and Eichler, E. E. (2003). An *Alu* transposition model for the origin and expansion of human segmental duplications. *The American Journal of Human Genetics*, *73*(4), 823–834.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. *et al.* (2004). The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, *91*(2), 355–358.

Barış, Ö., Karadayı, M., Yanmış, D. and Güllüce, M. (2013a). Genomic rearrangements and evolution.

Barış, Ö., Karadayı, M., Yanmış, D. and Güllüce, M. (2013b). Genomic rearrangements and evolution.

Baudis, M. (2007). Genomic imbalances in 5918 malignant epithelial tumors : an explorative meta-analysis of chromosomal cgh data. *BMC cancer*, *7*(1), 226.

Becker, K. G., Barnes, K. C., Bright, T. J. and Wang, S. A. (2004). The genetic association database. *Nature genetics*, *36*(5), 431–432.

Behe, M. and Felsenfeld, G. (1981). Effects of methylation on a synthetic polynucleotide : the b–z transition in poly (dg-m5dc). poly (dg-m5dc). *Proceedings of the National Academy of Sciences*, *78*(3), 1619–1623.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2012). Genbank. *Nucleic acids research*, p. gks1195.

Berman, H. M., Battistuz, T., Bhat, T., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S. *et al.* (2002). The protein data bank. *Acta Crystallographica Section D : Biological Crystallography*, *58*(6), 899–907.

Bernardi, G. (2005). Genome organization of vertebrates. *eLS*.

Blanchette, M. (2001). Evolutionary puzzles : An introduction to genome rearrangement. In *Computational Science-ICCS 2001* 1003–1011. Springer.

Brown, T. A. (2002). *Genomes* (2nd ed.). Wiley-Liss, Oxford.

Carbone, L., Harris, R. A., Vessere, G. M., Mootnick, A. R., Humphray, S., Rogers, J., Kim, S. K., Wall, J. D., Martin, D., Jurka, J. *et al.* (2009). Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *PLoS genetics*, *5*(6), e1000538.

Chen, C., Burton, M., Greenberger, E. and Dmitrieva, J. (1999). Population migration and the variation of dopamine d4 receptor (drd4) allele frequencies around the globe. *Evolution and Human Behavior*, *20*(5), 309–324.

Ciccarelli, F. D., von Mering, C., Suyama, M., Harrington, E. D., Izaurralde, E. and Bork, P. (2005). Complex genomic rearrangements lead to novel primate gene function. *Genome research*, *15*(3), 343–351.

Consortium, E. P. *et al.* (2004a). The encode (encyclopedia of dna elements) project. *Science*, *306*(5696), 636–640.

Consortium, I. H. G. S. *et al.* (2004b). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945.

Consortium, U. *et al.* (2008). The universal protein resource (uniprot). *Nucleic acids research*, *36*(suppl 1), D190–D195.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C. and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in burkitt lymphoma cells. *Proceedings of the National Academy of Sciences*, *79*(24), 7824–7827.

Darai, E., Kost-Alimova, M., Kiss, H., Kansoul, H., Klein, G. and Imreh, S. (2005). Evolutionarily plastic regions at human 3p21. 3 coincide with tumor breakpoints identified by the "elimination test". *Genomics*, *86*(1), 1–12.

Darai-Ramqvist, E., Sandlund, A., Müller, S., Klein, G., Imreh, S. and Kost-Alimova, M. (2008). Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome research*, *18*(3), 370–379.

Darling, A. C., Mau, B., Blattner, F. R. and Perna, N. T. (2004). Mauve : multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, *14*(7), 1394–1403.

De, S. and Michor, F. (2011). Dna secondary structures and epigenetic determinants of cancer genome evolution. *Nature structural & molecular biology*, *18*(8), 950–955.

de Gruijl, F. R., van Kranen, H. J. and Mullenders, L. H. (2001). Uv-induced dna damage, repair, mutations and oncogenic pathways in skin cancer. *Journal of Photochemistry and Photobiology B : Biology*, *63*(1), 19–27.

Diallo, A. B. (2009). *Inference of Insertion and Deletion Scenarios for Ancestral Genome Reconstruction and Phylogenetic Analyses : Algorithms and Biological Applications*. (Thèse de doctorat). McGill University.

Durkin, S. G. and Glover, T. W. (2007). Chromosome fragile sites. *Annu. Rev. Genet.*, *41*, 169–192.

Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, *4*(2), 143–153.

Felsenstein, J. (2005). Phylip (phylogeny inference package) version 3.6. distributed by the author. department of genome sciences, university of washington, seattle, 2005. *Efficiently Finding the Most Parsimonious Phylogenetic Tree*, *47*.

Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A. and Makova, K. D. (2012). A genome-wide analysis of common fragile sites : What features determine chromosomal instability in the human genome? *Genome research*, *22*(6), 993–1005.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer*, *4*(3), 177–183.

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P. and Boyle, P. (2008). Tobacco smoking and cancer : A meta-analysis. *International journal of cancer*, *122*(1), 155–164.

Ganley, A. R. and Kobayashi, T. (2008). Phylogenetic footprinting to find functional dna elements. In *Comparative Genomics* 367–379. Springer.

Gollin, S. M. (2005). Acquired chromosome abnormalities : the cytogenetics of cancer. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

Gören, E. (2014). The biogeographic origins of novelty-seeking traits. *University of Oldenburg, Department of Economics Working Papers*, *366*(14).

Gross, E., Meul, C., Raab, S., Propping, C., Avril, S., Aubele, M., Gkazepis, A., Schuster, T., Grebenchtchikov, N., Schmitt, M. *et al.* (2013). Somatic copy number changes in dpyd are associated with lower risk of recurrence in triple-negative breast cancers. *British journal of cancer*, *109*(9), 2347–2355.

Gu, W., Zhang, F. and Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics*, *1*(1), 4.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, *100*(1), 57–70.

Hardison, R. C. (2003). Comparative genomics. *PLoS biology*, *1*(2), e58.

Hengstschläger, M., Prusa, A., Repa, C., Deutinger, J., Pollak, A. and Bernaschek, G. (2005). Subtelomeric rearrangements as neutral genomic polymorphisms. *American Journal of Medical Genetics Part A*, *133*(1), 48–52.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F. *et al.* (2006). The ucsc genome browser database : update 2006. *Nucleic acids research*, *34*(suppl 1), D590–D598.

Horvath, J. E., Sheedy, C. B., Merrett, S. L., Diallo, A. B., Swofford, D. L., Green, E. D., Willard, H. F., Program, N. C. S. *et al.* (2011). Comparative analysis of the primate x-inactivation center region and reconstruction of the ancestral primate xist locus. *Genome research*, *21*(6), 850–862.

Hoskins, R. A., Smith, C. D., Carlson, J. W., Carvalho, A. B., Halpern, A., Kaminker, J. S., Kennedy, C., Mungall, C. J., Sullivan, B. A., Sutton, G. G. *et al.* (2002). Heterochromatic sequences in a drosophila whole-genome shotgun assembly. *Genome Biol*, *3*(12), 0081–0085.

Huelsenbeck, J. P., Ronquist, F. *et al.* (2001). Mrbayes : Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755.

Huret, J. L. and Senon, S. (2006). Atlas of genetics and cytogenetics in oncology and haematology.

Institute, N. H. G. R. (2013). *Human Genome Project.* Retrieved from `http://report.nih.gov/NIHfactsheets/ViewFactSheet.aspx?csid=45&key=H#H`

Kanai, Y. (2010). Genome-wide dna methylation profiles in precancerous conditions and cancers. *Cancer science*, *101*(1), 36–45.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014). Data, information, knowledge and principle : back to metabolism in kegg. *Nucleic acids research*, *42*(D1), D199–D205.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J. *et al.* (2003). The ucsc genome browser database. *Nucleic acids research*, *31*(1), 51–54.

Katapadi, V. K., Nambiar, M. and Raghavan, S. C. (2012). Potential g-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, *100*(2), 72–80.

Kazazian, H. H. (2004). Mobile elements : drivers of genome evolution. *Science*, *303*(5664), 1626–1632.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, *12*(6), 996–1006.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K. *et al.* (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*, *17*(5), 545–555.

Kost-Alimova, M., Kiss, H., Fedorova, L., Yang, Y., Dumanski, J. P., Klein, G. and Imreh, S. (2003). Coincidence of synteny breakpoints with malignancy-related deletions on human chromosome 3. *Proceedings of the National Academy of Sciences*, *100*(11), 6622–6627.

