The Evolution of Base Composition in Mammalian Genomes

Yves Clément

Dissertation zu Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin, August 2012

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Dr. Laurent Duret

Tag der Promotion: 21. November 2012

Pour mon père.

Acknowledgements

Disclaimer: as the collection of (too many for some, not enough for others) years of scientific work, a PhD thesis is the celebration of logic, facts, reason, deduction, induction, in one word: science. I would therefore recommend the reader to directly skip this part, as it will consist only of private jokes, obscure references and questionable humor, probably personified best by the following sentence which I collected to memory long before I sold my soul to impact factors and negative controls began to haunt my dreams but which surprisingly helped me retain some level of sanity as I wandered through the dark seas of PhD work: "A society without science is as incongruous as a fish without a bicycle" Pierre Desproges.

Where to begin? This little collection of experiments (yes, experiments) bares my name on its cover and only mine. I must however clear my conscience and spill the true fact that I could not in the name of "where have all my clean socks gone" have completed this piece of writing, infinitesimal teardrop in the ocean of humanity's knowledge, simple Mt Everest for me, without the help, support and psychological torture of a few people whose names will now cover these as short as possible lines.

First of, many thanks to past & present members of the swimming club (Sarah, Ina, Akdes, Morgane, Annalisa, Benju, Florian...) for making me discover the true meaning of the word motivation, thanks to Annalisa, my quitting partner whose Italian accent could bring sunlight to regions affected by months-long polar nights, Morgane for listening & talking and for making me realize there were more interesting publications than *Science* or *Nature*, Jonathan, any gypsy jazz fan is automatically a friend of mine, Julia L, who brought a new meaning to the verb "to wait", Julia G, living proof that people from different departments can be friends, Tina and her tiramisu, the Kicker crew (Alessandro, Masha, Annalisa, Stefan, Juliane), I had loads of fun trying to pretend I didn't care about loosing and I hope my swearing wasn't too much to bear, Mike, respectable bear contest opponent and recipient of next year's "most unlikely name in science" award, Roland, for his interest on french politics and for, and this is serious, introducing me to Karl Popper whose work I would like to influence me if I had any motivation at all.

In short, thanks to all past & present members of the Vingron Department: Matt, Stefan, Ruping, Stephanie, Jonas, Stephan, Sarah, Akdes (who could smile in the darkest hours of a Berlin winter), Holger and his music, Corinna, many more but my memory is failing me unfortunately.

Thanks to past & present ~~slaves of Peter~~ EvoGeners for making these years under the ~~tyrant ruling~~ exciting leadership of Peter ~~barely livable~~ some of the best of my life. Thanks to Federico and his beautiful mind, Paz and to the discussions we could always have, his radical approach to scientific communication (going as far as inventing a new language) was truly an inspiration, Barbara who helped me countless times with math-related problems and her finely distilled words of support of advice, and whom I paid back with cursing in french lessons, Brian and his "ass-kicking" approach to science, which helped me so many times and for his priceless english corrections, Irina, probably the only other lab member who laughs as loud as I do, Florian, to whom I shall come clean and admit the only reason I hang out with is because he, like me, speaks french.

Thanks to past & present office mate, Federico, Paz, Barbara, Florian, Rosa, one of the sanest crazy person I know and walking add for the movie "Coffee & Cigarettes".

Thanks to many people in the IMPRS, Martin Vingron on the top of my list for giving me a chance, for wasting money on me all these years and whose working knowledge of biology makes it the first time a biologist whish he knew as much about biology as a mathematician, Sharon, Marcel,

Ewa, Alena, Christian, Anne-Katrin and all the students there. Many thanks to the most awesome IMPRS-CBSC coordinators in the world, Hannes (who I deeply miss) and Kirsten. Thanks for helping me so many times on so many different issues.

I am very grateful to the various people who cut into their precious free time to proofread this thesis: Jonathan, Barbara, Mike, Annalisa and Kirsten. Thanks for all your comments, suggestions and ideas without which this thesis would have been much duller than it already is.

I would like to present my deepest apologies to all the people around me that I offended, insulted, ran into without acknowledging or just was quite mean to these last few months. The thesis writing left me in a strange state where I became unaware of any possible damage done to any form of life around me (the plant in my office unfortunately cannot testify on this), let alone human beings.

Sorry.

Let me also utter a few words about a certain individual. He haunts the halls of the institute and goes by the name of Peter F Arndt (if that's his real name). What to say about him? Well, apart from the fact that I am thankful he gave me a chance to work for him and with him when I was a more than questionable choice for a PhD candidate, that his openness for discussion and ideas has made this time more than interesting, that his guidance and advices always helped me navigate through the muddy waters of science, that his coffee machine has done so much for science it should get a Nobel Prize, apart from all that, I am afraid not much can be said.

Thanks to friends here in Berlin and all around the world, An honorary academy award goes to David M for the best many things, thanks to David F, my accomplice in crime during many Berlin nights, Benjamin and our pale imitations of the french tennis open. Many, many thanks to the fan club of cells that kill themselves, with whom I shared lab benches and drinks at university and with whom I made the most terrible career choice someone like me could possibly make, I hold these people responsible for what is happening to me right now. I thank all of you guys for all the fun we had, all the crazy moments and all the support in dark times.

And finally...

Mamy, JP, Isa, Marion et Julie, Philippe et Sylvie, Laurence, Philippe et les filles, je vous remercie pour un soutien sincère et constant. Enfin, mille remerciements à ma soeur Anne, à Laurent, Louane et Emeline et a ma chère Maman, merci du fond du coeur pour m'avoir inspiré, aidé, supporté parfois à bout de bras toutes ces années, sachez que si cette thèse existe, c'est en grande partie grace à vous.

And now for something completely different...

Contents

List of Figures

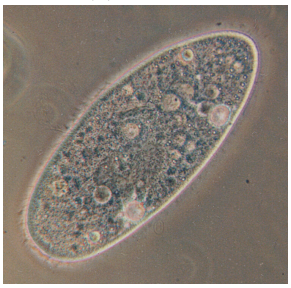
List of Tables

1. Introduction

A striking feature of living organisms is how diverse they are. They adapted to amazingly different and hostile environments and exhibit great variety in shape, size and lifestyles (Figure 1.1). All instructions necessary to generate a living organism and maintain its life is encoded in DNA sequences organized into a genome, which is transmitted from one generation to the next to allow new organisms to be formed. Studying an organism's genome can thus tell us how this organism functions. As more and more genomic sequences are collected and available for analysis, it is important to understand how genomes are organized, for example which regions are functional and what that function might be. Studying how genomes evolve through time can help identifying such regions and also highlight mechanisms acting on the evolution of genomes. As DNA sequences consist of four bases (A, T, G and C) grouped together in a double stranded molecule, the most basic property of any DNA sequence is its base composition, the proportion in the sequence of As, Ts, Gs and Cs. As such, genomic base composition has been studied for a long time and revealed evolutionary forces acting on genome evolution. This thesis presents analyses of base composition evolution in animal genomes and of the different mechanisms affecting it.

(a) House mouse

(b) Fruit fly



(c) Paramecium

Figure 1.1.: Pictures of animals. Pictures downloaded from wikipedia.org.

1.1. General introduction

1.1.1. Introduction to DNA and genomes

The genome is the recipe book to an organism: it contains the information needed to build an organism from a single-cell embryo to a multicellular adult in the case of multicellular organisms, or more generally to maintain the organism's function. The molecule containing all the genome's information is DNA (short for DeoxyriboNucleic Acid).

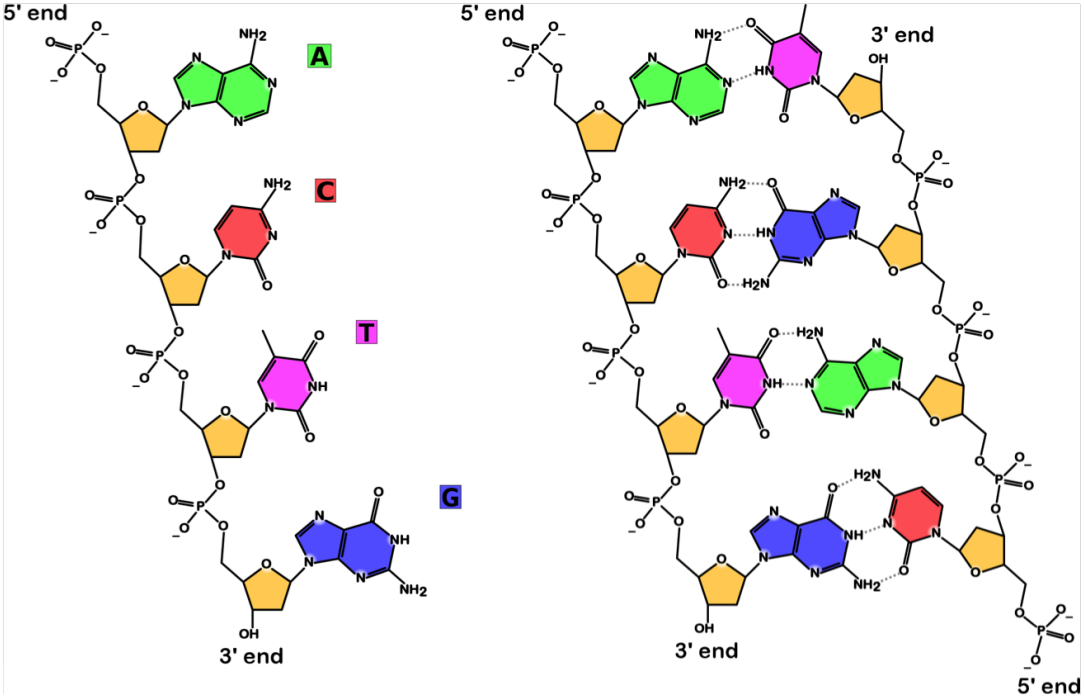


Figure 1.2.: Right panel: one strand of DNA. The deoxyribose groups are represented in orange and linked together by phosphate groups form the backbone of the DNA strand. The four types of bases are represented in green (A), red (C), purple (T) and blue (G). One can observe the 5' end of the strand formed by the phosphate group while the 3' end is formed by the OH residue of the deoxyribose. Left panel: double-stranded DNA. The DNA bases of each strand pair together through hydrogen bonds. A and T bases pair with two bonds while G and C bases pair with three. The two strands pair in an antiparallel manner: the 3' end of one strand will pair with the 5' end of the other. Figures adapted from wikipedia.org.

The chemical structure of DNA was determined nearly 60 years ago (Franklin and Gosling, 1953; Wilkins et al., 1953; Watson and Crick, 1953). One molecule of DNA consists of a backbone of sugar molecules (in this case deoxyribose) linked together by phosphate groups (Figure 1.2). Attached to this backbone are nucleobases, which will constitute the actual sequence of the DNA. There are four nucleobases found in DNA: cytosine (C) and thymine (T), two pyrimidines, and adenine (A) and guanine (G), two purines. Such association of sugar, phosphate groups and nucleobases (or bases for short) form a DNA strand. The bases of this strand will be decoded by the cell machinery and constitute the genetic information. This strand has an orientation, which is based on what is at its end: a phosphate group on the so-called 5' end

and a hydroxyl group from a deoxyribose on the 3' end. By convention, a strand is read and analyzed from the 5' end to the 3' end (Figure 1.2).

In living cells, DNA is not single-stranded. Two complement strands pair together to form a double-stranded DNA (Figure 1.2). This pairing is possible because bases on opposing strands pair together through hydrogen bonds: A bases pair with T bases with two hydrogen bonds while C bases pair with G bases with three hydrogen bonds (Figure 1.2). The two strands pairing up are anti-parallel: one is oriented from 5' to 3' while the other is oriented from 3' to 5'. The end product will form a double helix structure.

The fact that double-stranded DNA is made of two complement single strands has important consequences for the transmission of genetic material. During replication of DNA, the two strands are denatured or separated. Each strand will then serve as template for the synthesis of its complement strand. We will end up with two identical double-strands of DNA, which can then be transmitted to daughter cells during cell division.

The DNA contains the information necessary to the formation of an organism, either in protein-coding genes or in regions regulating the activity of such genes. The entire genetic information of an organism constitutes its genome. In eukaryotes, this genome is usually divided into chromosomes, each containing several million bases.

1.1.2. Introduction to evolution

Molecular evolution can be simply viewed as mutations occurring in the genetic material being transmitted from one generation to the next. These mutations happening in one individual can then affect entire populations and species.

Mutations can be of several types and scales (Figure 1.3). The simplest mutation is one single DNA base being replaced by another. A slightly more complex mutation occurs when one or several DNA bases which are either inserted in or deleted from a locus. These are referred to as indels. A segmental duplication occurs when a large region (e.g. several kbp to several Mbp) is copied and inserted in the same chromosome. Mobile DNA elements or transposable elements are elements that have the capacity to copy themselves and insert this copy elsewhere in the genome, thus affecting the locus in which they insert themselves. Finally, large section of chromosomes (like entire chromosomal arms) can translocate from one chromosome to another, causing chromosomal translocations or rearrangements. All these events happen at different rates, the single nucleotide replacements occurring at much higher frequencies than chromosomal rearrangements.

When a mutation occurs in one individual, it will carry the mutated allele and possibly transmit this allele to its offspring through reproduction. These offspring will in turn reproduce and transmit the allele to the following generations. By this process, the allele will spread in the population. When all individuals in a population possess the allele, it has reached fixation. The probability of a mutation to become fixed in a population depends on several things: the population size, the frequency

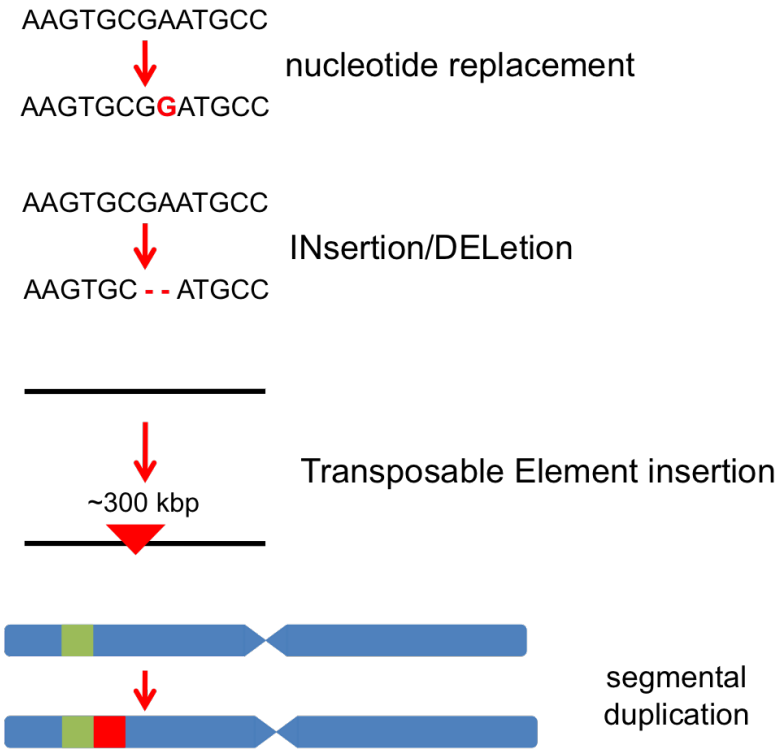


Figure 1.3.: Different types of mutations

of the mutated allele in the population (the proportion of individuals carrying this allele), and its selective coefficient. The latter represents how advantageous or disadvantageous an allele is. It is generally computed by comparing the number of offspring produced by individuals with a mutation to the number of offspring of individuals without. It is usually denoted s. A mutation without any effect on an individual's capacity to produce offspring will have a s of 0 and will be called a neutral mutation. When considering a population of diploid organisms of size N, we will have $2N$ copies of the genome. A newly arising mutations will have a frequency of

$$f = \frac{1}{2N} \ . \tag{1.1}$$

When a mutation occurs at a locus in an individual, there will be several alleles, or variants, of this locus observable in the population. The original allele present before any mutation is defined as the ancestral allele whereas the allele arising from a mutation will be defined as a derived allele. The fixation probability of a neutral allele will be equal to its frequency in the population (Kimura, 1968; Felsenstein, 2005). Therefore that of a newly arising mutation will be

$$P_{\text{fixation}} = f = \frac{1}{2N} \ . \tag{1.2}$$

When the mutation has a selective advantage or disadvantage, its fixation prob-

ability is then

$$P_{\text{fixation}} = \frac{1 - e^{-4Nsf}}{1 - e^{-4Ns}} \ , \tag{1.3}$$

where s is the allele's selection coefficient and f its frequency in the population (Kimura, 1968; Felsenstein, 2005). The sign of s will indicate whether the allele is deleterious (s is negative) or advantageous (s is positive). The fixation probability of a newly arising mutation will be

$$P_{\text{fixation}} = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \ . \tag{1.4}$$

A mutation that has reached fixation in the population will be called a substitution (Figure 1.4).

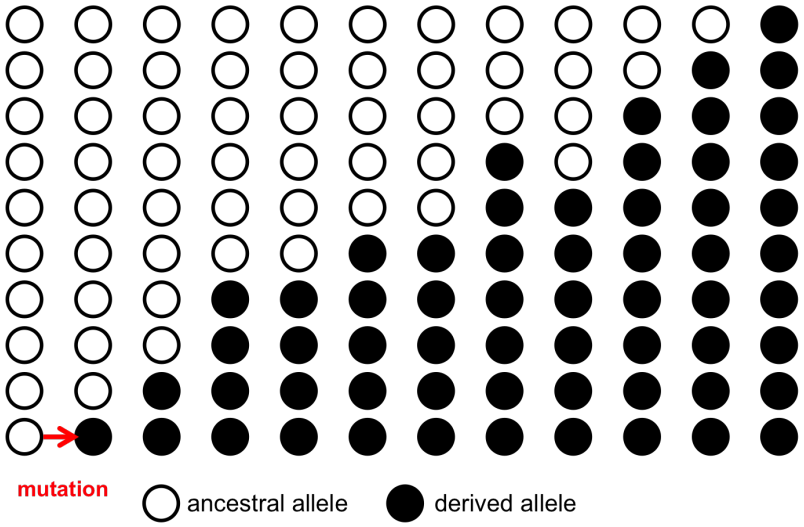


Figure 1.4.: Schematized fixation of a mutation in a population. The frequency of the derived allele increases until it reaches fixation when all individuals have this allele.

When we study the evolution of organisms and genomes, we want to find past mutations in genomic sequences. However, it is impossible to access past species and DNA sequences. The issue is then to be able to reconstruct past events only from information about the present. This is resolved by taking a comparative approach. The principle is to compare DNA sequences from several species of interest, identify similarities and differences and based on this information, reconstruct ancestral DNA sequences and changes that occurred between them and present-day sequences (Figure 1.5).

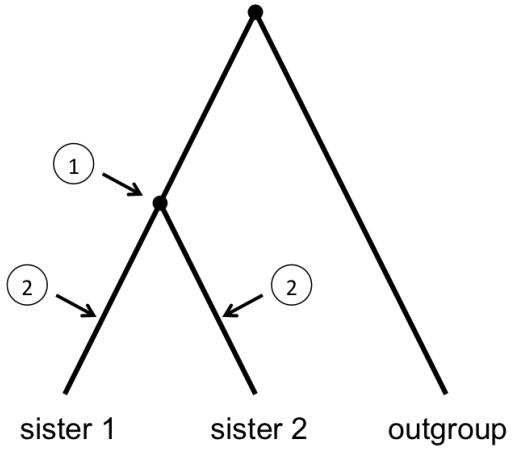


Figure 1.5: Comparative approach. By comparing the sequences of two sister species and a more distantly related outgroup, one can try to infer, given some assumptions on evolutionary processes, 1) the ancestral state of the two sister species and 2) the events in each of their branches.

1.2. Isochore structures

1.2.1. Large-scale variations of GC-content in mammalian genomes

DNA molecules are studied since the discovery of its double helix structure by Rosalind Franklin, Maurice Wilkins, James Watson and Francis Crick about 60 years ago (Franklin and Gosling, 1953; Watson and Crick, 1953; Wilkins et al., 1953). The first sequences of nucleic acid sequences (RNA or DNA) were obtained in the early seventies. Techniques were however costly, time and labor demanding and thus unpractical for genomic sequences. Thus, early characterizations of DNA sequences was done using more classical biochemical analyses.

A major property one can determine of DNA sequences is their base composition, the number or frequency or relative abundance of bases. A bases pair up with T bases while G and C bases pair together in double-stranded DNA. In a genomic DNA sequence, the number of G and C bases will be equal, a feature shared by A and T bases (this is known as Chargaff's first parity rule). Only the abundance of AT bases or GC bases needs to be quantified in genomic DNA, which can be further reduced to the quantification of one of these (the added frequency of all four bases is always 1). The GC-content was historically chosen as subject of measurement.

The method most widely used to study the GC-content of DNA sequences was ultra-centrifugation, first described by Meselson et al. (1957). Its principle was to prepare a density gradient of a certain medium that would separate after a centrifugation step DNA sequences according to their GC-content. Using this method, it was possible to determine the GC-content of genomic DNA sequences but also that of smaller fragments to reveal variations in GC-content within the genomic sequence. After chemically fragmenting genomic DNA, it was possible to characterize the GC-content of these fragments and generate GC-content distributions for mammalian genomes (Filipski et al., 1973; Macaya et al., 1976; Thiery et al., 1976; Cortadas et al., 1977; Macaya et al., 1978). These analyses revealed that the

GC-content distributions of mammalian genomes showed a non-homogeneous distribution with an excess of GC-rich regions. Moreover, the authors concluded that mammalian genomes were organized into isochores (Cuny et al., 1981): regions of approximately 200 kbp having a relatively homogeneous GC-content. More specifically, mammalian genomes could be divided into several individual and independent classes of isochores: L1 & 2 (Low), regions of low GC-content and the majority of mammalian genomes, and H1, 2 & 3 (High), regions with high GC-content and only a small fraction of mammalian genomes (Bernardi et al., 1985).

The initial sequencing of human and mouse genomes (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002) allowed for direct analysis of mammalian genome sequences. More importantly, the GC-content variations along chromosomes became obvious. Figure 1.6 shows GC-content computed in 10 kbp windows in a 3 Mbp region of the human genome. One can easily notice how the GC-content varies along this segment. Several questions are raised when studying such profiles: is there a simple way to quantify GC-content variations along chromosomes, what can we expect from random DNA sequences, is there a simple way to quantify GC-content variations along chromosomes and can we use such method to compare the genomes of different organisms.