Kruisselbrink, E., Guryev, V., Brouwer, K., Pontier, D. B., Cuppen, E. and Tijsterman, M. (2008). Mutagenic capacity of endogenous g4 dna underlies genome instability in fancj-defective C.*elegans*. *Current Biology*, *18*(12), 900–905.

Kumar, S., Tamura, K. and Nei, M. (2004). Mega3 : integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in bioinformatics*, *5*(2), 150–163.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. and Maglott, D. R. (2013). Clinvar : public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, p. gkt1113.

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M. and Lewin, H. A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome research*, *19*(5), 770–777.

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). Cpg islands as gene markers in the human genome. *Genomics*, *13*(4), 1095–1107.

Lemaitre, C., Tannier, E., Gautier, C. and Sagot, M.-F. (2008). Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC bioinformatics*, *9*(1), 286.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E. and Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC genomics*, *10*(1), 335.

Letunic, I. and Bork, P. (2007). Interactive tree of life (itol) : an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128.

Library, B. M. T. (2013). Retrieved from `http://ghr.nlm.nih.gov/chromosome/1`

Lukusa, T. and Fryns, J.-P. (2008). Human chromosome fragility. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1779*(1), 3–16.

Ma, J. (2011). Reconstructing the history of large-scale genomic changes : biological questions and computational challenges. *Journal of Computational Biology*, *18*(7), 879–893.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005). Entrez gene : gene-centered information at ncbi. *Nucleic acids research*, *33*(suppl 1), D54–D58.

Maizels, N. (2005). Immunoglobulin gene diversification. *Annu. Rev. Genet.*, *39*, 23–46.

Mongin, E. (2009). *An Evolutionary Approach to Long-Range Regulation.* (Thèse de doctorat). McGill University.

Mongin, E., Dewar, K. and Blanchette, M. (2009). Long-range regulation is a major driving force in maintaining genome integrity. *BMC evolutionary biology*, *9*(1), 203.

Mongin, E., Dewar, K. and Blanchette, M. (2011). Mapping association between long-range cis-regulatory regions and their target genes using synteny. *Journal of Computational Biology*, *18*(9), 1115–1130.

Moore, J. K. and Haber, J. E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in saccharomyces cerevisiae. *Molecular and Cellular Biology*, *16*(5), 2164–2173.

Moore, S. R., Papworth, D. and Grosovsky, A. J. (2006). Non-random distribution of instability-associated chromosomal rearrangement breakpoints in human lymphoblastoid cells. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *600*(1), 113–124.

Moufatich, F. E. (2008). *Lowest Common Ancestor(LCA)*. Technische Universität München, St. Petersburg .

Mrasek, K., Schoder, C., Teichmann, A.-C., Behr, K., Franze, B., Wilhelm, K., Blaurock, N., Claussen, U., Liehr, T. and Weise, A. (2010). Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. *International journal of oncology*, *36*(4), 929–940.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L. *et al.* (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, *309*(5734), 613–617.

Nadeau, J. H. and Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*, *81*(3), 814–818.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W. and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome research*, *15*(1), 137–145.

Pelucchi, C., Gallus, S., Garavello, W., Bosetti, C. and La Vecchia, C. (2006). Cancer risk associated with alcohol and tobacco use : focus on upper aerodigestive tract and liver. *Alcohol Research & Health*.

Peng, Q., Pevzner, P. A. and Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. *PLoS computational biology*, *2*(2), e14.

Pevzner, P. and Tesler, G. (2003a). Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Research*, *13*(1), 37–45.

Pevzner, P. and Tesler, G. (2003b). Genome rearrangements in mammalian evolution : lessons from human and mouse genomes. *Genome Research*, *13*(1), 37–45.

Pevzner, P. and Tesler, G. (2003c). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences*, *100*(13), 7672–7677.

PHYLIP (2014). The newick tree format. Online. Retrieved from `http://evolution.genetics.washington.edu/phylip/newicktree.html`

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. and Jensen, L. J. (2014). Diseases : Text mining and data integration of disease–gene associations. *bioRxiv*, p. 008425.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, *20*(1), 110–121.

Pontier, D. B., Kruisselbrink, E., Guryev, V. and Tijsterman, M. (2009). Isolation of deletion alleles by g4 dna-induced mutagenesis. *Nature methods*, *6*(9), 655–657.

Raphael, B. J., Volik, S., Yu, P., Wu, C., Huang, G., Linardopoulou, E. V., Trask, B. J., Waldman, F., Costello, J., Pienta, K. J. *et al.* (2008). A sequence-based survey of the complex structural organization of tumor genomes. *Genome biology*, *9*(3), R59.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J. *et al.* (2009). The ucsc genome browser database : update 2010. *Nucleic acids research*, p. gkp939.

Rivière, G. J., Gentry, W. B. and Owens, S. M. (2000). Disposition of methamphetamine and its metabolite amphetamine in brain and other tissues in rats after intravenous administration. *Journal of Pharmacology and Experimental Therapeutics*, *292*(3), 1042–1047.

Robinson, D. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1), 131–147.

Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., Wong, M. C., Maddren, M., Fang, R., Heitner, S. G. *et al.* (2013). Encode data in the ucsc genome browser : year 5 update. *Nucleic acids research*, *41*(D1), D56–D63.

Rowley, J. D. (2001). Chromosome translocations : dangerous liaisons revisited. *Nature Reviews Cancer*, *1*(3), 245–250.

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J. and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC genomics*, *5*(1), 99.

Sankoff, D. (2009). The where and wherefore of evolutionary breakpoints. *J Biol*, *8*, 66.

Santoro, M., Chiappetta, G., Cerrato, A., Salvatore, D., Zhang, L., Manzo, G., Picone, A., Portella, G., Santelli, G., Vecchio, G. *et al.* (1996). Development of thyroid papillary carcinomas secondary to tissue-specific expression of the ret/ptc1 oncogene in transgenic mice. *Oncogene*, *12*(8), 1821–1826.

Savelyeva, L. and Brueckner, L. M. (2014). Molecular characterization of common fragile sites as a strategy to discover cancer susceptibility genes. *Cellular and Molecular Life Sciences*, *71*(23), 4561–4575.

Shapiro, J. A. and von Sternberg, R. (2005). Why repetitive dna is essential to genome function. *Biological Reviews*, *80*(02), 227–250.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, *15*(8), 1034–1050.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H. *et al.* (2002). The bioperl toolkit : Perl modules for the life sciences. *Genome research*, *12*(10), 1611–1618.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J. *et al.* (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, *462*(7276), 1005–1010.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724.

Sun, D. and Hurley, L. H. (2010). Biochemical techniques for the characterization of g-quadruplex structures : Emsa, dms footprinting, and dna polymerase stop assay. In *G-Quadruplex DNA* 65–79. Springer.

Swofford, D. L. (1998). Phylogenetic analysis using parsimony.

Talbot, S. J. and Crawford, D. H. (2004). Viruses and tumours–an update. *European Journal of Cancer*, *40*(13), 1998–2005.

Tazima, Y., Kada, T. and Murakami, A. (1975). Mutagenicity of nitrofuran derivatives, including furylfuramide, a food preservative. *Mutation Research/Reviews in Genetic Toxicology*, *32*(1), 55–80.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R. *et al.* (2005). Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, *310*(5748), 644–648.

Van Loon, A., Botterweck, A., Goldbohm, R., Brants, H., Van Klaveren, J. and Van den Brandt, P. (1998). Intake of nitrate and nitrite and the risk of gastric cancer : a prospective cohort study. *British journal of cancer*, *78*(1), 129.

Vargason, J. M. and Ho, P. S. (2002). The effect of cytosine methylation on the structure and geometry of the holliday junction the structure of d (ccggtacm5cgg) at 1.5 å resolution. *Journal of Biological Chemistry*, *277*(23), 21041–21049.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. *et al.* (2001). The sequence of the human genome. *science*, *291*(5507), 1304–1351.

Wang, Y. and Leung, F. C. (2004). An evaluation of new criteria for cpg islands in the human genome as gene markers. *Bioinformatics*, *20*(7), 1170–1177.

Watson, J. D. (2003). The secret of life. *New York : Alfred Knopf.*

Wiley, E. O. and Lieberman, B. S. (2011). *Phylogenetics : theory and practice of phylogenetic systematics.* John Wiley & Sons.

Wong, H. M., Stegle, O., Rodgers, S. and Huppert, J. L. (2010). A toolbox for predicting g-quadruplex formation and stability. *Journal of nucleic acids*, *2010.*

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K. *et al.* (2004). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology*, *3*(1), e7.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C.-Z., Wala, J., Mermel, C. H. *et al.* (2013). Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, *45*(10), 1134–1140.