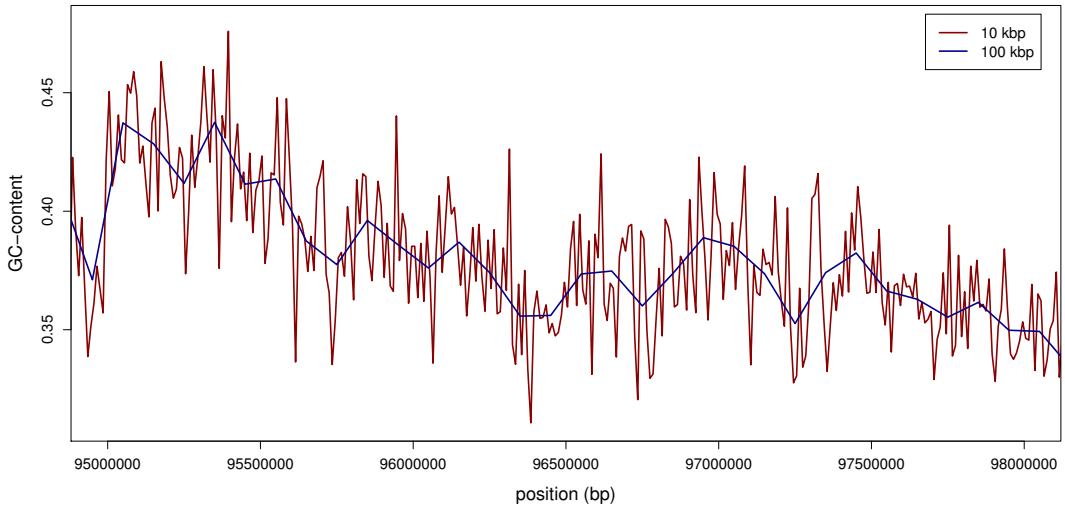


Figure 1.6.: GC-content profile in 10 and 100 kbp windows along a 3 Mbp segment of the human chromosome 1.

1.2.2. Association of isochore structures with genomic features

While it became clear that the isochore structure was an important feature of a genome, the association between GC-content and several genomic features was studied. Initially, the distribution of a handful of genes with respect to isochore classes was analyzed (Bernardi et al., 1985). It was found that genes were not homogeneously

distributed along mammalian genomes with respect to GC-content: GC-rich regions while representing 8% of the total genomic DNA contained half of the 24 genes studied at the time (Bernardi et al., 1985). Complete genome sequencing and gene annotation showed this tendency for genes to lie in GC-rich regions was widespread in mammalian genomes (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002; Gibbs et al., 2004). Moreover, it was shown that the length of genes was also linked to neighboring GC-content, with genes in GC-rich isochores being on average shorter than genes in other regions (Duret et al., 1995).

Transposable elements, DNA elements that have the ability to move around the genome (they had been dubbed 'jumping genes' early on), were found to have their distribution also linked to GC-content variations. Preliminary sequences of the human genome showed that different classes of elements were found in regions of different base composition along human sequences (Smit, 1999). Notably, short interspersed elements (hereafter designated as SINEs) like Alu elements were found mostly in GC-rich regions whereas long interspersed elements (hereafter designated as LINEs) were mostly found in GC-poor regions of the human genome. Subsequent analysis of the human and mouse complete genomes confirmed this observation (Lander et al., 2001; Mouse Genome Sequencing Consortium et al., 2002).

The link between GC-content and other processes, like meiotic recombination (the exchange of genetic material between homologous chromosomes during meiosis) or replication-timing (the relative time at which a particular region is duplicated) has been studied over the years. It has been shown that the number of chiasmata (points of exchange between two homologous chromosomes) is linked with GC-content in mammalian genomes (Eyre-Walker, 1993). More detailed measures of meiotic recombination in human genomes further confirmed the link between meiotic recombination and base composition variations (Fullerton et al., 2001; Kong et al., 2002).

Furthermore, when measuring timing of replication in mammalian cells, it was found that whether a regions is replicated early or late is linked with its base composition (Federico et al., 1998; Watanabe et al., 2002; Touchon et al., 2005; Costantini and Bernardi, 2008; Hiratani et al., 2008; Ryba et al., 2010).

The association between a number of genomic features and isochores structures led to the hypothesis that these structures had a level of importance for the genome's function. The question of their origin and maintenance was raised and tested, which led to the identification of evolutionary forces acting on mammalian genomes.

1.3. Origin and evolution of isochore structures

We know there are variations in the GC-content along mammalian and bird chromosomes that have been called isochore structures and that these structures are linked with a number of genomic features. We now ask the question of the origin of such variations: how did they came into place, what mechanisms were responsible for this and are these mechanisms still at work today. Three hypotheses have been put for-

ward to explain these variations: a bias in the mutation process, natural selection or a neutral process called GC-biased gene conversion (hereafter designated as gBGC). We will now review these hypotheses.

1.3.1. Mutational bias

According to the mutational bias hypothesis, isochore structures are caused by a variation along chromosomes of the bias of mutations towards AT or GC nucleotides (Wolfe et al., 1989; Filipski, 1988). This hypothesis can be tested by looking at single nucleotide polymorphisms (hereafter designated as SNPs). Such analysis in human and mouse revealed a weak bias in mutations favoring AT nucleotides and a strong fixation bias favoring GC nucleotides. These mutations spread more rapidly in populations than AT mutations (Eyre-Walker, 1999; Spencer et al., 2006).

These studies show that the mutational bias hypothesis alone cannot explain isochore structures along mammalian chromosomes and reveal a fixation bias favoring G and C alleles. This bias could be caused by natural selection.

1.3.2. Natural selection

Several hypotheses explaining isochore structures by natural selection have been proposed in the past. The main one confers a selective advantage of GC-rich regions in warm-blooded organisms like birds and mammals. There are three hydrogen bonds between a G and a C base whereas there are only two between A and T bases. As the two strands of DNA separate more easily at higher temperature, GC-rich sequences thus ensure more stability for DNA and RNA in warm-blooded organisms. Having a GC-rich genome will therefore be a clear selective advantage for such organisms (Bernardi, 2000, 2007). The fact that isochore structures were first discovered in birds and mammals, warm-blooded organisms, was also interpreted as evidence supporting this hypothesis (Bernardi, 2000).

However, analyses in cold-blooded organisms like crocodiles or turtle revealed isochore structures comparable to warm-blooded structures (Hughes et al., 1999; Hamada et al., 2002, 2003). Two conclusions were drawn from these results. First, as mammals, birds and reptiles share isochore structures, it is very likely that these appeared once in the ancestor or amniotes about 350 million years ago (Duret et al., 2002). Second, as cold-blooded organisms do have isochore structures, these do not bring a selective advantage to warm-blooded organisms. The hypothesis of adaptation to high body temperature therefore has to be rejected.

Finally we should consider the following. If we imagine a region of a certain length (e.g. 1 kbp) where a GC-changing mutation occurs, the region's GC-content will only change marginally. Because of such limited effect, it is very unlikely that this mutation will be seen by natural selection: the selective advantage of such mutation will be very low. For this reason, it is hard to imagine how GC-content is effectively under selective pressure.

Overall, these studies show that natural selection alone cannot explain the fixation bias favoring GC alleles nor isochore structures in amniotes. This means that a neutral process like gBGC was likely to be at the origin of these structures.

1.3.3. GC-biased gene conversion

The GC-biased gene conversion model has first been proposed by Galtier et al. (2001). This model states that meiotic recombination will increase the fixation probability of G and C alleles through a biased repair of mismatches occurring during gene conversion events. Over the years, it has gained substantial supporting evidence.

The GC-biased gene conversion model

The different molecular pathways of meiotic recombination have been well described in eukaryotes (Figure 1.7, see Marais, 2003 and Coop and Przeworski, 2007 for reviews).

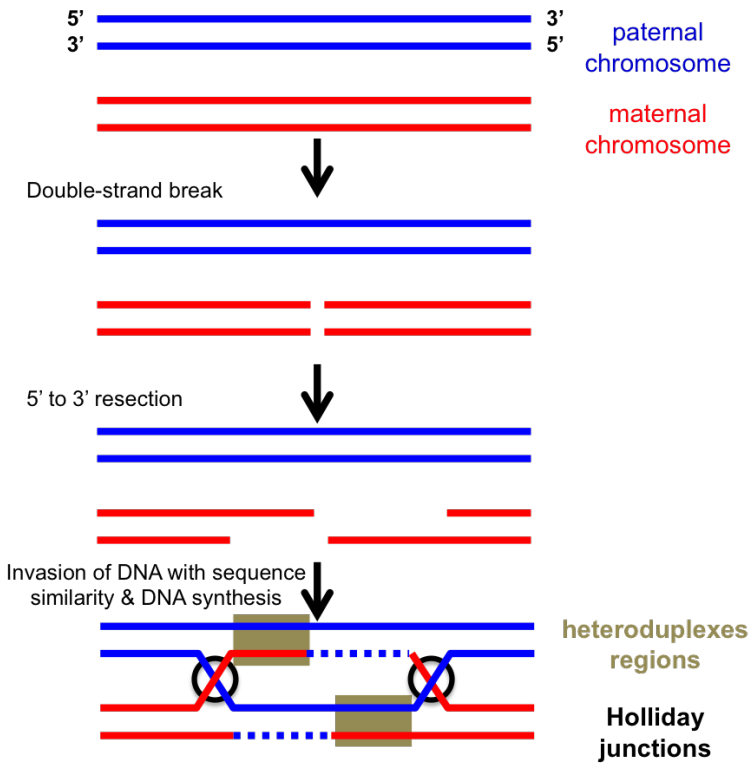


Figure 1.7.: Meiotic recombination. The repair of the double strand break will form Holliday junction which will be repaired into crossover or non-crossover events.

This process starts with a double strand break (hereafter designated as DSB) in one chromosome of a chromosomal pair, which product is digested and then repaired

by the invasion of the homologous region of the sister chromosome. This region will be used as a template for DNA synthesis and repair by gene conversion, the copy and paste of one DNA fragment into another (de Massy, 2003; Chen et al., 2007).

During this process, strands from two sister chromosomes are paired together, which may result in mismatches occurring if the corresponding locus is heterozygote. It has been shown that in mammals the mismatch repair mechanism is biased towards G and C bases: it will repair for example a G:T mismatch more often into G:C than into A:T (Figure 1.8, Brown and Jiricny, 1988; Bill et al., 1998).

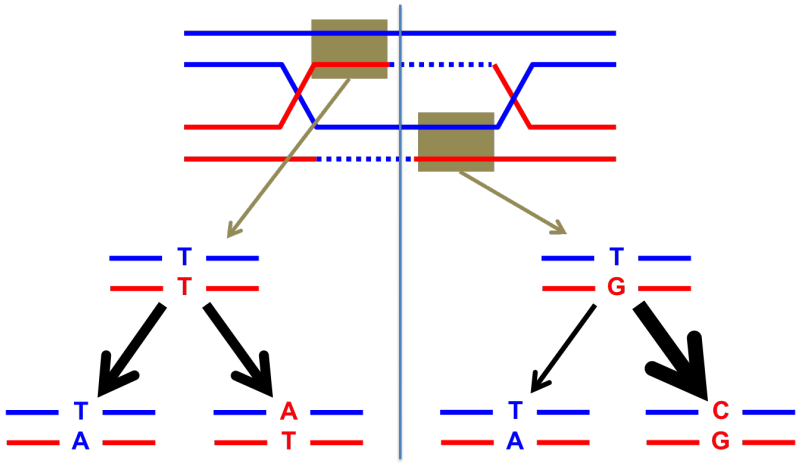


Figure 1.8.: Two cases of mismatch repair. Left: in the case of a T:T mismatch, the repair process is unbiased, both resolutions will have the same frequency. Right: in the case of a T:G mismatch, the biased repair process will lead to this mismatch being repaired more often into C:G than into A:T.

This biased repair will have important implications for population genetics as it will lead to an unequal segregation of alleles, G and C alleles segregating at higher frequencies than A and T alleles. This will result in a fixation bias (Nagylaki, 1983) favoring G and C alleles and disfavoring A and T alleles, which increases with meiotic recombination (for a review on gBGC, see Duret and Galtier, 2009). How much gBGC will affect substitution rates and genome evolution will depend on the length of gene conversion tracts, how often an A or T base is converted into a G or C base, and on the effective population size (Lynch, 2007). This model predicts a positive link between meiotic recombination and GC-content as well as a positive influence of meiotic recombination on substitutions that increase GC-content.

Evidence of GC-biased gene conversion affecting mammalian genomes

The link between GC-content and meiotic recombination has been put forward in various studies. First, a positive correlation between local rates of recombination and GC-content has been found in the human genome (Fullerton et al., 2001). Second, studies on the *Fxy* gene in mouse have shown an increase of GC-content associated with high recombination rates (Montoya-Burgos et al., 2003; Galtier and Duret, 2007). The 3' side of the gene underwent a translocation in the pseudoautosomal

region of the X chromosome (a region which experiences very high levels of meiotic recombination) about 1 to 3 million years ago in the *Mus musculus musculus* (or *Mus m. musculus* for short) lineage. This 3' side of the *Fxy* gene accumulated an extremely high number of non-synonymous substitution, all AT to GC, over this short period of time, whereas the same gene in different species where no translocation occurred only experienced a handful of non-synonymous substitutions, all distributed along the entire coding sequence (Montoya-Burgos et al., 2003; Galtier and Duret, 2007). These results highlight a positive link between meiotic recombination and GC-content, which is a prediction of the gBGC model.

The second prediction of the gBGC model is the influence of meiotic recombination on substitution patterns. These can be inferred by comparing homologous sequences, either orthologous sequences from closely related species or repeated elements such as transposable elements.

Studies of substitution patterns in non-coding regions of the human genome have shown a positive correlation between substitution patterns and crossover rates, a proxy measure of meiotic recombination (Meunier and Duret, 2004; Duret and Arndt, 2008). Similar studies of substitution patterns in retropseudogenes in human, rat and human (Khelifi et al., 2006) as well as in Alu retroelements in human (Webster et al., 2005) show a similar correlation. These studies analyze equilibrium GC-content or future GC-content (hereafter designated as GC*), a value computed from substitution patterns which represents the final GC-content value the sequences of interest evolve to, provided substitution patterns stay constant over time. This value is proportional to the ratio of A or T (Weak or W) to G or C (Strong or S) substitution rates divided by the sum of S $\rightarrow$ W and W $\rightarrow$ S substitution rates. It represents the relative importance of W$\rightarrow$S rates compared to S$\rightarrow$W rates:

$$\mathrm{GC}^* = \frac{(\mathrm{W} \rightarrow \mathrm{S})}{(\mathrm{W} \rightarrow \mathrm{S}) + (\mathrm{S} \rightarrow \mathrm{W})} \ . \tag{1.5}$$

Crossover rates are positively correlated with both GC-content and GC*. However, these rates are more strongly correlated with GC* than with GC-content. As GC* values are computed from substitution patterns and not from current GC-content, these results show that meiotic recombination influences GC-content evolution through the influence of substitution patterns, which is predicted by the gBGC model.

GC-biased gene conversion and natural selection

The GC-biased gene conversion process is a fixation bias: when meiotic recombination is high, G and C alleles spread more easily through the population than A and T alleles and thus will get fixed more rapidly. This fixation bias is neutral (it does not affect an organism's fitness, its ability to reproduce) but will look like natural selection: first evidence of a fixation bias favoring G and C alleles was interpreted as resulting from natural selection (Eyre-Walker, 1999). More recently, this process

was mistaken for positive selection in fast evolving regions of the human genome. Pollard et al. (2006b) scanned the human genome for regions that were conserved in a number of organisms but showed human-specific substitutions. These regions were called human accelerated regions (or HARs for short) as they showed a significant acceleration of evolutionary rates specific to the human lineage. Their functions were tested and some were found to code for RNA genes expressed during cortical development (Pollard et al., 2006b). However, most changes in these HARs are W to S changes (Pollard et al., 2006a), a clear signature of gBGC, which questioned the fact that natural selection was the sole factor influencing the evolution of HARs. Subsequent analyses confirmed that the most accelerated regions were enriched for A or T to G or C substitutions and showed that these regions were more likely to be found near recombination hotspots, regions of the genome experiencing extremely high amounts of recombination (Berglund et al., 2009). This suggested that gBGC had a strong role in affecting fast-evolving regions of the human genome. It also showed that signatures of gBGC can easily be identified as positive selection, and disentangling signatures of gBGC and positive selection is an issue that was addressed very recently (Kostka et al., 2012).

One can ask whether a neutral process like gBGC can be stronger than natural selection in some cases. Recombination is not homogeneously distributed along mammalian genomes, but is concentrated in hotspots, regions of 1 to 2 kbp experiencing high amounts of recombination (Myers et al., 2005; Paigen and Petkov, 2010). Clear and strong signatures of gBGC have been shown inside meiotic recombination hotspots (Spencer et al., 2006; Katzman et al., 2011). It is possible that inside these hotspots gBGC is strong enough to promote the fixation of deleterious alleles (i.e. alleles that are a disadvantage for the individuals possessing them). When analyzing different primates lineages for deleterious amino-acid changes caused by A or T to G or C substitution, it has been shown that such changes happened at the same time as an increase of GC-content in both synonymous positions and introns of the studied genes (Galtier et al., 2009), which was interpreted as gBGC promoting the fixation of deleterious alleles in primates. Furthermore, by studying the frequency of various alleles known to cause diseases, it was shown that these alleles segregate at higher frequencies when they lie close to recombination hotspots (Necşulea et al., 2011). Both examples show that gBGC can be stronger than natural selection and promote the spread and fixation of deleterious alleles, especially in regions experiencing high amounts of recombination.

Finally, evidence of gBGC affecting substitution patterns has been found in a wide number of organisms: mammals (Romiguier et al., 2010), drosophila and other metazoans (Galtier et al., 2006; Capra and Pollard, 2011), yeast (Harrison and Charlesworth, 2011) and other eukaryotes like angiosperms (Escobar et al., 2011). This indicates that gBGC is an important process in eukaryote genome evolution. Overall, it is now clear that gBGC is a major process influencing mammalian genome evolution.

However, some questions about gBGC and GC-content evolution are still unan-

swered despite recent discoveries. First, can we quantify and compare GC-content variations across different animal species. Second, gBGC has been well characterized only in primates. Measuring in detail the influence of gBGC on substitution patterns across the mouse genome will give us a more complete picture of how gBGC works in mammals. This could be completed by a measure of the relative influence of different processes on substitution patterns in mouse and human. Third, what is the fine-scale signature of gBGC inside meiotic recombination hotspots in mouse and human. What can this signature tell us about how meiotic recombination is working in mammalian genomes.

1.4. Thesis outline

In this thesis, we study the base composition of organisms, focusing on mammalian genomes. We try to first identify and then quantify processes affecting its evolution. Chapter 2 lists and details the experimental procedures used in subsequent chapters, notably the different sets of alignments and datasets used, as well as the maximum likelihood framework used to infer substitution patterns in different sets of alignments. The third chapter presents a detailed quantification of GC-content variations along a wide range of genomes. We determine GC-content variations for random DNA sequences and compare these results to those for the human genome. We then characterize GC-content for different groups of organisms like mammals, primates or amniotes. Chapter 4 presents analyses of substitution patterns and GC-content evolution in both mouse and human genomes. We measure substitution patterns from triple alignments in these lineages to compare GC-content evolution as well as the activity of GC-biased gene conversion. We also compute the relative influence of various genomic features on substitution patterns, highlighting major differences between the two lineages. Chapter 5 analyzes GC-content evolution and substitution patterns in the context of meiotic recombination hotspots in both human and mouse genomes. We determine GC-content evolution in a wide range of mouse lineages. We then measure the fine-scale evolutionary signature of meiotic recombination in both human and mouse genomes. Finally, from these signatures, we derive characteristics of meiotic recombination in mammalian genomes.

Publications Parts of Chapter 4 appeared in a publication in *Genome Biology and Evolution* (Clément and Arndt, 2011).

2. Materials & Methods

2.1. GC-content variance

We present here methods used in Chapter 3. Base composition of genomes was analyzed in several groups of species. In primates, the following genomes: human (*Homo sapiens*, version hg19), chimpanzee (*Pan troglodytes*, version *panTro3*) gorilla (*Gorilla gorilla*, version *gorGor3*), orangutan (*Pongo abelii*, version *ponAbe2*), macaque (*Macaca mulatta*, version *rheMac2*) and marmoset (*Callithrix jacchus*, version *calJac3*); in rodents, mouse (*Mus m. musculus*, version *mm9*) and rat (*Rattus norvegicus*, version *rn4*); in mammals, horse (*Equus caballus*, version *equCab2*), dog (*Canis lupus familiaris*, version *canFam2*), pig (*Sus scrofa*, version *susScr2*), cow (*Bos taurus*, version *bosTau6*), opossum (*Monodelphis domestica*, version *monDom5*) and platypus (*Ornithorhynchus anatinus*, version *ornAna1*); in birds, chicken (*Gallus gallus*, version *galGal3*), zebra finch (*Taeniopygia guttata*, version *taeGut1*) and turkey (*Meleagris gallopavo*, version *melGal1*) were analyzed. This was completed by the study of an amniote with anole lizard (*Anolis carolinensis*, version *anoCar2*), and with the study of zebrafish (*Danio rerio*, version *Zv9*) and drosophila (*Drosophila melanogaster*, version *dm3*).

For each genome, sequences for autosomes and sex chromosomes were downloaded from the Ensembl database (version 65, Flicek et al., 2012).

For each species of interest, the genomic sequence was first divided into non-overlapping windows of different sizes (hereafter designated as tiling): 1, 2, 5, 10, 20, 50, 100, 200, 500 kbp and 1 Mbp. For some genomic intervals the precise nucleotide sequence cannot be determined (e.g. centromeric or telomeric regions). To ensure that windows contain enough information to compute GC-content, only those containing at least 50% of nucleotides (e.g. 50 kbp of nucleotides in a 100 kbp window) were kept for analysis. The GC-content was then computed in each window. For each tiling the variance for the GC-content values of all windows was finally computed. At the same time, GC-content distributions for 10 kbp tiling were plotted.

2.2. Substitution patterns in primates and rodents

We present here methods used in Chapter 4. The link between substitution patterns and genomic features was studied in human and mouse genomes. To do so, we took a whole-genome approach. For each species the reference genome was divided into non-overlapping windows, then in each of these substitution patterns were computed

from multiple alignments available online. At the same time we measured in each window genomic features from publicly available datasets.

2.2.1. The maximum likelihood framework

Computing substitution rates in different lineages is done using a comparative approach, for example by studying sequence alignments of a region of interest. As sequence alignments only represent present-day states, the methodological challenge is to reconstruct past events from these alignments.

The simplest, most intuitive method to infer substitution rates in different lineages is maximum parsimony: finding the lowest number of substitutions to explain the data. However, using such simple method will cause some problems, specifically in cases where the base composition is not at equilibrium (Felsenstein, 1978; Eyre-Walker, 1998). Using a maximum likelihood-based method will yield better results (Felsenstein, 1981). In a maximum likelihood framework, one wants to estimate a model (M) that best explains data (D). The following function gives the likelihood (L): the probability of observing the data given the model.

$$L = P(D|M) \tag{2.1}$$

The goal is to estimate the model that maximizes the likelihood. In our case, the data consists of multiple alignments of DNA sequences and the model is parameterized by substitution frequencies along different branches of a tree. Substitution probabilities are usually represented by a rate matrix (denoted Q) where every entry $Q_{\beta\alpha}$ is the $\alpha \rightarrow \beta$ transition probability for an infinitesimal time interval and α and β are nucleotides. Below is an example of a rate matrix for the simple one parameter Jukes-Cantor model (Jukes and Cantor, 1969).

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -3q & q & q & q \\ q & -3q & q & q \\ q & q & -3q & q \\ q & q & q & -3q \end{pmatrix} \end{array} \tag{2.2}$$

Rate matrices are constrained to ensure the conservation of the total probability: every column sums up to 0. The probability of change for all nucleotides over a time interval t can be then described by the following differential equation

$$\frac{\partial}{\partial t}\rho(t) = Q\rho(t) \ , \tag{2.3}$$

where ρ is a vector of all 4 nucleotides. The solution to this differential equation is then

$$\rho(t) = P(t)\rho(0) \ , \tag{2.4}$$

where $P(t)$ is the matrix of transition probabilities over the time interval t. It is given by

$$P(t) = \exp(Qt) = \sum_{k=0}^{\infty} \frac{1}{k!}(Qt)^k \ . \tag{2.5}$$

In the equation above, one can see that when k is superior to 1, this will represent cases for which more than one substitution per site are observed. This is the main advantage of using a maximum likelihood based method over maximum parsimony as the latter can only infer one change per site per branch, which will lead to false estimations of substitution rates. Each entry of the matrix $P(t)_{\beta\alpha}$ will then represent the probability of a $\alpha \rightarrow \beta$ during the time interval t. The matrix of transition probabilities in the case of a Jukes-Cantor model will be

$$P(t) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{pmatrix} \end{array} , \tag{2.6}$$

where $p = (1 - \exp(-4qt))/4$.

In this case, p is the parameter that needs to be estimated to maximize the likelihood function. The likelihood is then the product of the probabilities for each site of our alignment. Traditionally, the logarithm of the likelihood is used, in order to sum over all sites of the alignment.

2.2.2. Inferring substitution patterns from multiple alignments

Substitution patterns were computed in both human and mouse genomes as follows. Whereas similar methods were used for both genomes, they were analyzed separately. All human and mouse autosomes were divided into 1 Mbp non-overlapping windows. Endero-Pecan-Ortheus (hereafter designated as EPO) 10 eutherian mammals multiple alignments available at the Ensembl database (version 56, Hubbard et al., 2009) corresponding to each window were downloaded and restricted to the analysis of the following species: human, chimpanzee and macaque for the analysis of the human lineage, mouse, rat and human for the analysis of the mouse lineage. For both analyses, all exons were masked from our alignments using the Ensembl database annotation (version 56, mouse genome version *mm9*, human genome version *hg19*). Repeated elements were not masked from our alignments.

The method used to infer substitution patterns from multiple alignments is based on maximum likelihood (Arndt et al., 2003a; Arndt and Hwa, 2005; Duret and Arndt, 2008). It does not assume that the substitution process is time-reversible, nor that sequence composition has yet reached equilibrium. It also takes into account the

fact that the methylated cytosine of a CpG dinucleotide is hypermutable: C→T and G→A mutations occur approximately ten times more frequently in CpGs than in non CpGs (Bird, 1978; Giannelli et al., 1999). Not taking this into account can lead to the inference of incorrect substitution patterns (Duret, 2006). The method adds an additional rate parameter to represent this CpG substitution process. Finally, this method computes an individual substitution matrix for each branch of the tree. The parameters that are estimated are then substitution rates for each branch of the tree as well as the nucleotide frequencies at the root of the tree.

In this analysis of substitution patterns in 1 Mbp windows, complementary rates are assumed to be equal (A→G = T→C, denoted AT→GC) for simplicity. As a result, 2 transition rates (AT→GC, GC→AT), 4 transversion rates (AT→CG, AT→TA, GC→TA, GC→CG) and 1 CpG rate (CpG→TpG/CpA) were computed. The rate matrix computed in each branch of the tree is shown below (columns are constrained to sum up to 0).

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \bullet & r_{GC\to TA} & r_{GC\to AT} & r_{AT\to TA} \\ r_{AT\to CG} & \bullet & r_{GC\to CG} & r_{AT\to GC} \\ r_{AT\to GC} & r_{CG\to CG} & \bullet & r_{AT\to CG} \\ r_{AT\to TA} & r_{GC\to AT} & r_{GC\to TA} & \bullet \end{pmatrix} \end{array} \tag{2.7}$$

Furthermore, AT→GC and AT→CG substitution rates were grouped together as Weak (W) → Strong (S) substitution rates (G and C bases on complementary strands are bound by 3 hydrogen bonds whereas A and T bases only by 2). Similarly, GC→AT and GC→TA substitution rates were grouped together as S→W substitution rates. A total substitution rate was also computed as the weighted sum of all substitution rates. A substitution pattern consists of all substitution rates. For each substitution pattern an equilibrium GC-content or future GC-content was computed (later designated as GC*), which is the expected final GC-content if the sequence evolves with a constant substitution pattern through time. It can be viewed as both the summary value of the substitution pattern and a proportional value to the ratio between W→S and S→W substitution rates.

2.2.3. Retrieving genomic features

The following genomic features were computed in each window: GC-content, the distance to the telomere, the CpG dinucleotide odds ratio (the observed CpG frequency divided by the expected CpG frequency, later designated as CpGodds), exon density (proportion of base pairs occupied by exons in a window, later designated as Exons) as well as SINE, LINE and LTR transposable element densities (later designated as SINEs, LINEs and LTRs). Crossover rates (hereafter designated as CO) were extracted from high quality genetic maps available for the human genome (International HapMap Consortium et al., 2007) and the mouse genome (Shifman et al., 2006). These rates were computed as the weighted average of crossover rates

of chromosomal regions that overlap the window. It was possible to extract sex-averaged crossover rates in the human genome and sex-averaged as well as male and female-specific crossover rates in the mouse genome. Because in the mouse lineage the crossover rates and the distance to the telomere exhibit a non-normal distribution (Figure A.1), the logarithms of crossover rates (hereafter designated as LCO) as well as the distance to the telomere (hereafter designated as LDT) were computed. Replication-timing values (hereafter designated as RepTime) were computed from high resolution replication-timing profiles available for mouse embryonic stem cells (Hiratani et al., 2008) and human embryonic stem cells (Ryba et al., 2010), as the weighted median of replication-timing values of chromosomal regions that overlap the window. All genomic positions in the genetic maps and replication-timing profiles were converted to the versions of the human genome (*hg19*) and mouse genome (*mm9*) from which the alignments were computed using the liftOver tool available at UCSC (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

Finally, windows were filtered as follows: Windows with less than 100 kbp of sites where all three species have an aligned nucleotide were discarded, as well as those which overlapped centromeric regions and those without enough information to compute crossover rates or other genomic features (for example, windows with no genomic markers for crossover rates).

2.2.4. Multivariate analysis

To investigate the link between substitution patterns and genomic features, multivariate analysis was performed using both the Relative Contribution to Variability Explained (hereafter designated as *RCVE*) method and Principal Component Regression (hereafter designated as PCR). In both cases, the goal was to see how much all features together predict substitution patterns and how much each individual feature predicts substitution rates compared to the others.

Relative Contribution to Variability Explained

The $RCVE$ method is based on linear modeling. For each substitution rate, a linear model was first built, where the substitution rate is explained by 9 genomic features (GC-content, crossover rates, distance to telomeres, replication-timing, exons' density, transposable elements (SINEs, LINEs and LTRs) densities and CpG odds ratio).

For all methods described below, each variable was normalized such that its mean is equal to 0 and its standard deviation is equal to 1. For each of the following variables (GC*, W→S, S→W, W→W, S→S, CpG rate and Total substitution rate), a linear model was built where we investigated the relationship between the variable Y (the substitution rate) and the 9 genomic features

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_9 X_9 \ , \tag{2.8}$$

where Y is the dependent variable, X_k is the k^{th} feature and β_k the slope of the regression line between Y and the k^{th} feature. The coefficient of determination (R^2) of this full model (where all 9 features are included, designated as R^2_{full}) was extracted. The relative contribution to variability explained ($RCVE$) was then computed for each feature as follows. The goal is to find out how much a feature contributes to the model by measuring how much taking out (or shuffling) this feature will affect the model. The feature was first shuffled. The linear model was then rebuilt by replacing the original feature with the shuffled feature. The R^2 of this modified model (designated as $R^2_{\text{reduced}(k)}$) was extracted. The $RCVE$ value of the k^{th} feature was finally computed using the following formula

$$RCVE_k = \frac{R^2_{\text{full}} - R^2_{\text{reduced}(k)}}{R^2_{\text{full}}} \ . \tag{2.9}$$

Also, in each full linear model the slope between the variable Y and the k^{th} feature was extracted.

Principal Component Regression

The link between substitution patterns and genomic features was further investigated using principal component regression (principal component analysis followed by linear regression) as described below. First, principal component analysis was carried out in both human and mouse genomes on the 9 genomic features. The goal of principal component analysis is to transform feature variables into new uncorrelated variables containing exactly the same amount of variance. It first builds a covariance matrix of all 9 features as follows:

$$M = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_9 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_9 \end{array} \\ \begin{pmatrix} var(X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_9) \\ cov(X_2,X_1) & var(X_2) & \cdots & cov(X_2,X_9) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_9,X_1) & cov(X_9,X_2) & \cdots & var(X_9) \end{pmatrix} \end{array} , \tag{2.10}$$

where each entry M_{ij} represents the covariance of the two features X_i and X_j. Such matrix is a representation of all variance contained in these 9 features. The eigenvectors and associated eigenvalues of this matrix are then computed. Eigenvalues represent how much of the total variance the associated eigenvector explains. Eigenvectors are all orthogonal with respect to each other and are ranked based on their eigenvalues, the first explaining most of the variance. Entries of eigenvectors are normalized such as the sum of squared values of each vector is equal to 1. The 9 genomic features are then transformed into new variables by projecting them onto the eigenvectors. The end result will be 9 variables called principal components (PC1 to PC9), which are linear combinations of the 9 genomic features and are all

independent from each other.

Two independent projections were performed for the mouse and human lineages. As components are independent from each other, linear regressions were then performed, using the principal components previously computed as features and substitution rates computed in each lineage as variables.

$$Y = \beta_0 + \beta_k \mathrm{PC}k \tag{2.11}$$

We finally extracted the R^2 for each linear regression.

All statistics were performed using R (http://www.r-project.org/). The R package *pls* was used to perform principal component regression (Mevik and Wehrens, 2007). The R code of Drummond et al. (2006) was used to generate figures and tables for principal component regression.

2.3. GC-content evolution and meiotic recombination hotspots

We present here methods used in Chapter 5.

2.3.1. Mouse Lineage

Double strand break hotspots

Recently, genomic locations of double strand breaks (hereafter designated as DSB) hotspots were determined in the mouse genome using chromatin immuno-precipitation follow by next-generation sequencing (ChIP-seq). This methods identifies regions of the genome where proteins of interest are bound. DSB hotspots coordinates that were recently mapped in the *Mus m. musculus* genome were downloaded (Smagulova et al., 2011).

Substitution patterns across the *Mus m. musculus* genome

Substitution patterns across the *Mus m. musculus* genome were analyzed as follows. The genome was divided into 1 Mbp non-overlapping windows. For each window, *Mus m. musculus* - *Mus musculus castaneus* (or *Mus m. castaneus* for short)- *Mus spretus* triple alignments were built as follows. First, the genomic consensus sequences of *Mus spretus* and *Mus m. castaneus* that were recently sequenced and mapped to the reference *Mus m. musculus* genome were downloaded (version *mm9*, Keane et al., 2011). Genomic sequences of these two species have the same coordinates than *Mus m. musculus*, which allows for direct sequence comparison. For each window, the corresponding sequence of the reference *Mus m. musculus* genome was then obtained from the Ensembl database (version 62, Flicek et al., 2011, the

corresponding sequences of *Mus m. castaneus* and *Mus spretus* were extracted. Finally, all exons were masked from the *Mus m. musculus* reference sequence (Ensembl annotation).

Substitution rates were computed from these alignments using the same maximum likelihood-based method as in previous sections (Arndt et al., 2003a; Arndt and Hwa, 2005; Duret and Arndt, 2008). Similarly, 7 substitution rates: 2 transition rates and 4 transversion rates as well as 1 CpG rate were computed. A or T $\rightarrow$ G or C substitution rates were grouped together as Weak (W) $\rightarrow$ Strong (S) substitution rates, and G or C $\rightarrow$ A or T substitution rates as S$\rightarrow$W substitution rates. Similarly to previous analyses, a GC* value was finally computed for each window.

The divergence time of *Mus m. musculus* and *Mus spretus* is estimated to be around 2 million years (Veyrunes et al., 2005), that of *Mus m. musculus* and *Mus m. castaneus* around 500,000 years (Geraldes et al., 2008). Moreover, the nucleotide divergence (number of positions where bases are different in both species divided by number of positions where both species have a nucleotide) of *Mus m. musculus* and *Mus m. castaneus* is about 0.010, that of *Mus m. musculus* and *Mus spretus* about 0.022 and that of *Mus m. castaneus* and *Mus spretus* about 0.021 (for comparison, the nucleotide divergence of human and chimpanzee is around 0.013). Substitution rates were therefore computed by comparing both the *Mus m. musculus* and *Mus m. castaneus* lineages and using *Mus spretus* as an outgroup.

In each window, crossover rates as well as DSB hotspot density were computed as follows. Sex-averaged as well as male and female-specific crossover rates were extracted from high-quality genetic maps available for the *Mus m. musculus* genome (Shifman et al., 2006; Cox et al., 2009). Crossover rates were computed in each window as the weighted average of crossover rates of chromosomal regions that overlap the window. At the same time, DSB hotspot density values were calculated as the proportion of the windows overlapping a DSB hotspot. These density values were used as a proxy measure of DSB activity. Windows containing less than 100 kbp of sites where the three species have a nucleotide as well as windows not containing enough information to compute crossover rates or DSB density values (for example, windows not overlapping any genetic marker in genetic maps nor overlapping any DSB hotspot) were filtered out.

All analyses were done using R (http://www.r-project.org/). Because almost all DSB hotspots have identical length, DSB hotspot density will not behave like a fully continuous variable but like a discrete variable. This makes the computation of *p-values* unreliable for correlation coefficients. To solve this issue, s linear model was computed where DSB hotspot density is the explanatory variable, the *p-value* of the slope between the explanatory variable and the response variable was then extracted.

Substitution patterns around double strand break hotspots

Substitution patterns around DSB hotspots were analyzed as follows. First, all DSB hotspots were merged together using the hotspot center position as a reference position. Then, sequences around middle points were divided into 2,000 non-overlapping windows of 100 bp, 1,000 on the 5' side and 1,000 on the 3' side of middle points. For each window, *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* triple alignments were built using the same methodology as described above. Similarly, substitution patterns were computed using the same maximum likelihood-based method as described above. This time we wanted to compare complement rates, 14 rates were computed: 4 transition rates, 8 transversion rates as well as 2 CpG rates. The rate matrix computed in each branch of the tree is shown below (columns are constrained to sum up to 0).

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \bullet & r_{C\to A} & r_{G\to A} & r_{T\to A} \\ r_{A\to C} & \bullet & r_{G\to C} & r_{T\to C} \\ r_{A\to G} & r_{C\to G} & \bullet & r_{T\to G} \\ r_{A\to T} & r_{C\to T} & r_{G\to T} & \bullet \end{pmatrix} \end{array} \tag{2.12}$$

To visualize trends in substitution patterns around DSB hotspot middle points, local polynomial regression were performed as follows. First, all windows on the 5' side or on the 3' side of reference positions were grouped into two separate groups. A local polynomial regression was then fitted in each group of windows, using the windows' positions as predictor values and GC*, W→S or S→W substitution rates as response values, giving us fitted values. The smoothing was done over 25 neighbor windows, which corresponds to a span parameter of 0.025 (25/1000 = 0.025) for the local polynomial regression.

2.3.2. Human Lineage

Recombination hotspots

Genomic coordinates of meiotic recombination hotspots were defined in the human genome as follows. The HapMap genetic map available for the human genome was downloaded (*Homo sapiens*, version *hg19*, International HapMap Consortium et al., 2007). Inside this map, any genomic interval between two markers having a recombination rate of at least 10 cM/Mb was considered as a recombination hotspot. Adjacent intervals fulfilling this criterion were merged together. More than 25,000 hotspots were obtained this way.

PRDM9 binding sites

Each hotspot was scanned for the consensus PRDM9 binding motif (CCNCCNTNNCCNC, Myers et al., 2008, 2010) on both + and − strands. Hotspots containing no

motif or more than one motif were filtered out, keeping only hotspots containing one binding motif. More than 6,000 recombination hotspots were kept this way. All these hotspots were finally pooled together, using PRDM9 binding sites as a reference position, and using the same orientation as the binding site.

Alignments and substitution patterns

Substitution patterns around PRDM9 binding sites were analyzed as follows. First, sequences around binding sites were divided into 2,000 non-overlapping windows of 100 bp, 1,000 both on the 5' end and on the 3' end of binding sites. In each window, all human - chimpanzee - gorilla alignments from EPO primates multiple alignments were then downloaded from the Ensembl database (version 62, Flicek et al., 2011). All exons were masked in the human alignments (Ensembl annotation). Substitution patterns were finally computed in both the human and chimpanzee lineages using the same method as in the mouse lineage by comparing human and chimpanzee and using gorilla as an outgroup.

Trends in substitution patterns around PRDM9 binding sites were visualized by performing the same analysis as around DSB hotspots middle points in *Ms m. musculus* (see section 2.3.1 for more details).

3. Isochore structures and GC-content variation

3.1. Introduction

Until recently and the complete sequencing and publication of mammalian genomes, direct statistical study of base composition in genomes was not possible. Instead, this was done indirectly through biochemical methods like ultracentrifugation or by using the GC-content of synonymous positions in genes as a proxy measure. As more and more species have their genome sequenced, it is possible to measure genome-wide features of base composition like how much it varies within a genome. Furthermore, we can also compare several organisms with each other, thus revealing how base composition evolves in different lineages.

3.2. GC-content along one human chromosome

The simplest way to study base composition in one genome is to look at how it varies along one chromosome. Figure 3.1 shows GC-content values in a 3 Mbp segment of the human chromosome 1 for different tilings (10 kbp and 100 kbp), as well as values for the same segment after shuffling nucleotides in the segment.

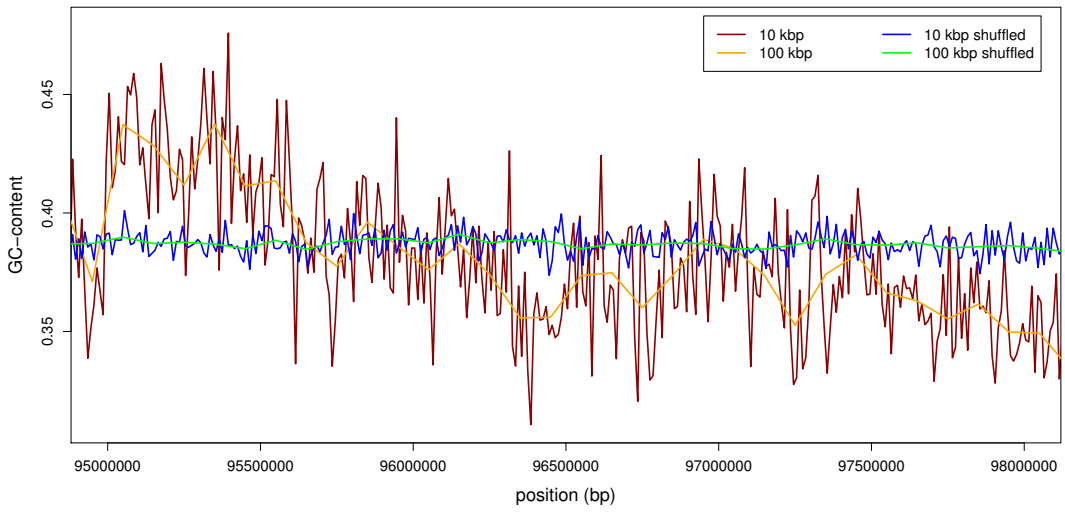


Figure 3.1.: GC-content profile along a 3 Mbp segment of the human chromosome 1.

We first observe that GC-content variation changes with the tiling: it varies less for larger tilings. Second, we notice more variation in the genomic GC-content than in that of the shuffled sequence.

We conclude several things from these observations. First, as GC-content of genomic sequences varies more than that of random sequences, it seems that there is a position-dependent process acting on GC-content evolution. Second, as the GC-content variation of shuffled sequences changes with tiling, it seems there is a link between GC-content variance and tiling. We ask ourselves what is the random expectation for GC-content variance and its link with tiling.

3.3. GC-content variance in random sequences

3.3.1. Analytical solution

Generating a random DNA sequence of specific length and base composition can be modeled as a binomial process. In this case, we represent success by adding a G or C base to the sequence and failure by adding an A or T base. Likewise, N will be the sequence's length and f_{GC} the probability of adding a G or a C to the sequence (this parameter will then represent the frequency of G & C in the sequence or GC-content). According to the properties of a binomial process, the mean number of G and C bases (N_{GC}) in the sequence will then be

$$E(N_{\mathrm{GC}}) = N f_{\mathrm{GC}} \ , \tag{3.1}$$

the variance of the number of G and C bases will be

$$\begin{aligned} &Var(N_{\mathrm{GC}}) = E(N_{\mathrm{GC}} - E(N_{\mathrm{GC}}))^2 \\ &= E(N_{\mathrm{GC}}^2) - E(N_{\mathrm{GC}})^2 \ . \end{aligned} \tag{3.2}$$

According to the properties of the binomial process, this variance can be rewritten

$$Var(N_{\mathrm{GC}}) = N f_{\mathrm{GC}}(1 - f_{\mathrm{GC}}) \ . \tag{3.3}$$

The observed frequency of G & C bases $\hat{f}_{\mathrm{GC}}$ can be rewritten as follows: $\hat{f}_{\mathrm{GC}} = \frac{N_{\mathrm{GC}}}{N}$. The mean of $\hat{f}_{\mathrm{GC}}$ is now

$$E(\hat{f}_{\mathrm{GC}}) = \frac{E(N_{\mathrm{GC}})}{N} \ . \tag{3.4}$$

The variance of $\hat{f}_{\mathrm{GC}}$ can be decomposed as follows:

$$\begin{aligned}
& Var(\hat{f}_{\mathrm{GC}}) = E\left[\hat{f}_{\mathrm{GC}} - E(\hat{f}_{\mathrm{GC}})\right]^2 \\
&= E\left[\frac{N_{\mathrm{GC}}}{N} - \frac{E(N_{\mathrm{GC}})}{N}\right]^2 \\
&= E\left[\frac{N_{\mathrm{GC}}^2}{N^2} - \frac{2N_{\mathrm{GC}}E(N_{\mathrm{GC}})}{N^2} + \frac{E(N_{\mathrm{GC}})^2}{N^2}\right] \\
&= E\left[\frac{1}{N^2}\left(N_{\mathrm{GC}}^2 - 2N_{\mathrm{GC}}E(N_{\mathrm{GC}}) + E(N_{\mathrm{GC}})^2\right)\right] \\
&= \frac{1}{N^2}E[N_{\mathrm{GC}}^2 - 2N_{\mathrm{GC}}E(N_{\mathrm{GC}}) + E(N_{\mathrm{GC}})^2] \\
&= \frac{1}{N^2}E[N_{\mathrm{GC}} - E(N_{\mathrm{GC}})]^2 \\
&= \frac{1}{N^2}Var(N_{\mathrm{GC}}) \\
&= \frac{1}{N^2}Nf_{\mathrm{GC}}(1 - f_{\mathrm{GC}}) \\
&= f_{\mathrm{GC}}(1 - f_{\mathrm{GC}})N^{-1} \ .
\end{aligned} \tag{3.5}$$

We therefore conclude that for random DNA sequences GC-content variance depends both on the inverse of the sequence's length (N^{-1}) and on its GC-content (f_{GC}).

The fact that N is to the power -1 makes this a power law. Such law can be represented as follows: $y = \exp(\alpha) \times x^{\beta}$ or $\log(y) = \alpha + \beta \log(x)$, the base composition's variance can thus be viewed as a power law: $Var(\hat{f}_{\mathrm{GC}}) = f_{\mathrm{GC}}(1 - f_{\mathrm{GC}}) \times N^{-1}$ or $\log(Var(\hat{f}_{\mathrm{GC}})) = \log(f_{\mathrm{GC}}(1 - f_{\mathrm{GC}})) - \log(N)$, where y is the GC-content variance, α is $\log(f_{\mathrm{GC}}(1 - f_{\mathrm{GC}}))$, x is the sequence's length and β is -1.

These equations allow us to compare observations for genomic sequences to random sequences exhibiting identical base compositions.

3.3.2. Comparison with human genomic sequence

Historically, the human genome was the first mammalian genome to be fully sequenced (Lander et al., 2001), we therefore analyzed this genome first. We downloaded the latest available genomic sequence for human (*Homo sapiens*, version *hg19*) and analyzed all chromosomes. Figure 3.2 shows the GC-content variance and distribution for this genome as well as that of shuffled genomic sequence.

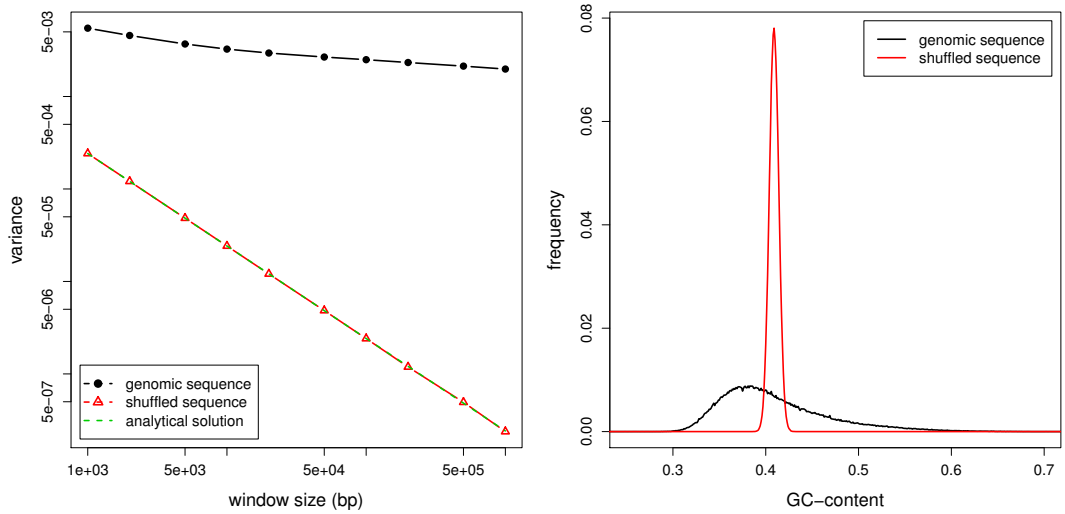


Figure 3.2.: Left panel: GC-content variance plotted against tiling for human genomic or shuffled sequences, as well as the analytical solution for a GC-content of 0.41. Right panel: GC-content distribution for human genomic shuffled sequences for a tiling of 10 kbp.

We first see that human GC-content is more variable than what is expected by chance. These variations are called isochore structures (Bernardi, 2000; Eyre-Walker and Hurst, 2001). Second, GC-content variance decreases as tiling increases. However, this decrease is weaker than what is expected by chance: we see more variance in large tilings than in small ones compared to random sequences.

We can now analyze and compare several genomes with each other.

3.4. GC-content variance & distribution in genomic sequences

We first want to have a global view of base composition across different major lineages in animals. To do so, we analyzed the following groups: mammals with human (*Homo sapiens*, version *hg19*), birds with chicken (*Gallus gallus*, version *galGal3*), reptiles with anole lizard (or lizard, *Anolis carolinensis*, version *anoCar2*), fish with zebrafish (*Danio rerio*, version *Zv9*) and insects with drosophila (*Drosophila melanogaster*, version *dm3*). Figure 3.3 shows phylogenetic relationships as well as divergence times for these species.

One can clearly see that different groups show very different base compositions, for both variance and distribution, for example, the human genome shows the highest GC-content variance values whereas the lizard genome shows the lowest (Figure 3.4). In subsequent sections, we first compare base composition within different clades (primates, mammals, birds) to precisely characterize these groups and then compare them. Doing so, we hope to highlight major evolutionary forces acting in different lineages and how they influence GC-content evolution.

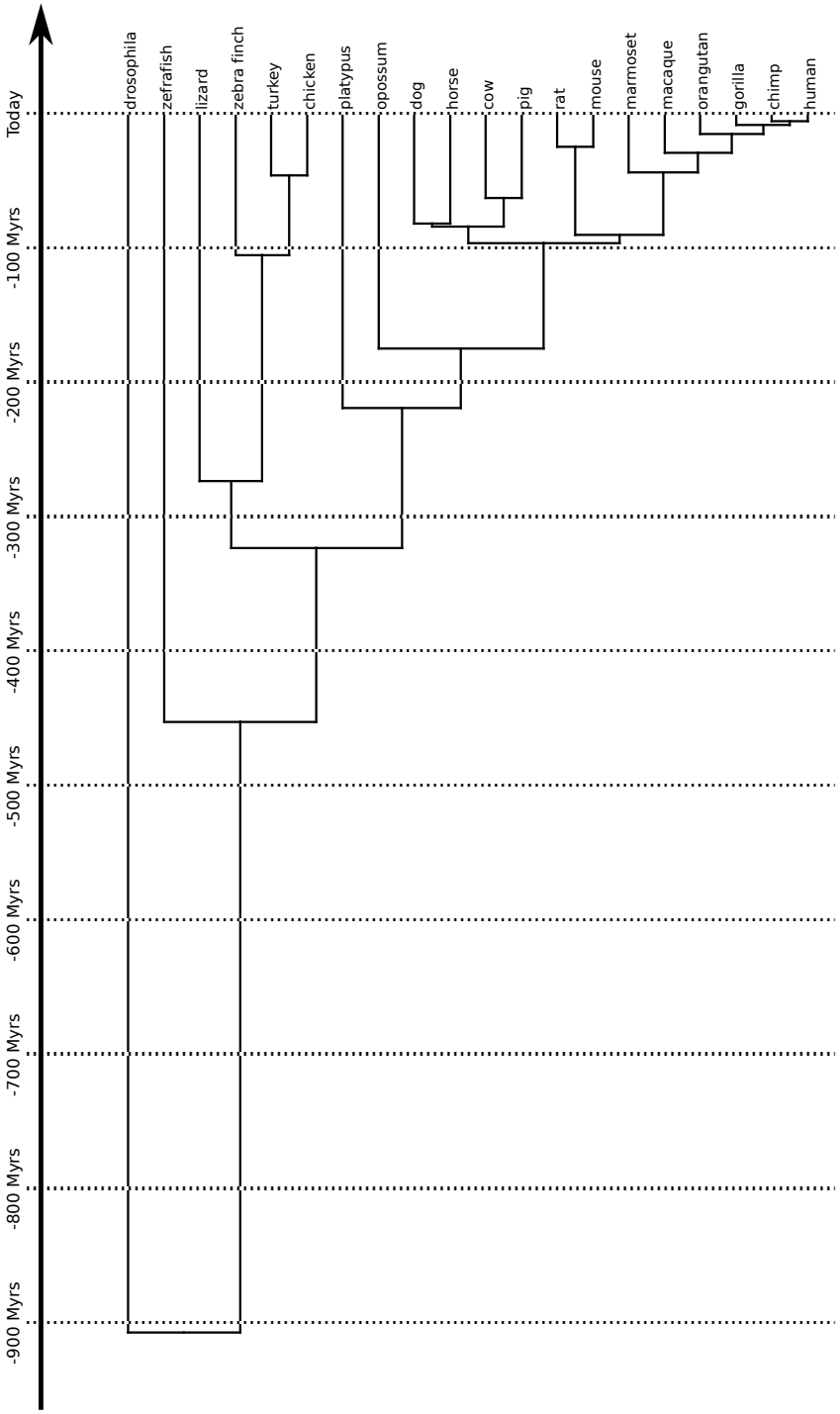


Figure 3.3.: Phylogenetic tree of all species studied in this chapter. The Dendroscope program was used to draw the tree (Huson et al., 2007). Divergence time were downloaded from TimeTree.org (Hedges et al., 2006; Kumar and Hedges, 2011).

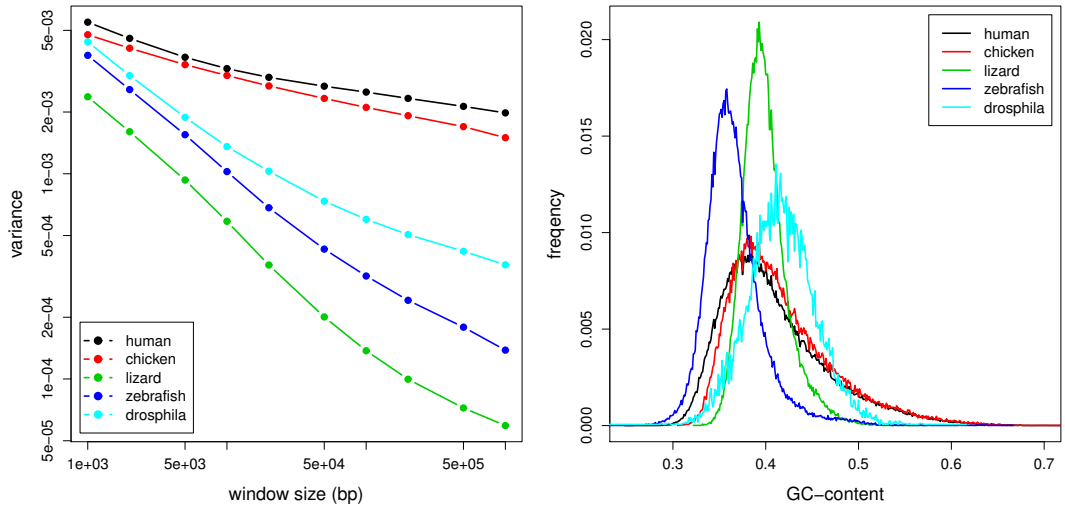


Figure 3.4.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

3.4.1. Primates

We describe GC-content variance and distribution in six primate genomes: human (*Homo sapiens*, version hg19), chimpanzee (*Pan troglodytes*, version *panTro3*), gorilla (*Gorilla gorilla*, version *gorGor3*), orangutan (*Pongo abelii*, version *ponAbe2*), macaque (*Macaca mulatta*, version *rheMac2*) and marmoset (*Callithrix jacchus*, version *calJac3*). Figure 3.5 shows GC-content variance plotted against tiling (left panel) as well as GC-content distributions (right panel) for these genomes.

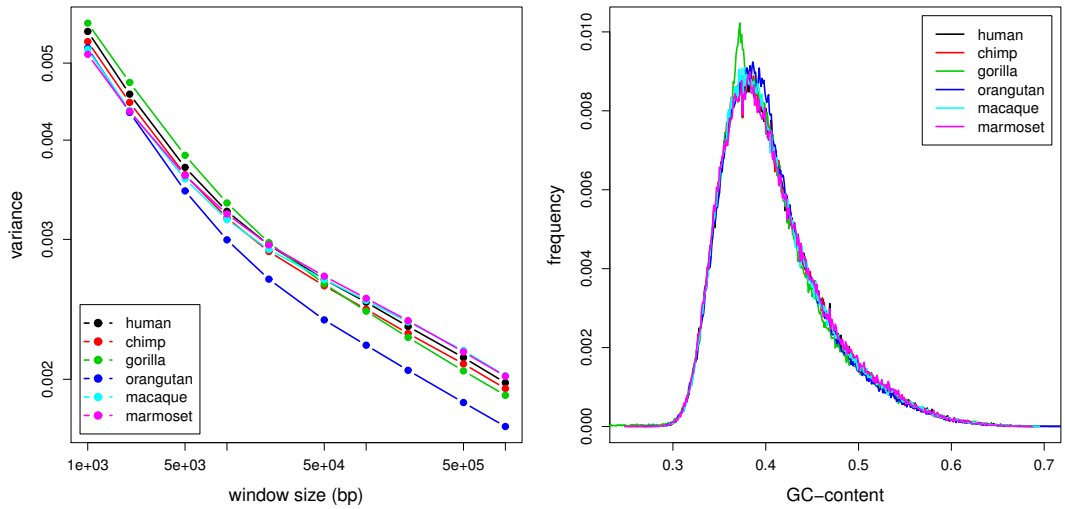


Figure 3.5.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

GC-content variance and distribution results show that primates' genomes are extremely similar with respect to base composition. We note that orangutan GC-content variance is slightly lower than other primate species. We also note that gorilla has an excess of regions with a GC-content of 0.35 relative to other primates. This similarity is expected as primate species have short divergence times, there is not enough time for base composition to change radically between species (Figure 3.3, Fabre et al., 2009). However, even marmoset, the most distantly related species to human (40 Myrs, Figure 3.3, Fabre et al., 2009), GC-content variance and distribution are indistinguishable from other primates, meaning base composition stayed stable during primate evolution.

We next looked at GC-content distributions. Again, GC-content profiles are very similar with two exceptions, gorilla showing an excess of sequence at 0.36 and orangutan showing a slight excess of sequences from 0.40 to 0.43.

3.4.2. Rodents and mammals

We describe here results for the mouse (*Mus m. musculus*, version *mm9*) and rat (*Rattus norvegicus*, version *rn4*) genomes. Figure 3.6 shows the results for these genomes.

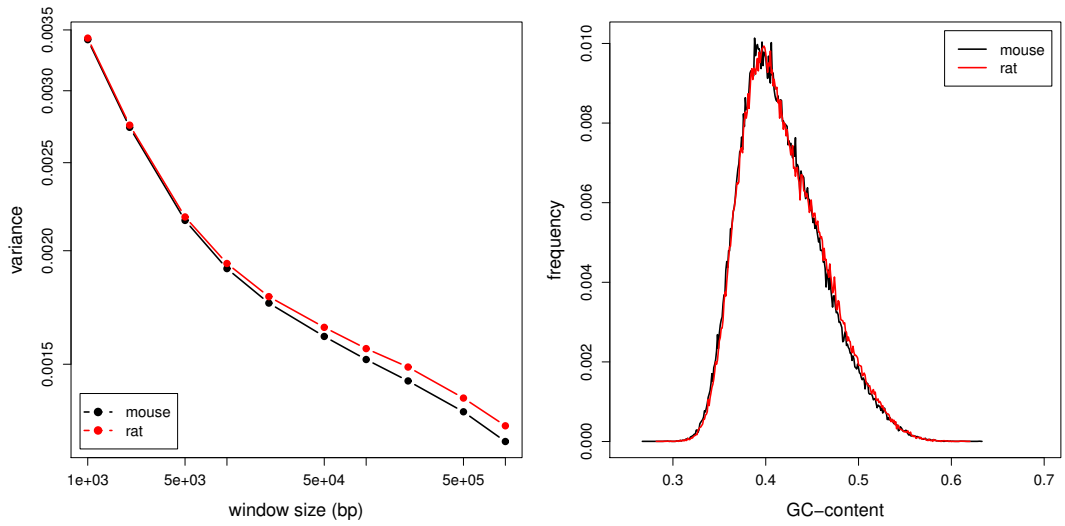


Figure 3.6.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

GC-content distribution and variance are very similar between mouse and rat genomes, which is expected as these two species diverged about 20 Myrs ago (Figure 3.3, Poux et al., 2006; Huchon et al., 2007). Similar to primates, base composition did not experience much changes in rodents.

We compared the human and mouse genomes to the following: horse (*Equus caballus*, version *equCab2*), dog (*Canis lupus familiaris*, version *canFam2*), pig (*Sus*

scrofa, version *susScr2*), cow (*Bos taurus*, version *bosTau6*), opossum (*Monodelphis domestica*, version *monDom5*) and platypus (*Ornithorhynchus anatinus*, version *ornAna1*). Figure 3.7 shows GC-content variance (left panel) and distribution (right panel) for these genomes

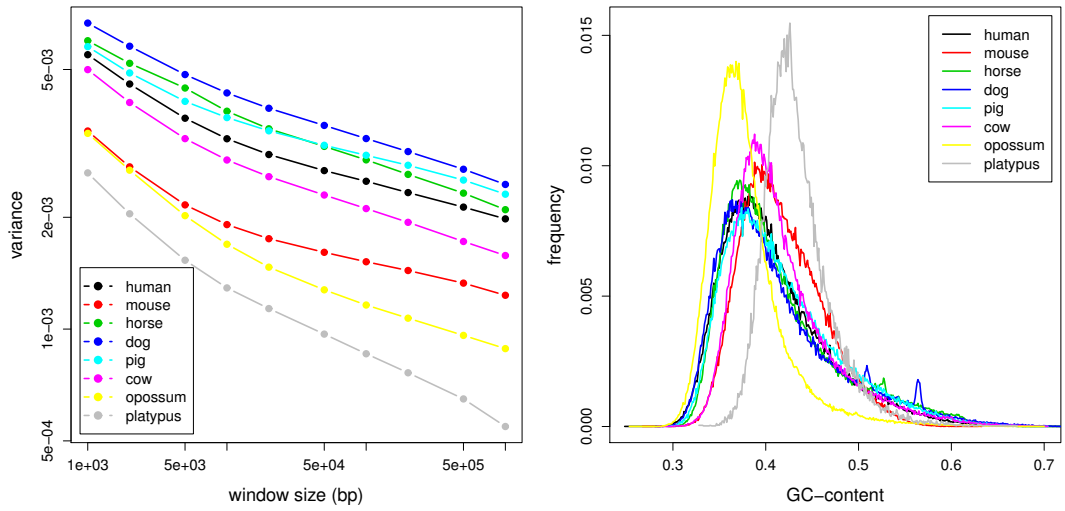


Figure 3.7.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

We first observe that with the exception of the cow genome, all eutherians have a higher GC-content variance than human, regardless of tiling. The cow GC-content variance is slightly lower than that of human, the dog genome has the highest variance. Second, both platypus and opossum have lower GC-content variance than mouse, platypus having the lowest values. It is possible that low variance in platypus is due to the fact that GC-content variance is computed over fewer windows in platypus compared to opossum (439 and 3578 1 Mbp windows respectively). Interestingly, in both these genomes GC-content variance decreases more as tiling increases compared to other mammalian genomes.

GC-content distributions show that pig, dog, horse and human have similar profiles. Interestingly, the dog genome shows relative to other genomes an excess of regions with a GC-content of 0.50 and 0.56 whereas the pig genome shows a slight excess of regions with a GC-content of 0.52. Overall, the cow genome shows a profile similar to mouse. Both platypus and opossum have a more homogeneous GC-content distribution compared to eutherians, opossum having a much lower mean GC-content than platypus (0.39 and 0.43 respectively).

To study the evolution of base composition in mammals, one might ask first whether its ancestral state can be determined. Several studies did find that the mammalian ancestral isochore structure is very close to the human structure (Galtier and Mouchiroud, 1998; Romiguier et al., 2010). We therefore hypothesize here that the base composition variance and distribution observed in the human genome represent

that of the mammalian common ancestor, and base compositions observed in other lineages represent derived states. As a result, we ask what caused base composition to evolve in different mammalian lineages.

There are three major evolutionary processes that can affect base composition: base substitutions, insertion and deletion events (hereafter designated as indels) and transposable elements (hereafter designated as TE) dynamics. It has been shown that base substitutions are affected by a neutral process linked with meiotic recombination called GC-biased gene conversion (hereafter designated as gBGC) which favors the fixation of mutations from A or T (weak or W) bases to G or C (strong or S) bases and disfavors the fixation of mutations from S to W (Galtier et al., 2001; Duret and Galtier, 2009): regions experiencing high levels of recombination will have increased rates of W→S substitution and decreased rates of S→W substitution. Furthermore, a link between chromosomal organization and meiotic recombination has been found: as a minimum of one crossover per chromosomal arm per meiosis is necessary to ensure correct migration of chromosomes (Petronczki et al., 2003), short chromosomal arms will experience more recombination than long chromosomal arms (Kaback, 1996; de Villena and Sapienza, 2001; Coop and Przeworski, 2007). Changes in chromosome size and number will therefore affect base composition evolution.

The mouse genome has 19 autosomes that are all telocentric (i.e. the centromere is located at one of the telomeres) whereas the human genome has 22 autosomes with the majority of them being metacentric (i.e. the centromere is located in the middle of the chromosome). As a result, mouse has fewer chromosomal arms and experiences less recombination than the latter (Jensen-Seaman et al., 2004; Li and Freudenberg, 2009). Furthermore, chromosome-wide recombination rates (the total recombination rate for each chromosome) vary more in human than in mouse or rat (the variance for these rates is 0.100, 0.017 and 0.033 for these genomes respectively, Jensen-Seaman et al., 2004). It is then possible to explain difference in genome-wide GC-content variance by differences in chromosome-wide recombination rates variance: base composition variance will increase with chromosome-wide recombination rates (Eyre-Walker, 1993). Agreeing with this prediction, dog has more variable recombination rates than human (0.125, Wong et al., 2010) whereas recombination rates variance for cow is between mouse and human (0.039, Arias et al., 2009). However, chromosomes are divided into one or two arms and one crossover is required for each arm of each chromosome. Chromosome-wide observations have to be taken cautiously as chromosomal arms have not been characterized for genomes other than mouse, rat and human.

The precise study of substitution patterns or TE dynamics across genomes can show us how GC-content is evolving in different lineages. Results in human show that GC-rich regions see their GC-content decreasing while GC-poor regions are close to equilibrium (Arndt et al., 2003b; Meunier and Duret, 2004; Khelifi et al., 2006; Duret and Arndt, 2008), which might cause base composition to depart from its ancestral-like structure. However, this will happen very slowly as GC-content is evolving slowly in primates (Duret and Arndt, 2008; Romiguier et al., 2010). Substitution patterns

in mouse show that GC-content is also globally decreasing (Khelifi et al., 2006; Clément and Arndt, 2011): the GC-content in both GC-poor and GC-rich regions is decreasing whereas GC-medium regions are close to equilibrium. Transposable elements dynamics is different for both human and mouse genomes. Whereas in the mouse genome both GC-rich SINE elements and GC-poor LINE elements are preferentially inserted in GC-rich regions, in human both classes of elements are inserted mostly in GC-poor regions (Yang et al., 2004).

3.4.3. Birds

We compared the human and mouse genomes to three genomes of birds: chicken (*Gallus gallus*, version *galGal3*), zebra finch (*Taeniopygia guttata*, version *taeGut1*) and turkey (*Meleagris gallopavo*, version *melGal1*). Figure 3.8 show the results for these genomes.

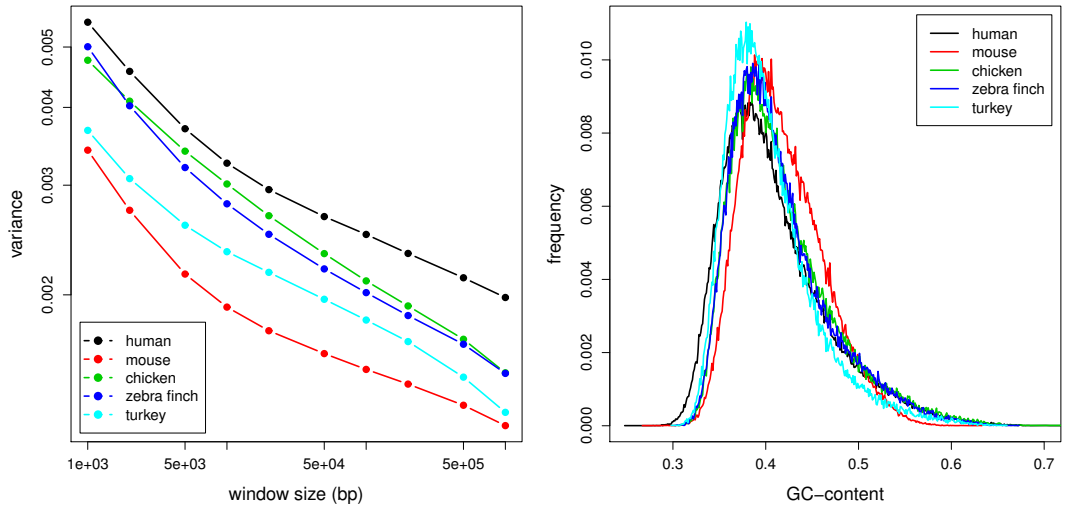


Figure 3.8.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

All bird genomes have their GC-content variance above mouse and below human, regardless of tiling. Chicken and zebra finch have similar variance, slightly lower than human whereas turkey variance is lower.

GC-content distributions show similar trends: chicken and zebra finch have very similar GC-content distributions, different from turkey. This is surprising as chicken and turkey are more closely related to each other (49 Myrs divergence time, Pereira and Baker, 2006) than with zebra finch (122 Myrs divergence time, Figure 3.3, Pereira and Baker, 2006). We can interpret this in two different ways. First, the base composition shared between chicken and zebra finch represents the ancestral bird composition and that of turkey evolve differently since its divergence with chicken. Second, the turkey base composition represents the bird ancestral composi-

tion with chicken and zebra finch evolving independently towards an identical base composition. To test these possibilities, one might have to study features influencing GC-content evolution in different bird lineages.

Birds differ from other mammals in their chromosomal organization as birds chromosomes are grouped into long macrochromosomes and small microchromosomes, the latter experiencing much higher amounts of recombination compared to the former (Groenen et al., 2009). As a result, the chromosome-wide recombination rates' variance is extremely high compared to mammalian genomes (41.62, Groenen et al., 2009). Moreover, the GC-content of GC-poor regions is decreasing whereas that of GC-rich regions is increasing (Webster et al., 2006). These facts would suggest a high GC-content variance in bird genomes, something that we do not observe. As GC-rich CR1 elements are found in both GC-rich and GC-poor regions of the chicken genome and GC-rich MIR elements in GC-medium regions of the genome (Abrusán et al., 2008), transposable elements dynamics do not provide an explanation for the fact that birds genomes have lower GC-content variance than expected.

3.4.4. Amniotes & Reptiles

Using the recently sequenced genome of the anole lizard (Alföldi et al., 2011), we finally compared the human, mouse, dog, opossum, platypus and chicken genomes to the anole lizard genome (*Anolis carolinensis*, version *anoCar2*). Figure 3.9 shows results for these genomes.

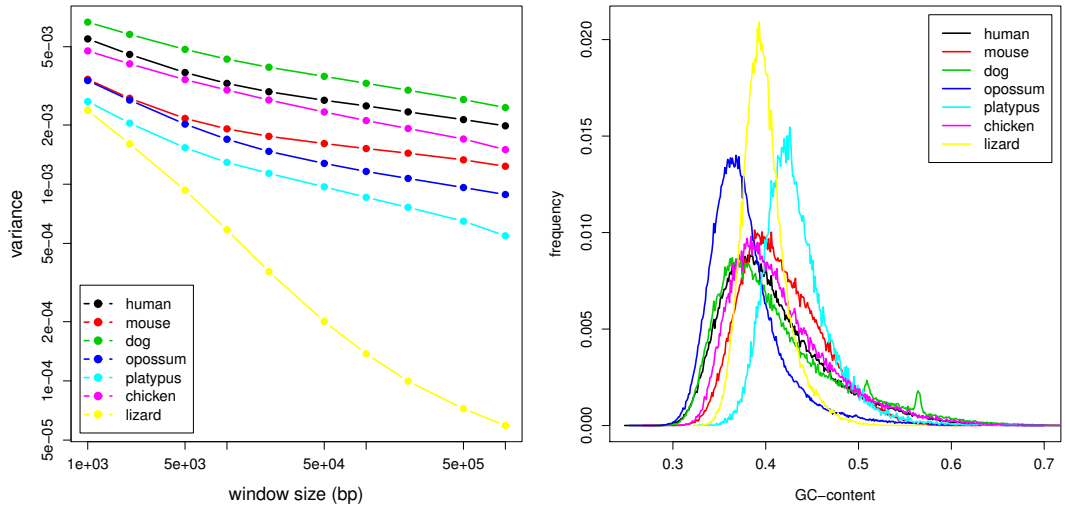


Figure 3.9.: Left panel: GC-content variance plotted against tiling used for the analysis. Right panel: GC-content distribution for a tiling of 10 kbp.

Compared to other amniotes (bids and mammals), the anole lizard has a very different base composition. First, it has less variance compared to other genomes. We note that this variance decreases much more as tiling increases than other genomes.

Second, the lizard has a much more homogeneous GC-content distribution compared to other amniotes. Its mean GC-content, however, is similar to eutherian and birds (0.40).

The anole lizard genome represents the only sequenced genome of an amniote that is neither a mammal nor a bird. Base composition in this genome shows unexpected trends. Despite a mean GC-content close to human (0.40), the GC-content variance is much lower than other genomes studied here, especially at large tilings. As its variance is lower than that of xenopus (Fujita et al., 2011) or zebrafish or drosophila (Figure 3.4), this feature seems to be specific to the lizard genome rather than representing an ancestor state of amniotes. Moreover, there is very little correlation between base composition and both intergenic and intron length whereas inverse correlations are found in mammals and birds (Fujita et al., 2011). Finally, the evolution of the GC-content at third positions of codons (GC_3) in different amniote lineages shows that the base composition of the lizard genome is decreasing but not as much as other amniotes like opossum (Fujita et al., 2011). What causes this decline of base composition variance in lizard is still unknown, however a decline in strength of gBGC in the lizard lineage is one of the likeliest causes.

The base composition in the lizard genome appears to be an exception among reptiles as shown by GC_3 analysis in turtles and crocodiles (Hughes et al., 1999, 2002; Janes et al., 2010). The low number of reptile full genome sequences prevents us from studying base composition evolution in detail, although the sequencing of several crocodile genomes will make comparative studies possible in this group (John et al., 2012).

3.5. Conclusion

In this chapter we described base composition variations along mammalian chromosomes. We first found that for random sequences GC-content, its variance and window size are mathematically linked, which enabled us to compute expectations for GC-content variance for sequences of a particular GC-content We then compared these random expectations to genomic sequences of various animal taxa to find that generally mammalian genomes have more GC-content variance than random sequences. Moreover, we showed that different taxa exhibit isochore structures, with mammalian and bird genomes having the most variable structures whereas amniotes like anole lizard have the least variable base composition. Finally, we linked differences in base composition between species to changes in genomic features such as chromosomal organization and meiotic recombination.

4. Substitution patterns in primates and rodents

4.1. Introduction

The analysis of GC-content showed major differences across mammalian genomes, notably between human and mouse, indicating the GC-content is evolving differently in both species. Of all three major evolutionary forces influencing GC-content evolution, base substitution are the most widely studied and can help us understand how GC-content evolved. We aim here to first measure substitution patterns across human and mouse lineages and then identity the different forces influencing substitution patterns. We also aim to discover if the same forces are at work in both lineages and if so, how important are these different relative to each other.

4.2. GC-content evolution and GC-biased gene conversion in human and mouse lineages

It is now known that GC-biased gene conversion (hereafter designated as gBGC) is a major factor affecting substitution patterns in the human lineage (Galtier et al., 2009; Duret and Arndt, 2008; Pozzoli et al., 2008; Tyekucheva et al., 2008; Arndt et al., 2005; Webster et al., 2005; Meunier and Duret, 2004). As most studies on gBGC focused on primates, one can ask how much this process influences substitution patterns in the mouse genome. As murid rodents and primates differ greatly in their chromosomal organization as well as their recombination profiles (see Chapter 3 for more details), we expect gBGC to have a different impact in both lineages leading to a different GC-content evolution.

4.2.1. GC-content is decreasing in the mouse genome

We computed substitution patterns and GC* (equilibrium GC-content) in both human and mouse lineages in 1 Mbp windows using triple alignments (see the Materials and Methods chapter for more details). After filtering out windows without at least 100 kbp of sites where all three species of the triple alignments share a nucleotide and those overlapping centromeric regions, we obtained 1527 windows containing more than 479 Mbp of analyzable sites in the mouse genome and 2570 windows

containing more than 1800 Mbp of analyzable sites in the human genome. Results show that human and mouse GC-content is evolving very differently (Figure 4.1).

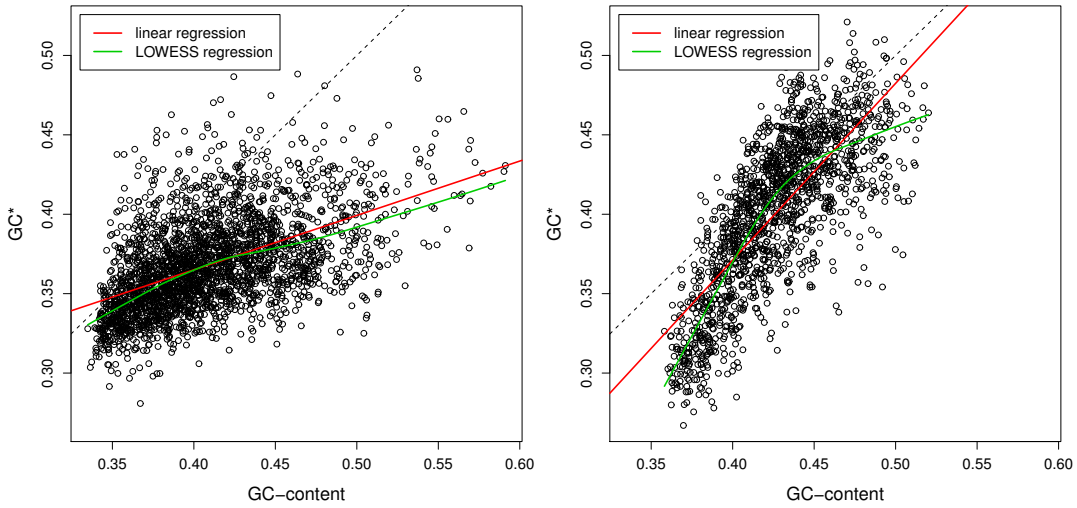


Figure 4.1.: GC* plotted against GC-content for the human (left panel) and the mouse (right panel) lineages. Dashed lines represent the $x = y$ relationship.

We found a linear relationship between GC-content and GC* in the human lineage. In GC-rich regions, GC-content is decreasing (GC* is lower than GC-content) whereas in GC-poor regions, GC-content is at equilibrium (GC* is equal to GC-content). In the mouse lineage, the relationship between GC-content and GC* is not linear, illustrated by the fact that the local LOWESS regression and the linear regression between the two variables do not match (Figure 4.1). We see that the GC-content is decreasing in GC-rich regions but also in GC-poor regions. The GC-content in GC-intermediate regions (GC-content equal to 0.42) is at equilibrium.

Our results show that both human and mouse lineages exhibit different modes of GC-content evolution. We also show that the GC-content of GC-rich regions is decreasing in both lineages, confirming previous results which called this phenomenon the erosion of GC-rich isochores (Duret et al., 2002; Belle et al., 2004; Smith and Eyre-Walker, 2002). Moreover, it has been suggested that this murid shift was caused by meiotic recombination rates being less variable in the mouse genome (Eyre-Walker, 1993). We do indeed observe that mouse crossover rates (a proxy measure of meiotic recombination rates) are less variable than human crossover rates (variance = 0.50 and 0.69 for mouse and human crossover rates respectively). These previous studies have shown, however, that the GC-content of GC-poor regions is increasing in murid rodents, whereas we show that the GC-content is decreasing in these regions. We can explain these differences by the small number of genes these studies relied on, analyzing the GC-content at synonymous positions (GC_3).

It has been hypothesized that the decline of GC-rich isochores in primates and murid rodents has been caused by chromosomal fusions at the time of mammalian

radiation, more than 80 Myrs ago (Duret et al., 2002). However, since this decline is not shared across all mammals (Romiguier et al., 2010), it is likely that different factors influenced GC-content evolution in both human and mouse lineages. We therefore have to specifically compare primate and murid rodent GC-content evolution and substitution patterns.

4.2.2. gBGC is weaker in the mouse lineage compared to the human lineage

We applied the same methodology as previous studies and analyzed the link between GC-content, GC* and crossover rates (hereafter designated as CO) (Meunier and Duret, 2004; Duret and Arndt, 2008). As crossover rates are not normally distributed in both human and mouse genomes (Figure A.1), we used the logarithm of crossover rates (hereafter designated as LCO) for the remainder of the chapter.

We observe a positive correlation between GC-content and crossover rates in both lineages (Tables 4.1 & 4.2). The cause and effect relationship between these two features cannot be inferred from this correlation alone: GC-content could influence recombination (as it has been shown in yeast, Gerton et al., 2000), recombination could influence GC-content or there could be no cause and effect relationship between the two. To solve this issue, we calculated correlation coefficients between GC* and crossover rates. We see that these correlations are stronger than between GC-content and crossover rates (Tables 4.1 & 4.2).

We draw two conclusions from these results. First, since GC* values are computed from substitution patterns and not from current GC-content, these results show that in the mouse lineage, as well as the human lineage, meiotic recombination has an effect on GC-content evolution by acting on substitution patterns. This is consistent with the influence of gBGC on substitution patterns. We repeated this analysis in the mouse lineage for male and female-specific crossover rates (Table 4.2), as well as using Spearman's correlation coefficients, and obtained similar results (Tables A.1 & A.2). We also obtained similar results when using the logarithm of the distance to telomeres (hereafter designated as LDT) as it is known to be a proxy measure of meiotic recombination rates (Tables 4.1, 4.2, A.1 & A.2, Duret and Arndt, 2008). The correlation between LDT and recombination is negative, accordingly we observe negative correlations between LDT, GC-content and GC*. Second, our results suggest that the influence of meiotic recombination on substitution patterns is weaker in the mouse lineage than in the human lineage, since correlation coefficients are lower in the mouse lineage. Also, in the mouse lineage, the correlation coefficients between crossover rates and GC-content and between crossover rates and GC* are much closer than in the human lineage (Table 4.1 & 4.2). One possible explanation is that the mouse genome has lower meiotic recombination rates than the human genome (human median crossover rate = 1.16 cM/Mb, mouse median sex-averaged crossover rate = 0.65 cM/Mb, Figure A.1).

	Sex-averaged	**LDT**
	R	R
GC-content	0.354^{***}	-0.453^{***}
GC*	0.631^{***}	-0.604^{***}

Table 4.1.: Pearson correlation coefficients between substitution rates, crossover rates and LDT in human. ***p-value $< 10^{-10}$

	Sex-averaged	**Male-specific**	**Female-specific**	**LDT**
	R	R	R	R
GC-content	0.186^{***}	0.240^{***}	0.085^{*}	-0.304^{***}
GC*	0.196^{***}	0.275^{***}	0.110^{*}	-0.305^{***}

Table 4.2.: Pearson correlation coefficients between substitution rates, crossover rates and LDT in mouse. *p-value < 0.05; ***p-value $< 10^{-10}$

Furthermore, in the mouse lineage we can see that male-specific crossover rates correlate more strongly with current GC-content or GC* than sex-averaged or female crossover rates do (Tables 4.2 & A.2). This indicates that male recombination has more influence on substitution patterns than female recombination in the mouse lineage, as well as was previously observed in the human lineage (Duret and Arndt, 2008; Webster et al., 2005). We therefore focused on male-specific crossover rates in the mouse lineage for the remainder of the chapter.

The effective population size (N_e) of mice is around 30 times greater than that of humans: it is estimated to be around 20,000 in humans and around 600,000 in mouse (Keightley et al., 2005). gBGC should therefore be stronger in the mouse lineage compared to the human lineage because gBGC has a bigger impact in species with larger effective population sizes (Nagylaki, 1983). However, the effect of gBGC appears to be weaker in mouse lineage compared to the human lineage. We cannot claim, however, that gBGC is generally absent in the mouse lineage as there are reported cases showing clear evidence of gBGC inside the mouse genome (Montoya-Burgos et al., 2003).

There are 4 possible explanations for the fact that gBGC is weaker in the mouse lineage compared to the human lineage. First, recombination rates are lower in the mouse genome compared to the human genome (Jensen-Seaman et al., 2004; Li and Freudenberg, 2009), which will cause gBGC to be weaker in the mouse lineage compared to the human lineage. Second, it cannot be excluded that recombination events are repaired more often into crossovers than non-crossovers in the human genome compared to the mouse genome. This can cause crossover rates to be a less accurate proxy of meiotic recombination in the mouse genome compared to the human genome. Third, it is possible that the heteroduplex length that forms during gene conversion is shorter in mouse than in human. This will cause gBGC to affect fewer bases in mouse compared to human. Finally, the mismatch repair mechanism can be less biased towards G and C bases in the mouse genome compared to the

human genome. This will cause the fixation bias favoring G and C bases to be lower in mouse compared to humans. Unfortunately, we currently cannot test the last three hypotheses.

We would like to point out that the fact that recombination rates evolve rapidly in mouse species (Dumont et al., 2011) could affect our results. One way to solve this issue would be to study substitution patterns in the mouse lineage by comparing two closely related mouse species, using rat as an outgroup. The recent publication of the genomic sequence of several mouse laboratory strains, including the two subspecies *Mus m. castaneus* and *Mus spretus* (Keane et al., 2011) can help solve this issue. Chapter 5 presents results obtained using these new genomic sequences.

4.2.3. Substitution patterns are under the influence of male-specific recombination

Our results show that, in the mouse lineage, male-specific crossover rates are a better predictor of substitution patterns than female-specific crossover rates. This might imply that male-specific recombination have more impact on substitution patterns than female-specific recombination does. This has been previously reported in the human lineage (Duret and Arndt, 2008; Webster et al., 2005) and appears to be shared with dog and sheep whereas in pig and opossum female-specific recombination has more impact on substitution than male-specific recombination (Popa et al., 2012). We can put forward two hypotheses to explain these observations. First, the distribution of recombining regions along chromosomes is different for male and female-specific recombination, both in the human genome and in the mouse genome (Myers et al., 2005; Paigen et al., 2008). Female recombining regions are more numerous and more homogeneously distributed along chromosomes than male recombining regions. On the other hand, male recombination hotspots are more active. This more heterogeneous distribution of recombination in males may lead to the fact that male-specific recombination rates predict substitution patterns better than female-specific recombination rates. Second, meiotic recombination events cause the formation of Holliday Junctions which are resolved either into crossovers (COs) or non-crossovers (NCOs) (Smith and Nicolas, 1998; de Massy, 2003; Baudat and de Massy, 2007). Genetic maps available for the human and mouse genomes do not have enough resolution to show non-crossovers. It is possible that crossovers represent a greater proportion of recombination in males than in females. One alternative is to measure the frequency of double strand breaks in genomic regions and use these as a proxy measure of meiotic recombination. The recent publication of double strand breaks hotspots (Smagulova et al., 2011) enables us to perform such analysis, which is presented in Chapter 5.

Pink and Hurst (2011) reported a more complex relationship between substitution rates, crossover rate, GC-content and replication-timing in mouse introns. They showed a correlation between replication-timing and meiotic recombination that is negative for male-specific recombination but positive for female-specific re-

combination. They argue that this different sign of correlation is responsible for an underestimation of the influence of female-specific recombination on substitution rates. However, these authors measured total substitution rates but not rates affecting GC-content evolution (W→S and S→W substitution rates). Thus, it is unclear how the correlation between female-specific crossover rates and replication timing will affect measures of the influence of female-specific recombination on GC-content evolution.

4.3. Multiple genomic features influence substitution patterns

Several studies have shown a link between substitution rates and genomic features, such as replication-timing (Pink and Hurst, 2010; Chen et al., 2010) or CpG content (Walser et al., 2008; Walser and Furano, 2010). Thus, crossover rates are not the sole feature influencing substitution patterns in mouse and human. We investigated the link between the 9 following features and substitution patterns in both mouse and human lineages: GC-content (hereafter designated as GC), LCO (logarithm of crossover rates), LDT (distance to telomeres), replication-timing (hereafter designated as RepTime), exon density (fraction of a window occupied by exons, hereafter designated as Exons), transposable elements densities (hereafter designated as SINEs, LINEs and LTRs) and CpG odds ratio (the observed CpG frequency normalized by the expected CpG frequency, hereafter designated as CpGodds). Male crossover rates were used in the mouse lineage. We used the logarithm of all densities (Exon, LINEs, SINEs and LTRs) in both lineages.

We first computed correlation coefficients between each feature and substitution rates (Tables A.3 & A.4). However, since genomic features are also correlated with each other, this can affect correlations between genomic features and substitution rates as follows. Let's imagine two features A and B which are strongly correlated. A correlation between A and substitution rates for example will thus be affected by the correlation between A and B. As a result, we will not be able to decisively conclude whether there is a link between A and substitution rates. For example, we see that in the mouse lineage GC* and SINEs density are strongly correlated ($R = 0.734$, p-value $< 10^{-15}$). However, since both are strongly correlated with Exons ($R = 0.625$ and 0.711 respectively, both p-values $< 10^{-15}$), this can affect correlations with substitution rates. Indeed, the correlation between GC* and SINEs density controlling for Exons is lower ($R = 0.528$, p-value $< 10^{-15}$) (the R code used to compute partial correlation is taken from Drummond et al. (2006)). We therefore analyzed the link between substitution patterns and genomic features using a multivariate approach to be able to take all links between features into account.

4.3.1. Relative Contribution to Variability Explained

The first approach to study how much individual genomic features influence substitution patterns is based on linear modeling. The idea is to build a linear model where a variable (the substitution rate) is explained by the 9 genomic features listed above, and then to measure the individual contribution of each feature to the model. We applied this method in both the human and mouse lineages for GC* values, W→S, S→W, W→W and S→S substitution rates as well as the CpG→TpG/CpA substitution rate (hereafter designated as CpG Rate) and the total substitution rate (here after designated as Total Rate). This method, named *RCVE* (for Relative Contribution to Variability Explained), has been used before to analyze substitution patterns as well as transposable elements dynamics in primates (Kvikstad et al., 2007; Tyekucheva et al., 2008; Kvikstad and Makova, 2010).

For each substitution rate, we built a linear model where the substitution rate is the response variable and the 9 genomic features the explanatory variables. We extracted the coefficient of determination (R^2) from this model (named R^2_{full}). In order to assess the significance of each feature in the model, we then shuffled the labels of the feature of interest and recomputed the linear model using this shuffled feature. We extracted the coefficient of determination of this reduced model (named R^2_{reduced}). This value was finally normalize by the R^2_{full} of the full model to compute the *RCVE* value corresponding to the feature (See the Material & Methods chapter for more details)

$$RCVE = \frac{R^2_{\text{full}} - R^2_{\text{reduced}}}{R^2_{\text{full}}} \ . \tag{4.1}$$

Results for the human and mouse lineages are shown in Tables 4.3 and 4.4 respectively. The *RCVE* values for all genomic features in both lineages are plotted in Figure 4.2.

	GC	**LCO**	**LDT**	**RepTime**	**Exons**	**SINEs**	**LINEs**	**LTRs**	**CpGodds**	R^2
GC*	*NA*	0.248	0.084	0.008	0.002	*NA*	0.006	0.027	0.006	0.597
W→S	0.014	0.208	0.086	0.008	0.037	0.024	*NA*	0.001	0.036	0.598
S→W	0.018	0.005	0.001	0.028	0.086	0.029	0.003	0.048	0.024	0.581
W→W	0.063	0.040	0.009	0.016	0.122	0.014	*NA*	0.017	0.049	0.458
S→S	0.047	0.095	0.067	0.019	0.085	0.012	*NA*	0.004	0.094	0.354
CpG Rate	0.006	*NA*	*NA*	0.013	0.054	0.004	*NA*	0.028	0.011	0.658
Total Rate	*NA*	0.094	0.039	0.028	0.078	0.032	*NA*	0.021	0.079	0.570

Table 4.3.: *RCVE* results for the human lineage. *NA*: $RCVE < 0.001$

We see that the W→S substitution rate is predicted by different features in the mouse lineage and the human lineage. In the human lineage, crossover rates (designated as LCO) is the strongest predictor of the W→S substitution rate (*RCVE* = 0.208, Figure 4.2, Table 4.3). The slope of the regression line between this substitution rate and LCO is positive (Slope = 0.424, Figure A.3, Table A.7), which

can be interpreted as a positive influence of meiotic recombination on substitution rates, possibly mediated by gBGC. In contrast, in the mouse lineage, this rate is not predicted by LCO nor by LDT ($RCVE < 0.001$ for both features, Table 4.4): the strongest predictor of this substitution rate is the CpG odds ratio (designated as CpGodds, $RCVE = 0.767$, Figure 4.2, Table 4.4). The association between the CpG odds ratio and this substitution rate is positive as the slope of the regression line is positive (Slope = 1.076, Figure A.3, Table A.8). This confirms that in the mouse lineage, gBGC is weak and has a small impact on substitution rates compared to other features.

	GC	**LCO**	**LDT**	**RepTime**	**Exons**	**SINEs**	**LINEs**	**LTRs**	**CpGodds**	R^2
GC*	0.021	*NA*	*NA*	0.004	*NA*	0.004	0.005	*NA*	0.248	0.860
W→S	0.005	*NA*	*NA*	0.006	0.022	0.034	*NA*	0.023	0.767	0.541
S→W	0.056	*NA*	*NA*	0.010	0.009	0.038	0.014	0.009	0.061	0.791
W→W	0.126	0.001	0.003	0.038	0.019	0.096	0.028	0.011	*NA*	0.596
S→S	0.113	*NA*	*NA*	0.009	0.041	0.091	0.022	0.017	0.041	0.608
CpG Rate	0.013	*NA*	*NA*	0.006	0.045	0.008	0.003	0.005	0.029	0.456
Total Rate	0.022	*NA*	*NA*	0.005	0.059	0.148	0.007	0.061	0.178	0.409

Table 4.4.: *RCVE* results for the mouse lineage. *NA*: $RCVE < 0.001$

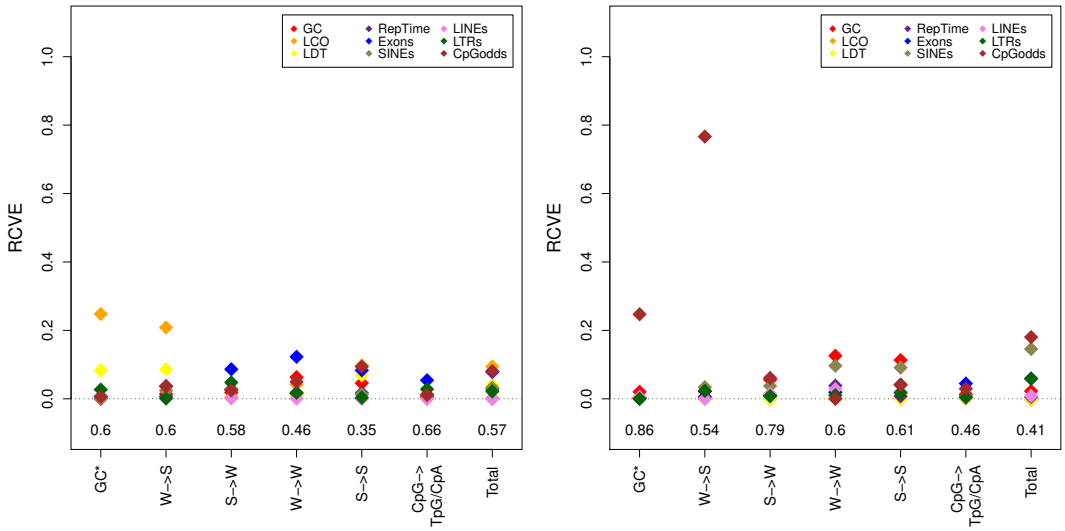


Figure 4.2.: *RCVE* results for substitution patterns in the human (left panel) and mouse (right panel) lineages. Total R^2 values for each linear model is indicated below.

Other substitution rates are also predicted by different features in both the mouse human lineages. In the human lineage, exons density is the best predictor of S→W substitution rates as well as the CpG→TpG/CpA rate ($RCVE = 0.086$ and 0.054 respectively, Figure 4.2, Table 4.3). This can be interpreted as an effect of transcription and gene expression on mutation patterns: more highly expressed genes will have lower mutation rates. However, in the mouse lineage, the S→W rate is most strongly

predicted by CpG odds ratio and GC-content ($RCVE = 0.061$ and 0.056 respectively, Figure 4.2, Table 4.4). Finally, like in the human lineage, the CpG→TpG/CpA rate is most strongly predicted by exon density in the mouse lineage ($RCVE = 0.045$, Figure 4.2, Table 4.4).

Results obtained using the *RCVE* method have to be tempered by the fact that this method does not perform well if genomic features are inter-correlated. We expect the sum of all *RCVE* values for one substitution rate to be equal to 1. It is, however, very rarely the case (for example, *RCVE* values sum up to 0.197 for the S→W substitution rate in the mouse lineage, which indicates that genomic features are highly correlated, Table 4.4). We therefore have to use another method to analyze the relative influence of genomic features on substitution patterns in both the human and mouse lineages.

4.3.2. Principal Component Regression

As the genomic features studied are correlated in both the human and mouse genomes (Tables A.5 & A.6), results coming from methods based on linear modeling such as *RCVE* have to be taken cautiously and improved. One way to solve this issue is to de-correlate the features before doing any analysis.

Principal component analysis (hereafter designated as PCA) provides an interesting tool to perform such task: this method will project each variable on the same number of axes, decomposing the data into an identical number of principal components, each being orthogonal from one another and being a linear combination of the different features. We can then study the correlation between each component and variables of interest (such as substitution patterns) individually using linear modeling. We therefore merged principal component analysis and linear regression into principal component regression (hereafter designated as PCR).

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
GC	-0.406	0.110	-0.041	0.123	-0.032	0.335	-0.098	**0.632**	**-0.534**
LCO	-0.149	**0.585**	**-0.539**	**-0.544**	-0.033	-0.138	0.009	0.077	0.154
LDT	0.197	**-0.606**	-0.294	-0.412	0.153	0.400	0.018	0.302	0.249
RepTime	0.326	0.357	0.285	-0.115	-0.358	**0.633**	-0.341	-0.114	0.123
Exons	-0.335	-0.237	-0.355	0.173	**-0.721**	0.158	0.234	-0.270	0.053
SINEs	-0.390	-0.190	-0.163	-0.013	0.195	0.003	**-0.790**	-0.350	-0.012
LINEs	0.383	-0.083	-0.171	0.198	-0.419	**-0.451**	-0.440	0.443	0.097
LTRs	0.307	0.196	**-0.569**	**0.593**	0.324	0.275	0.033	-0.098	0.019
CpGodds	-0.403	0.117	0.185	0.292	0.089	0.068	0.006	0.301	**0.776**
% of variance	0.544	0.171	0.072	0.062	0.049	0.038	0.025	0.023	0.017

Table 4.5.: Entries of the eigenvectors for all 9 principal components in the human genome. Features that contribute for at least 20% of the component are indicated in bold. Entries of each eigenvector were normalized such as $\sum \text{entries}^2 = 1$.

We first carried out one PCA in each lineage on the 9 genomic features listed above, in order to transform them into 9 independent (or orthogonal) principal

components (designated as PC, see the Materials & Methods chapter for more details). Figure A.4 shows the eigenvectors of the first two principal components in the human and mouse lineages, Tables 4.5 and 4.6 detail the eigenvectors for each principal component.

The percentage of variance in both tables indicates the proportion of the total variance of all genomic features contained in the corresponding principal component. We see that while some components do not have major contributors (like the first components), others are dominated by one or two features (for example PC6 in the mouse genome, dominated by the CpG odds ratio, Table 4.6). It is interesting to note that the first two components are quite similar in both genomes. No feature truly dominates the first component whereas both LCO and LDT are major contributors to the second component. However, the second component differs in both genomes with LTRs being a major contributor in the mouse lineage, unlike in the human genome. On the other hand, other components differ greatly between the two genomes.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
GC	-0.418	-0.030	0.009	0.097	-0.302	0.265	-0.397	-0.378	**-0.595**
LCO	-0.150	**0.485**	**0.493**	**-0.703**	-0.026	0.060	0.032	-0.009	-0.001
LDT	0.159	**-0.612**	-0.253	**-0.644**	-0.267	-0.174	-0.136	-0.034	-0.019
RepTime	0.387	0.235	-0.009	0.021	0.078	-0.116	-0.779	0.396	-0.103
Exons	-0.355	-0.265	-0.033	-0.152	**0.791**	0.217	-0.273	-0.076	0.163
SINEs	-0.419	-0.132	-0.060	-0.030	-0.088	0.138	0.176	**0.828**	-0.242
LINEs	0.415	-0.101	0.148	-0.059	0.416	-0.058	0.307	-0.021	**-0.722**
LTRs	0.036	**-0.483**	**0.816**	0.225	-0.115	-0.048	-0.118	0.065	0.122
CpGodds	-0.391	0.059	0.015	0.054	0.107	**-0.901**	-0.040	-0.049	-0.116
% of variance	0.529	0.140	0.108	0.080	0.045	0.035	0.034	0.018	0.009

Table 4.6.: Entries of the eigenvectors for all 9 principal components in the mouse genome. Features that contribute for at least 20% of the component are indicated in bold. Entries of each eigenvector were normalized such as $\sum \text{entries}^2 = 1$.

We then used these components to build multivariate linear regressions for W→S and S→W substitution rates as well as GC* values, where the substitution rate is the response variable and the components are the predictors, and computed how much of the variable's variance each principal component predicts (see the Materials and Methods chapter for details). It should be noted that we used principal component analysis only to de-correlate different genomic features. As such, we took all principal components into account in the linear modeling, even components that explain an almost insignificant proportion of all features' total variance.

4.3.3. CpG odds ratio is the main predictor of W→S substitution rates in the mouse lineage

We observe that in both human and mouse lineages, substitution patterns are predicted by different features. In the human lineage, the W→S substitution rate is

most strongly predicted by PC2, which is mostly composed of LCO and LDT, two proxy measures of meiotic recombination ($R^2 = 0.555$, Figure 4.3, Tables 4.5 & 4.7). This result reflects the influence of gBGC on W→S substitution. In the mouse lineage, PC6, which is dominated by CpG odds ratio rather than by measures of meiotic recombination, most strongly predicts the W→S substitution rate ($R^2 = 0.380$, Figure 4.3, Tables 4.6 & 4.8) This result reflects the fact that gBGC only has a very limited impact on W→S substitutions in the mouse lineage. Other principal components like the first component, also explain a small proportion of the variance of the W→S substitution rate in the mouse lineage ($R^2 = 0.10$, Figure 4.3, Table 4.8). All these results confirm results obtained with the *RCVE* method.

	PC1 R^2	**PC2** R^2	**PC3** R^2	**PC4** R^2	**PC5** R^2	**PC6** R^2	**PC7** R^2	**PC8** R^2	**PC9** R^2	**Total** R^2
GC*	0.290^{***}	0.243^{***}	0.006^{*}	0.010^{**}	0.011^{**}	0.037^{***}	*NA*	0.001	0.001	0.598^{***}
W→S	0.007^{*}	0.555^{***}	0.002	0.001	0.001	0.013^{**}	0.002^{*}	*NA*	0.018^{***}	0.600^{***}
S→W	0.373^{***}	0.148^{***}	0.012^{**}	0.002^{*}	0.026^{***}	0.003^{*}	0.003^{***}	*NA*	0.016^{**}	0.583^{***}
W→W	0.209^{***}	0.177^{***}	0.013^{**}	*NA*	0.029^{***}	0.002^{*}	0.001	*NA*	0.029^{***}	0.460^{***}
S→S	0.025^{***}	0.275^{***}	0.014^{**}	*NA*	0.007^{*}	0.006^{*}	*NA*	*NA*	0.029^{***}	0.356^{***}
CpG Rate	0.605^{***}	0.036^{***}	0.002^{*}	*NA*	0.014^{**}	*NA*	*NA*	0.001	0.001	0.659^{***}
Total Rate	0.002^{*}	0.513^{***}	0.011^{**}	0.004^{*}	0.010^{**}	0.001	0.001	0.013^{**}	0.017^{***}	0.572^{***}

Table 4.7.: Results of principal component regression on substitution patterns in the human lineage. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$
NA: $R^2 < 0.001$

	PC1 R^2	**PC2** R^2	**PC3** R^2	**PC4** R^2	**PC5** R^2	**PC6** R^2	**PC7** R^2	**PC8** R^2	**PC9** R^2	**Total** R^2
GC*	0.713^{***}	0.001	0.001	0.001	0.002	0.118^{***}	0.008^{*}	*NA*	0.018^{**}	0.861^{***}
W→S	0.100^{***}	0.023^{**}	0.024^{**}	0.006^{*}	0.001	0.380^{***}	0.003^{*}	0.006^{*}	*NA*	0.544^{***}
S→W	0.724^{***}	0.001	0.003^{*}	*NA*	0.003^{*}	0.007^{*}	0.014^{**}	0.004^{*}	0.036^{***}	0.792^{***}
W→W	0.468^{***}	0.009^{*}	0.003^{*}	0.003^{*}	0.001	0.019^{**}	0.031^{***}	0.012^{*}	0.053^{***}	0.598^{***}
S→S	0.441^{***}	0.015^{**}	0.007^{*}	0.001	0.002	0.082^{***}	0.013^{**}	0.009^{*}	0.040^{***}	0.611^{***}
CpG Rate	0.439^{***}	0.004^{*}	0.001	*NA*	0.012^{*}	0.001	*NA*	*NA*	0.003^{*}	0.459^{***}
Total Rate	0.205^{***}	0.023^{**}	0.030^{***}	0.006^{*}	0.004^{*}	0.117^{***}	0.002	0.017^{**}	0.008^{*}	0.412^{***}

Table 4.8.: Results of principal component regression on substitution patterns in the mouse lineage. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$
NA: $R^2 < 0.001$

Our results show that in the mouse lineage, CpG odds ratio (the observed CpG frequency divided by the expected CpG frequency) is the main predictor of W→S substitution rates, unlike in the human lineage.

One might be tempted to interpret these results as due to CpG odds ratio being a proxy measure of meiotic recombination. A link between DNA methylation (which occurs on cytosines of CpG dinucleotides) and meiotic recombination has already been described in the human genome (Sigurdsson et al., 2009). Moreover, in the mouse lineage, we observe an association between male crossover rates and CpG odds ratio (partial correlation = 0.14, p-value $< 10^{-7}$ when controlling for GC-content).

However, our results show us that CpG odds ratio predicts W→S substitution rates independently of meiotic recombination.

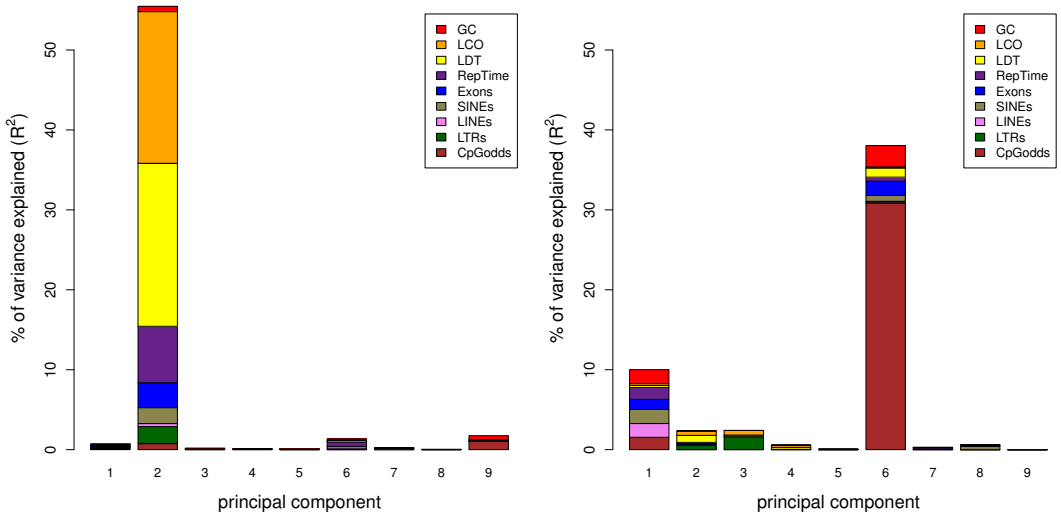


Figure 4.3.: Principal component regression for W→S substitution rates in the human (left panel) and the mouse (right panel) lineages. The height of each bar represents how much of the variable's variance is explained by the corresponding component. Each colored area is proportional to the relative importance of the corresponding feature inside a component.

There are two explanations for our results. First, it is possible that recombination decreases the CpG→TpG/CpA rate by protecting CpG dinucleotides from decaying into TpG or CpA dinucleotides. If this was the case, one should see a negative influence of meiotic recombination on the CpG rate. Since meiotic recombination is not the strongest predictor of the CpG →TpG/CpA substitution rate (Table 4.8), meiotic recombination does not seem to protect CpG dinucleotides.

Second, meiotic recombination could occur mostly in CpG-rich regions, e.g. CpG islands. However, no link between recombination hotspots and CpG islands has been proposed in the mouse or the human genome (Paigen et al., 2008; Myers et al., 2005). Also, the consensus DNA motif associated with hotspot activity is not CpG-rich (Myers et al., 2008). Furthermore in the mouse lineage, the PCA results show that meiotic recombination and CpG odds ratio contribute to two independent components, and only the latter component predicts W→S substitution rates (Tables 4.6 & 4.8). This shows that in the mouse lineage, CpG odds ratio predicts substitution patterns independently from meiotic recombination.

We cannot tell, however, if CpG content has a direct influence on W→S substitution rates or if CpG content serves as a proxy measure for genomic features we did not include in our model or if there is no cause and effect relationship between CpG content and W→S substitution rates. Authors have proposed that, in the human lineage, CpG content and substitution rates are associated through different mechanisms such as chromatin opening linked to gene expression or error-prone repair of T:G mismatches by different DNA polymerases (Walser and Furano, 2010).

Moreover, they have found no evidence that this association is mediated through fixation probabilities of mutations. The relationship between CpG content, substitution rates and other genomic features needs to be further investigated in both human and mouse lineages.

We have found that unlike in the human lineage, gBGC is weak in the mouse lineage and that CpG odds ratio, not meiotic recombination is the strongest predictor of W→S substitution rates. This reveals that isochore structures evolve differently in both human and mouse lineages and seems to indicate that this is the result of substitution patterns being under different influences in those lineages.

4.3.4. S→W substitution rates are predicted by a combination of features in both human and mouse lineages

In the human lineage, the S→W substitution rate is most strongly predicted by the first two principal components ($R^2 = 0.373$ and 0.148 respectively, Figure 4.4, Tables 4.7). In contrast, in the mouse lineage, the S→W substitution rate is most strongly predicted by the first principal component ($R^2 = 0.724$, Figure 4.4, Table 4.8).

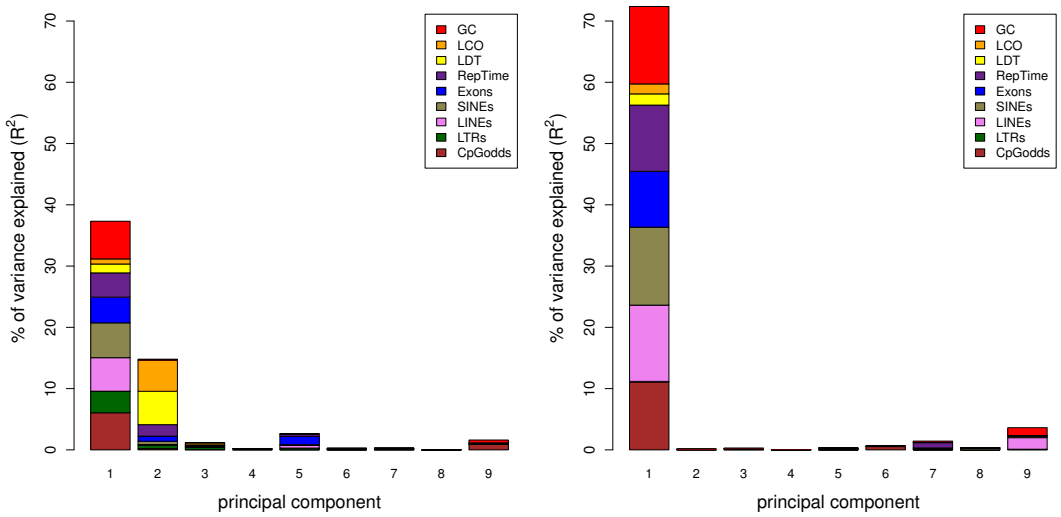


Figure 4.4.: Principal component regression for S→W substitution rates in the human (left panel) and the mouse (right panel) lineages. The height of each bar represents how much of the variable's variance is explained by the corresponding component. Each colored area is proportional to the relative importance of the corresponding feature inside a component.

Principal component regression results show that S→W substitution rates in both lineages are mostly predicted by a component which is a combination of different features (GC-content, exon density, replication-timing, transposable element densities). These results can be interpreted in different ways. First, it is possible that natural selection affects the fixation probabilities for the substitution rates we computed. Because we masked regions affected by natural selection in our windows (exons),

we assume that it does not play a role on substitutions and that nucleotides are evolving neutrally in our windows.

It is also possible that meiotic recombination influences the fixation probabilities of substitution rates through gBGC. However, because meiotic recombination is not the strongest predictor of these substitution rates and it constitutes only a small fraction of this component, we assume that meiotic recombination has a low impact on fixation probabilities for S→W substitution rates and that these substitution rates are equal to mutation rates and therefore interpret these results as the influence of mutation on substitution patterns. We cannot tell, however, if the features predicting S→W substitution rates have a direct impact on substitution patterns or if the associations we observe are not cause and effect associations.

4.3.5. Comparison of *RCVE* and PCR results

We compared results obtained with the *RCVE* and PCR methods.

We first see that results for W→S substitution rates in both human and mouse lineages are similar, with meiotic recombination being their best predictor in the former whereas CpG odds ratio being their best predictor in the latter. We can clearly see both features standing out as best predictors in both lineages in *RCVE* methods as well as PCR results.

On the other hand, we see that results disagree for other substitution rates at a first glance. Notably, whereas according to *RCVE* exons density and GC-content best predict S→W substitution rates in the human and mouse lineages respectively, a combination of features best predicts these rates in both lineages according to PCR. We notice that even though the aforementioned features have the best *RCVE* values, other features have close if not similar values. These results already hint towards a complex relationship between S→W substitution rates and genomic features. PCR results clearly show such complex relationships, as the components used in linear regressions are linear combinations of genomic features.

In the light of these results, it appears that *RCVE* and PCR methods complement each other quite well. When a genomic feature clearly best predicts substitution rates, it will be spotted by both methods. On the other hand, in the case of complex relationships, no clear predictor will appear in *RCVE* results when at the same time a component comprising of several features will appear as the best predictor in PCR results. Using both methods will yield best results when investigating the links between genomic features and evolutionary rates like insertion and deletion rates or substitution rates.

4.3.6. The effect of replication-timing on substitution patterns

Chen et al. (2010) reported a link between substitution patterns and genomic features such as meiotic recombination, GC-content and replication-timing. Their re-

sults differ significantly from ours as they report replication-timing as a strong predictor of substitution rates, notably S→W rates. These differences can be explained by the following reasons.

First, Chen et al. (2010) use replication-timing data based on massively parallel sequencing of replicating DNA labeled with BrdU, a fluorescent marker of newly synthesized DNA, whereas we used timing data from a microarray-based comparison of early and late replicating DNA (Ryba et al., 2010). Second, although based on linear regressions, our multivariate methods also differ. Third, these authors analyzed substitution rates and features in 100 kbp long windows, whereas we analyzed 1 Mbp long windows. Finally, their analysis comprised of 4 genomic features (GC-content, crossover rates, distance to telomeres and replication-timing) whereas we added 5 other features to our study (exons density, transposable elements density and CpG odds ratio).

The different results could then easily be interpreted as different influences of features on substitution rates at different scales, replication-timing having a stronger effect when studying small window sizes. The different number of features could also be a good explanation for these differences. It should be noted that R^2 values associated with substitution rates where replication-timing has a strong effect (S→W substitution rates notably) are quite low, whereas R^2 values we obtained for similar substitution rates are relatively higher (around 0.20 and 0.58 respectively). This can be explained mainly to the fact that we included more features in our study and therefore were able to obtain a more complete picture of how much different features influence substitution rates. The differences in window size can also explain this, as substitution rates inferred in smaller windows will tend to be more variable and therefore decrease R^2 values in linear models.

4.3.7. Outgroup choice

The method we use to infer substitution rates in one lineage uses triple alignments: it compares two sister species and uses an outgroup to infer the two sister's ancestral state. We compared human to chimpanzee and used macaque as an outgroup for the analysis of the human lineage. As at the time of analysis the only two rodents genomes fully sequenced were mouse and rat, we compared these two species and used human as an outgroup to study substitution patterns in the mouse lineage. The mouse - rat - human divergence time is between 85 and 95 million years (Myrs), whereas that of mouse and rat is between 16 and 19 Myrs (Poux et al., 2006; Huchon et al., 2007). Although using human as an outgroup may lead to incorrect inference of substitution rates in the mouse lineage, this was chosen as an outgroup as it was the closest available high coverage genome to mouse and rat. One of the closest related species to mouse and rat, whose complete genome was published and aligned to other placentals at the time of the analysis, was the guinea pig (*Cavia porcellus*). It is however a 6.79x low coverage genome (Ensembl version 56, Hubbard et al., 2009). Furthermore, the divergence time between mouse, rat and guinea pig is around 60

Myrs (Poux et al., 2006; Huchon et al., 2007), which is close to the mouse, rat and human divergence time. Preliminary results obtained using guinea pig or kangaroo rat as outgroups were very similar to results obtained using human as an outgroup (data not shown). Moreover, mouse rat human triple alignments are much cleaner and contain more sites where the three species share a nucleotide than alignments with guinea pig or kangaroo rat. We therefore used mouse - rat - human triple alignments to infer substitution patterns in the mouse lineage. Also, the method we used to infer substitution rates (Arndt et al., 2003a; Arndt and Hwa, 2005; Duret and Arndt, 2008) is based on maximum likelihood, which makes it robust to long lineage as it allows multiple substitutions at each site. It also does not imply that the substitution process is time-reversible nor that it is at a stationary state, both assumptions that can bias the inference of substitution patterns (Squartini and Arndt, 2008). Finally, it infers one substitution pattern for each of the four branches of the rooted tree ((sister 1, sister 2), outgroup).

4.3.8. Differences in branch length and genetic map resolutions

Our results could be affected by the different time-spans that substitution patterns reflect in both human and mouse lineages: human and chimpanzee diverged around 6 Myrs ago whereas mouse and rat diverged between 16 and 19 Myrs ago. Crossover rates computed in the mouse genome may not well reflect past recombination as mouse and rat genomes underwent frequent chromosomal rearrangements, which affected their chromosomal recombination patterns. Moreover, the outgroup for the analysis of the mouse lineage (human) is very distant whereas the outgroup for the analysis of the human lineage (macaque) is much closer: mouse and human diverged between 85 and 95 Myrs ago whereas human and macaque diverged between 27 and 33 Myrs ago (Poux et al., 2006; Huchon et al., 2007). Another potential source of bias is the different densities of genetic maps available for human an mouse: the mouse maps contain between 10,000 and 11,000 markers on autosomes (approximately one marker every 250 kbp, Shifman et al., 2006) whereas the human map contains more than 3 million markers (International HapMap Consortium et al., 2007). To control for all these sources of bias, we performed the following analyses. We computed substitution patterns in the branch between the human-macaque ancestor and human (hereafter designated as the *HCM* branch), using mouse as an outgroup. At the same time, we computed new crossover rates as follows: we generated a low-density human genetic map by sampling 11,000 random markers from the original map and re-computed crossover rates as described in the Materials & Methods chapter.

Results obtained for this *HCM* branch are very similar to results obtained with the branch between the human-chimpanzee ancestor branch (hereafter designated as the *HC* branch). First, even though correlation coefficients between crossover rates, GC-content and GC* are slightly lower for the *HCM* branch than for the *HC* branch, the correlation between crossover rates and GC* is stronger than the correlations

between crossover rates and GC-content (Table 4.9). Moreover, LDT correlates more strongly with GC-content and GC* in the *HCM* branch than in the mouse lineage (Table 4.9). Second, principal component regression results of the *HC* branch and the *HCM* branch were very similar: in this branch, the second component is the main predictor of W→S substitution rates, whereas the first component is the main predictor of S→W substitution rates (Figures A.6 to A.8, Table A.9). We therefore conclude that our results are not affected by the different time-spans between human and mouse lineages nor by different density of genetic maps.

	***HC* Branch**		***HCM* Branch**	
	Crossover rates	LDT	Crossover rates	LDT
GC-content	0.416***	-0.453***	0.340***	-0.436***
GC*	0.676***	-0.604***	0.389***	-0.582***

Table 4.9.: Pearson correlation coefficients (R) between crossover rates, distance to telomeres, GC-content and GC* in the *HC* and *HCM* branches. Crossover rates in the *HC* branch were computed using the original HapMap genetic map. Crossover rates in the *HCM* branch were computed using the low-density genetic map as described in the main text. All crossover rates are sex-averaged crossover rates. ***p-value $< 10^{-10}$

4.3.9. Cryptic variations of mutation rates

It has been shown that the mutation process is not homogeneous: mutations rates vary along the human genome. This has been called the cryptic variation of the mutation process (Hodgkinson et al., 2009) and can cause a bias in our substitution pattern inference and affect our results. We tested this possibility by conducting sequence evolution simulations and comparing evolution of three classes of sequences (GC-rich, GC-medium and GC-poor) in the presence or absence of hypermutable sites (see the Materials & Methods chapter for more details).

Inferring substitution patterns in the presence of cryptic variations of mutation rates

Random sequences of 500 kbp were generated, with a GC-content of either 0.25 (GC-poor), 0.50 (GC-medium) or 0.75 (GC-rich). 25,000 hypermutable sites were randomly chosen within these sequences, which have a total mutation rate 10 times higher than normal sites. These sequences served as the root sequences and were made evolve along each of the four branches of the following rooted tree: ((sister1, sister2), outgroup), Figure 4.5).

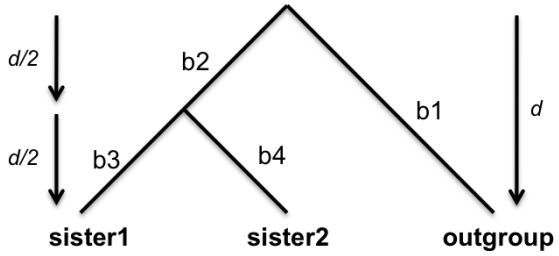


Figure 4.5.: Rooted tree used for sequence evolution simulation.

A Jukes-Cantor model was used to compute transition probabilities (see the matrix of transition probabilities $P(t)$ in section 2.2.1).

Within each branch, transition probabilities were computed for hypermutable sites and normal sites such as transition probabilities in hypermutable sites are ten times higher than in normal sites and the total divergence is equal to 0.15 in branch 1 and to 0.075 in branches 2, 3 and 4. By doing so, both sisters and the outgroup are simulated to evolve for the same amount of time ($d/2 + d/2 = d$, Figure 4.5). Substitution patterns were then inferred in branches 3 and 4 using the same maximum likelihood-based method as before using sequences obtained in the simulations. The GC-content of GC-rich, GC-medium and GC-poor regions evolved to mean values of 0.70, 0.50 and 0.30 respectively in both sister species.

Cryptic variations of mutation rates do not affect our interpretations

In the absence of hypermutable sites, we observe that for all three categories of sequences, the GC* is correctly inferred to be of 0.50 (Figure 4.6). In the presence of hypermutable sites, we observe that if the GC-content is at equilibrium, there is no effect on GC* inference. However, GC* values are under-estimated if the GC-content is lower than the GC* and over-estimated if the GC-content is higher than the GC*: given our simple Jukes-Cantor mutation model, all sequences should evolve towards a GC* of 0.50. The method we use computed a GC* of 0.60 for GC-rich regions and a GC* of 0.40 for GC-poor regions (Figure 4.6).

This bias due to cryptic variation of the mutation rate is linear with respect to GC-content. However, the linear transformation of a variable does not affect correlation coefficients. Also, we observe that the relationship between GC-content and GC* is non linear in the mouse lineage (Figure 4.1). At the light of this, we believe that correlations coefficients between GC-content, GC* and crossover rates are not affected by cryptic variation of the mutation rate and that this process alone therefore cannot explain results obtained in the mouse lineage.

4.4. Conclusion

In this chapter, we investigated substitution patterns and GC-content evolution in both human and mouse lineages and were able to highlight major differences be-

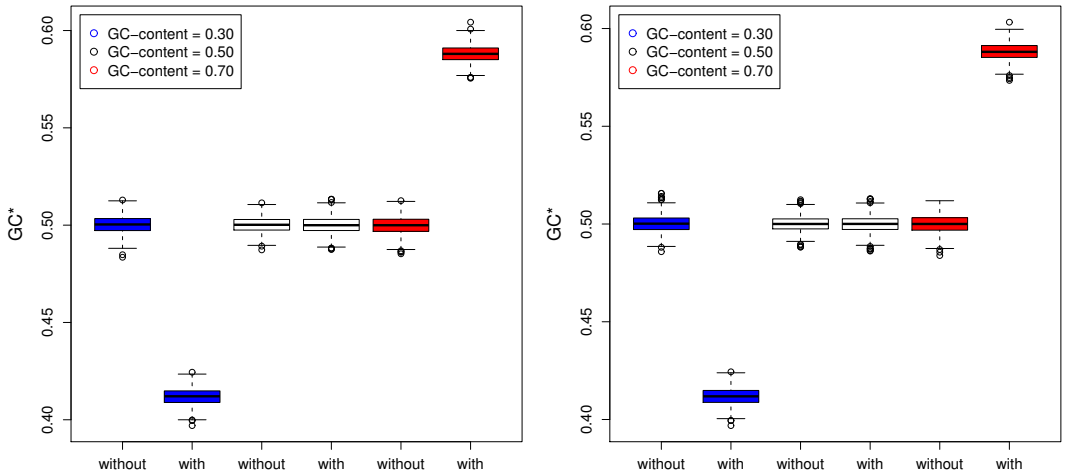


Figure 4.6.: GC* estimations after sequence evolution simulation in the presence or absence of cryptic variation in mutation rates in two sister lineages.

tween these two lineages. We first showed that despite having different GC-content evolution dynamics, both lineages have the GC-content decreasing in GC-rich regions. Also, we showed using multivariate analysis that substitution patterns are under different influences in both lineages. GC-biased gene conversion has a major influence on A or T to G or C substitution rates in human. In the mouse lineage, this process is active but not a major influence of substitution rates: CpG odds ratio is the major predictor of A or T to G or C substitution rates.

5. GC-content evolution and meiotic recombination hotspots

5.1. Introduction

Fine-scale characterizations of meiotic recombination in human and mouse genomes have shown that this process is unevenly distributed along genomes. Most recombination events are localized in short regions of 1 to 2 kbp, which have been called recombination hotspots. In this chapter, we investigate the link between these recombination hotspots and substitution patterns. We studied substitution patterns inside recombination hotspots in both human and mouse lineages and derived from these properties of meiotic recombination.

5.2. Double strand breaks and GC-content evolution across mouse genomes

5.2.1. Double strand breaks predict GC-content evolution in *Mus m. musculus*

The influence of meiotic recombination through gBGC on substitution patterns and GC-content evolution has been shown in both the human and the *Mus m. musculus* lineages (see previous chapter and Meunier and Duret, 2004; Duret and Arndt, 2008; Clément and Arndt, 2011). However, these analyses were done using crossover rates as a proxy measure of meiotic recombination. Double strand breaks (hereafter designated as DSBs) and recombination lead to the formation of Holliday Junctions, which can be repaired into non-crossover or crossover events (the swapping of chromosomal arms between the two chromosomes of each pair) (de Massy, 2003; Baudat and de Massy, 2007). As a result, crossover events represent only a fraction of all recombination events (Baudat and de Massy, 2007), and it is possible that the influence of gBGC on substitution patterns in mammals is stronger than previously measured. Since DSB is a better proxy measure of meiotic recombination, it allows us to test whether both crossover and non-crossover events lead are associated with gBGC. This can help us reevaluate the influence of meiotic recombination and gBGC on substitution patterns across the *Mus m. musculus* genome.

We took advantage of the recent mapping of DSB hotspots in the *Mus m. musculus* genome using chromatin immunoprecipitation followed by sequencing (hereafter

designated as ChIP-Seq) (Smagulova et al., 2011). This study identified more than 9,000 regions of approximately 3 kbp that experience high levels of DSB. We also took advantage of the recent sequencing of the genome of several mouse strains, including *Mus m. castaneus* and *Mus spretus*, two subspecies of *Mus m. musculus* (Keane et al., 2011). As these sequences were mapped on the *Mus m. musculus* genome, this allowed direct comparison of genomic sequences of different subspecies. Moreover, as recombination evolves rapidly in mouse species (Dumont et al., 2011), studying the link between recombination and substitution patterns using closely related mouse subspecies represents a clear advantage.

Substitution patterns in 1 Mbp windows across the *Mus m. musculus* genome were computed from *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* triple alignments. From these substitution patterns, GC* values, or future GC-content values, were computed which can be considered as final GC-content values providing substitution patterns stay constant over time. At the same time, for each window crossover rates from high density genetic maps available for the *Mus m. musculus* genome were extracted, as well as DSB hotspot density (proportion of a window covered by a DSB hotspot, see the Materials & Methods chapter for more details). The logarithm of both crossover rates and DSB hotspot density was used. 1580 windows were obtained, containing on average more than 960 kbp of sites where all three species have a nucleotide (e.g. none of the tree species have a gap).

First, we observe that DSB hotspot density correlates more with GC* than with current GC-content in the *Mus m. musculus* lineage (Table 5.1). As GC* values are computed from substitution patterns, we conclude that DSB hotspot density influences GC-content evolution by influencing substitution patterns in *Mus m. musculus*, something which is reminiscent of previous studies (see previous chapter as well as Meunier and Duret, 2004; Duret and Arndt, 2008; Clément and Arndt, 2011). Second, DSB density correlates more with GC-content or GC* values than sex-averaged or sex-specific crossover rates do (Table 5.1), showing that DSB density is a better proxy measure of meiotic recombination than crossover rates.

	Sex-averaged	**Male**	**Female**	**DSB Density**
	R	R	R	R
GC-content	0.244^{***}	0.276^{***}	0.088^{*}	0.298^{***}
GC*	0.310^{***}	0.300^{***}	0.152^{**}	0.404^{***}

Table 5.1.: Pearson correlation coefficients between GC-content, GC* and crossover rates or DSB density in the *Mus m. musculus* lineage. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$

It is possible that these differences in correlation coefficients reflect different levels of noise in crossover rates and DSB hotspot density and not the fact that DSB density is a better proxy measure of recombination than crossover rates. To evaluate the differences in noise level between crossover rates data and DSB data, we estimated how much noise needs to be added to DSB hotspot density to decrease the correlation coefficient to that of male crossover rates. GC*, male crossover rates and DSB

density variables were first normalized such that their mean is 0 and their standard deviation is 1. A noisy DSB density variable was then created (hereafter designated as noisyDSB) by simply adding random noise to DSB density. The random noise was generated from a normal distribution of mean 0 and of standard deviation x. This standard deviation was optimized such that the correlation of noisyDSB and GC* was on average equal to that of male crossover rates and GC*. This procedure was repeated 10,000 times. Results show that noisyDSB variance has to be 1.45 times higher than that of DSB hotspot density in order for its correlation coefficient with GC* to be equal to that of male crossover rates and GC* (Figure B.1). Overall, it is possible that differences in correlation coefficients are due to both different levels of noise between DSBs data and crossover rates, or to the fact that both NCOs and COs influence GC-content evolution through gBGC.

These results show that gBGC is likely to be associated with both crossover and non-crossover events and that the influence of gBGC in the *Mus m. musculus* lineage is much more important than previously estimated, something that we expect to be true also in the primate lineage. It should be noted that DSB hotspot density correlates better with male-specific crossover rates than with sex-averaged or female-specific crossover rates, which is expected since DSBs were mapped in male cells (Smagulova et al., 2011).

5.2.2. Different substitution patterns at different timescales in *Mus m. musculus*

The influence of gBGC on genome evolution was recently shown in the *Mus m. musculus* genome by analyzing substitution patterns (see previous chapter and Clément and Arndt, 2011). Substitution patterns are usually computed by comparing 2 sister species with an outgroup. Until recently, the closest sister species of *Mus m. musculus* with its genome sequenced was *Rattus norvegicus*, from which *Mus m. musculus* diverged about 19 million years (Myrs) ago (Veyrunes et al., 2005; Poux et al., 2006). Therefore rates computed in the *Mus m. musculus* lineages summarize events that occurred between *Mus m. musculus* and its last common ancestor with rat. As a result, rates computed in this long branch might not represent modern evolution in *Mus m. musculus*. The recent sequencing and mapping of *Mus m. castaneus* and *Mus spretus* genomes, two mouse subspecies, gives us the opportunity to measure substitution rates on a very short timescale (Keane et al., 2011). Furthermore, it allows us to compute rates in several mouse lineages and reveal possible changes in substitution patterns in different mouse species.

We compared the evolution of GC-content in the *Mus m. musculus* genome for two different time-scales by analyzing substitution rates in two different sets of triple alignments: *Mus m. musculus* - *Rattus norvegicus* - *Homo sapiens* (see previous Chapter) and *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* (see before). When we plot GC* values over current GC-content values, we observe that, across the *Mus m. musculus* genome, GC-content is globally decreasing, as GC* values are

lower than GC-content values (Figure 5.1). One striking observation is that GC* values computed in the branch between the *Mus m. musculus* - *Mus m. castaneus* ancestor (hereafter designated as *MC*, Figure 5.2 and *Mus m. musculus* are much lower than values computed in the branch between the *Mus m. musculus* - *Rattus norvegicus* ancestor (hereafter designated as *MCSR*, Figure 5.2) and *Mus m. musculus* (Figures 5.1 and 4.1, Clément and Arndt, 2011). However, two confounding effects can influence the differences in substitution patterns between different timescales.

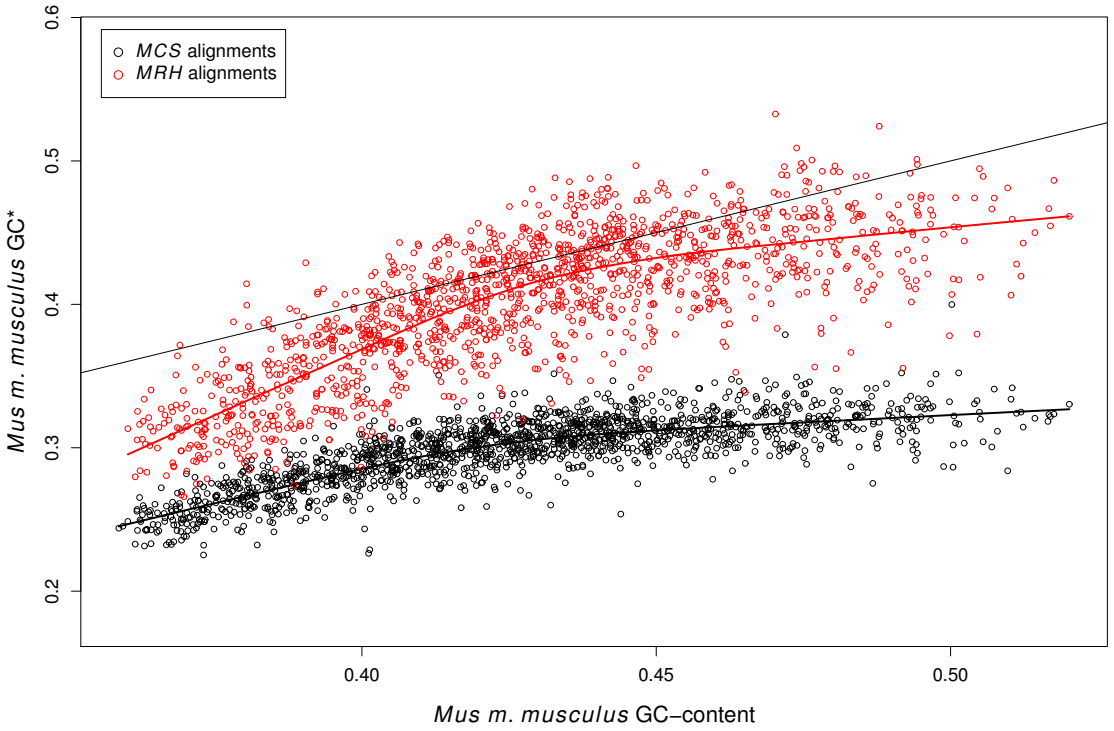


Figure 5.1.: GC* values plotted against genomic GC-content values in the *Mus m. musculus* genome for the branch leading to *Mus m. musculus* in *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* alignments (black) or *Mus m. musculus* - *Rattus norvegicus* - *Homo sapiens* (red) triple alignments. Curves represent LOWESS local regressions. The straight line represents the $x = y$ relationship.

First, as rates in the branch between *MCSR* and *Mus m. musculus* are computed using *Homo sapiens* as an outgroup, it is possible that such a distant outgroup can lead to biases in substitution patterns estimation. Second, as fewer positions in the *Mus m. musculus* genome share a nucleotide with *Rattus norvegicus* and *Homo sapiens* than with *Mus m. castaneus* and *Mus spretus*, it is possible that this first category of positions evolve differently than the second.

To resolve these issues, we built *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* - *Rattus norvegicus* - *Homo sapiens* multiple alignments for each 1 Mbp window across the *Mus m. musculus* genome as follows. For each window, we first retrieved the corresponding EPO 12 amniotes multiple alignments available at the

Ensembl database (version 62, Flicek et al., 2011) and restricted these to the analysis of *Mus m. musculus*, *Rattus norvegicus* and *Homo sapiens* (Figure 5.2). For each alignment, we then mapped the corresponding *Mus m. castaneus* and *Mus spretus* sequences onto the alignments. After filtering out windows containing less than 100 kbp of sites where all 5 species have a nucleotide as well as windows for which there was not enough information to compute crossover rates or DSB hotspot density, we obtained 1463 windows containing on average more than 300 kbp of sites where all 5 species have a nucleotide.

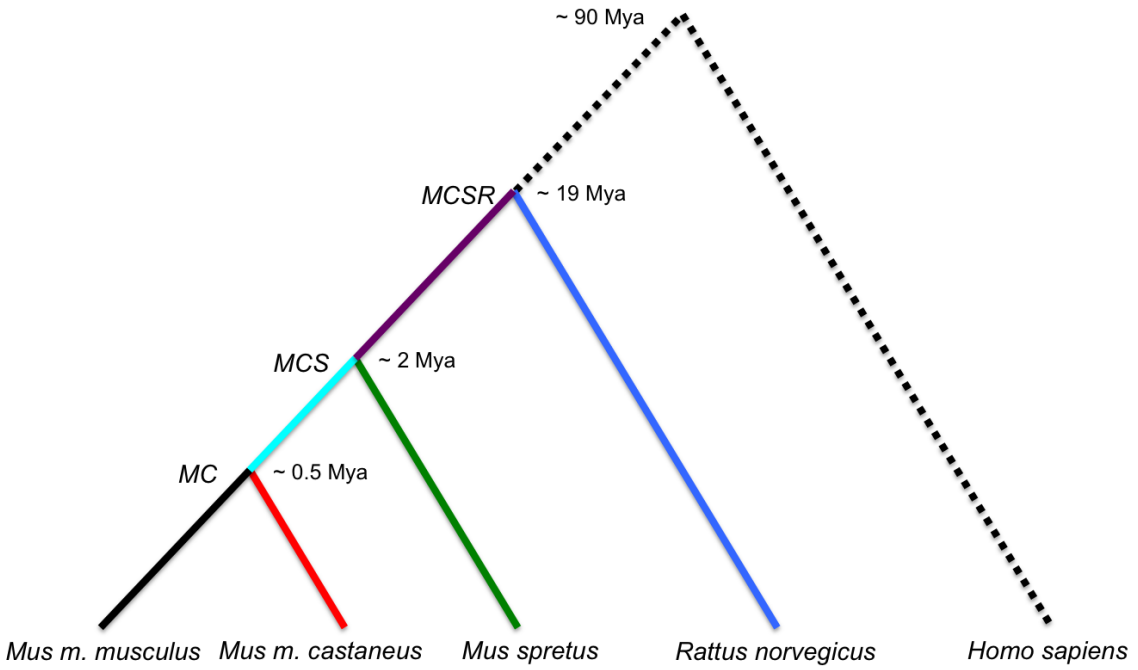


Figure 5.2.: Phylogenetic relationships between *Mus m. musculus*, *Mus m. castaneus*, *Mus spretus*, *Rattus norvegicus* and *Homo sapiens*. Divergence times are extracted from Veyrunes et al. (2005), Poux et al. (2006), Huchon et al. (2007) and Geraldes et al. (2008).

To control for the choice of outgroup, *Mus m. musculus* were compared to *Mus m. castaneus* sequences, using either *Mus spretus* or *Homo sapiens* as an outgroup, in multiple alignments built above. Results show that, although GC* values obtained using *Homo sapiens* as an outgroup are more variable than those obtained using *Mus spretus* (1.89×10^{-3} and 7.47×10^{-4} respectively), GC* values from both outgroups are within the same range (Figure B.2a). This shows that the use of a distant outgroup does not explain the observed differences in GC* values.

We further controlled whether different classes of sites evolve differently in the *Mus m. musculus* genome. We computed substitution rates by comparing *Mus m. musculus* and *Mus m. castaneus* using *Mus spretus* as an outgroup in both *Mus m. musculus* triple alignments and in 5 species alignments mentioned above.

Results show that, although GC* values obtained with multiple alignments are also more variable than GC* values obtained with triple alignments (7.47×10^{-4}

and 5.52×10^{-4} respectively, something that can be explained by the fact that using fewer sites to infer substitution rates will increase the amount of random noise), those values are in the same range (Figure B.2b). This demonstrates that sites for which *Homo sapiens* or *Rattus norvegicus* have a nucleotide do not evolve differently than other sites in the *Mus m. musculus* genome.

5.2.3. Substitution patterns changed recently in mouse lineages

We propose that different substitution patterns for long and short timescales in *Mus m. musculus* reflect recent shifts in substitution patterns in mouse lineages. In order to detect such shifts, one has to analyze substitution not only in terminal lineages but also in internal branches. To do this, we computed substitution patterns in all branches of the following 5 species tree: ((((*Mus m. musculus*, *Mus m. castaneus*), *Mus spretus*), *Rattus norvegicus*), *Homo sapiens*). This enables us to study branches between different ancestors of mouse species (Figure 5.2). Results are shown in Figure 5.3.

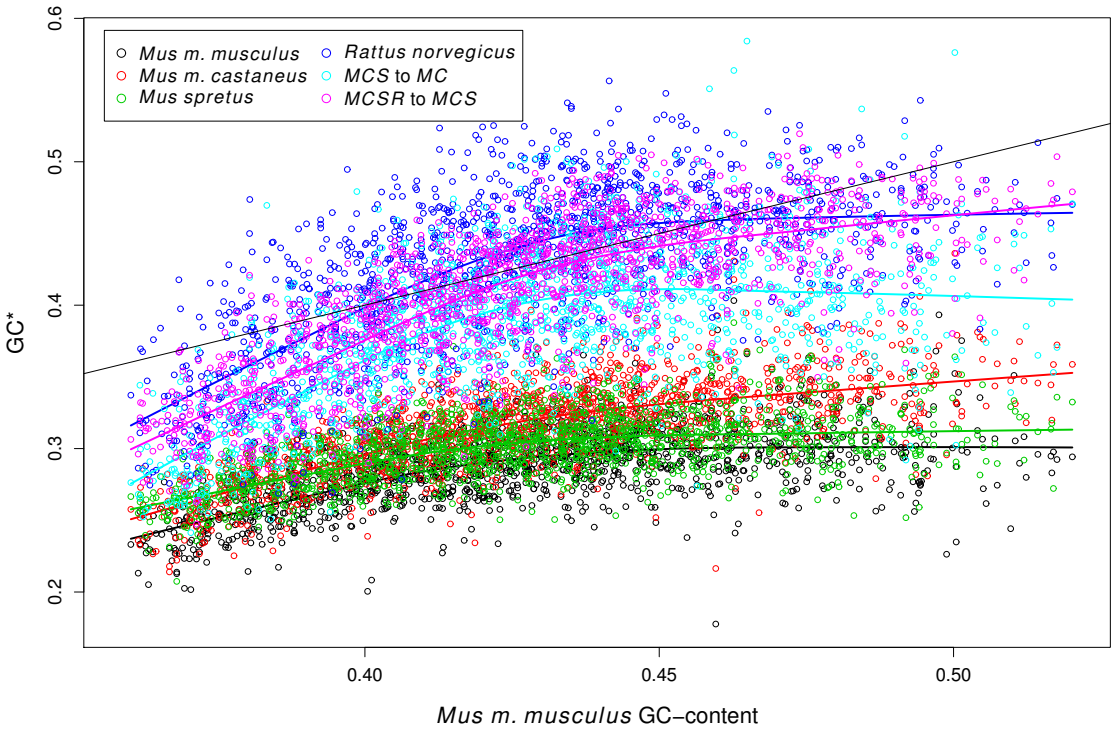


Figure 5.3.: GC* values against *Mus m. musculus* GC-content for selected branches of *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* - *Rattus norvegicus* - *Homo sapiens* multiple alignments (see Figure 5.2 for color coding). Curves represent LOWESS local regressions. The straight line represents the $x = y$ relationship.

We observe that all three mouse species (*Mus m. musculus*, *Mus m. castaneus* and

Mus spretus) have low GC* values in the same range, although *Mus m. castaneus* has slightly higher values than the other two species. Second, we see that both *Rattus norvegicus* and the branch between the *Mus m. musculus* - *Rattus norvegicus* ancestor (hereafter designated as *MCSR*) and the *Mus m. musculus* - *Mus spretus* ancestor (hereafter designated as *MCS*) have very high GC* values, comparable to values observed for *Mus m. musculus* in *Mus m. musculus* - *Rattus norvegicus* - *Homo sapiens* alignments (Figure 5.1). Finally, we observe that the branch between *MCS* and the *Mus m. musculus* - *Mus m. castaneus* ancestor (hereafter designated as *MC*) also exhibits high GC* values, though slightly lower than the previous 2 branches.

5.2.4. Possible effects of polymorphisms and incomplete lineage sorting

Several processes can affect these results. First, it is possible that the low GC* value we observe are due to segregating polymorphism: a fraction of the differences we observe between *Mus m. musculus*, *Mus m. castaneus* and *Mus spretus* correspond to polymorphic SNPs that are segregating in populations but that eventually will not get fixed. The average nucleotide diversity in mice is estimated to be between 0.0005 and 0.005 (Ideraabdullah et al., 2004; Keightley et al., 2005; Harr, 2006; Salcedo et al., 2007). Because of gBGC, S→W SNPs have a lower fixation probability than W→S ones. Thus, the estimate of GC* will tend to be lower for SNPs than for fixed positions. In our results, most changes observed in internal branches are likely to be fixed (because they are more ancient), whereas a significant fraction of changes observed in short terminal branches may be due to SNPs. This effect is therefore expected to contribute to the decrease in GC* in short terminal branches compared to internal branches. To control for this effect, we masked all SNP positions in *Mus m. musculus* (Ensembl version 65) and recomputed substitution patterns from these alignments. This procedure removes about 1% of analyzable sites (2,200 sites on average per window). Results show that, although GC* values are slightly higher in *Mus m. musculus* after masking SNPs, indicating segregating SNPs do affect GC*, this effect is very limited as GC* values are low for all three terminal branches (Figures B.4a & B.5a).

Second, it is possible that the phylogeny of sequences does not always correspond to the phylogeny of species. This process is known as incomplete lineage sorting (or ILS for short) and is expected to be stronger for short branches and in taxa with large effective population sizes. Indeed, about 12% of loci in *Mus spretus* do not place this species as an outgroup of *Mus m. musculus* or *Mus m. castaneus* (Keane et al., 2011).

To test the effect of ILS on GC* estimations, the following simulations were performed. First, random DNA sequences of length 300 kbp and of GC-content 0.25 (GC-poor), 0.50 (GC-medium) and 0.75 (GC-rich) were generated. Then, these sequences were evolve along all branches of the following tree: ((sister 1, sister 2),

outgroup), using a simple Jukes-Cantor model of parameter d for the branch to the outgroup and $d/2$ for the other branches (see Figure 4.5 and section 4.3.9 for more details). In these sequences, ILS was then performed on 14% of randomly chosen sites by replacing the nucleotide in one of the sister species by the nucleotide in the outgroup. Substitution patterns were finally in these sequences as well as in sequences not affected by ILS.

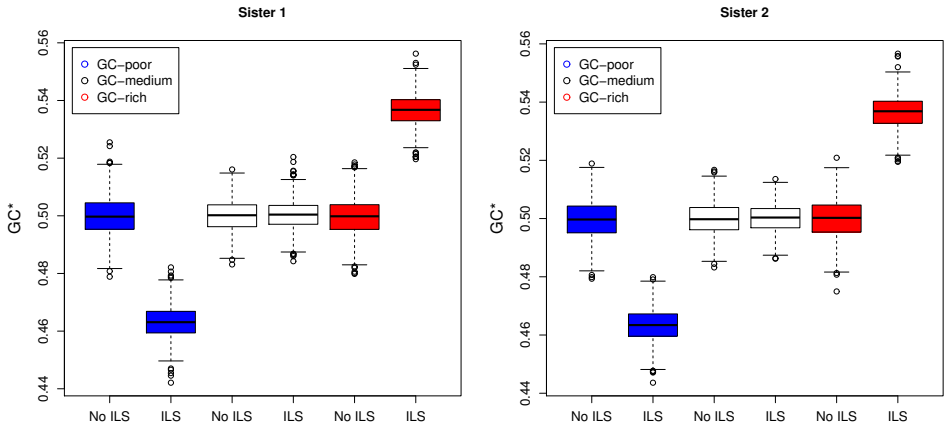


Figure 5.4.: GC* values computed in 2 sister lineages after simulating sequence evolution with a simple Jukes-Cantor model for GC-poor (blue), GC-medium (white) and GC-rich (red) sequences, with and without incomplete lineage sorting.

Results show that when GC-content is not at equilibrium, ILS does affect GC* estimations: it is underestimated when GC* is higher than GC-content and overestimated when GC* is lower than GC-content (Figure 5.4). As in our results, GC* values are lower than GC-content in the terminal branches, we argue that our computed GC* values are a conservative estimate and expect the 'real' GC* values to be lower than what we observe. In short, ILS cannot explain the low GC* values we observe in mouse subspecies.

5.2.5. Comparison of rates between branches

We interpret these results as follows. First of all, it is clear that, since different branches exhibit different substitution patterns, these changed in mouse species since the split with *Rattus norvegicus*. Because GC* values are much higher in the branch between *MCS* and *MC* than in the three mouse species, we infer that at least two independent shifts in substitution patterns are inferred to have occurred in mouse species: one before the split between *Mus m. musculus* and *Mus m. castaneus* and one in the *Mus spretus* lineage.

By comparing substitution rates in the *MCSR* to *MCS* branch to rates in other branches, we can determine what caused these shifts in GC*. To do so, substitution rates in the former were first plotted to rates in different lineage. Linear regression

for each rates were then computed. As the minimum value for substitution rates is necessarily 0, an intercept of 0 was forced for each linear regression (Figure B.3). In such analysis, as AT→TA and GC→CG substitution rates are neutral with respect to gBGC, changes in these rates are considered to reflect changes that are independent from this process. Substitution rates in both *Rattus norvegicus* and *MCS* to *MC* branches are very similar to rates in the *MCSR* to *MCS* branch, something that is expected given that GC* are also similar in these branches (Figures 5.3 and B.3a). *Mus m. castaneus* and *Mus spretus* branches exhibit similar changes. W (A or T) → S (G or C) substitution rates decreased substantially in both branches whereas S→W rates remained stable (Figures B.3d and B.3e). In the *Mus m. musculus* branch, it appears that both W→S and S→W rates changed, the former decreasing and the latter increasing (Figure B.3c). Despite these differences, all three branches have similar GC* values (Figure 5.3). Additionally, the differences in rates in *Mus m. musculus - Mus m. castaneus - Mus spretus* alignments and *Mus m. musculus - Rattus norvegicus - Homo sapiens* alignments are explained by the fact that rates in the large branch (*MCSR* to *Mus m. musculus* are dominated by rates in the *MCSR* to *MCS* branch (17 out of 19 Myrs).

5.2.6. Shifts in substitution patterns and isochores

A global increase in GC_3 (the GC-content of codons' third positions) along the mouse and rat branches, notably for GC-poor genes has been reported (Romiguier et al., 2010), which does not agree with our results. This may be due to the fact that these authors did not take CpG hypermutability into account to estimate ancestral GC. To test whether this could explain the differences between our results and theirs, substitution patterns were computed in all branches of the 5 species tree shown in Figure 5.2 without taking CpG hypermutability into account. Results show that GC* estimations do increase in the absence of CpG hypermutability (Figures B.4b & B.5b). However, this effect seems to be stronger for GC-rich regions. Differences between results of Romiguier et al. (2010) and ours therefore cannot be explained by CpG hypermutability alone.

It has been shown that GC-rich isochores are declining in mammalian genomes (Duret et al., 2002; Belle et al., 2004). We previously showed that in *Mus m. musculus*, the GC-content of both GC-rich and GC-poor regions is decreasing, and that GC-medium regions are at base composition equilibrium (see previous chapter and Clément and Arndt, 2011). In this chapter, using newly available genomic data, we however show that substitution patterns changed in mouse species, with GC* values being much lower as a result, and that this change occurred very recently. It is likely that this was caused by a decrease of gBGC's strength in mouse species. However, as it is impossible to retrieve past recombination rates, this cause cannot be precisely identified.

These genome-wide results show how DSBs and recombination influence substitution patterns and GC-content evolution across the *Mus m. musculus* genome. We

investigate in the next section how this influence is mediated in the close vicinity of DSBs.

5.3. Characterization of gene conversion in DSB hotspots in mouse

Characteristics of meiotic recombination associated gene conversion (such as length of gene conversion tracts) have been determined only for a handful of recombination loci in mouse and human (Guillon and de Massy, 2002; Jeffreys and Neumann, 2002; Jeffreys and May, 2004; Paigen et al., 2008; Webb et al., 2008; Wu et al., 2010). It is known that gene conversion tracts exhibit large variability in length, with an average length of several hundred base pairs (bp). The recent high-throughput mapping of DSB hotspots in the mouse genome (Smagulova et al., 2011) enables us to do large-scale detection and characterization of gene conversion around DSBs as well as measure the evolutionary traces of recombination and gene conversion.

We investigated which regions around DSBs are affected by gene conversion by looking for traces of gBGC, a neutral process favoring the fixation of G and C alleles: regions experiencing gene conversion will be under the influence of gBGC. This was done by first pooling all DSB hotspots using their middle points as a reference position, and then computing substitution rates in 2,000 non-overlapping windows of 100 bp using *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* triple alignments (for more details, see the Materials and Methods chapter). These pooled windows contain a total of more than 1,800 Mbp of analyzable sites (sites where all three species have a nucleotide), with an average of 900 kbp per window. From substitution patterns, GC* values (equilibrium GC-content) were computed in each window, a quantity which is correlated to the strength of gBGC.

5.3.1. Gene conversion is centered on DSB hotspots' middle points in mouse

Results show an increase of GC* relative to the background in the *Mus m. musculus* lineage, which is centered on DSB hotspots' middle points (Figure 5.5). This increase affects a region of approximately 1.5 kbp. Since DSB products are repaired through gene conversion, which will lead to a biased repair of mismatches occurring in heteroduplexes (Duret and Galtier, 2009), we infer that this increase of GC* is due to gBGC around DSB hotspot middle point. We also infer this 1.5 kbp region to represent gene conversion tracts in *Mus m. musculus*.

This increase of GC* affects only a small fraction of the 200,000 kbp that were analyzed, GC* values remain stable outside of this 1.5 kbp region (Figure 5.5). To increase resolution, we did further analyses by zooming in on this region (Figure 5.6).

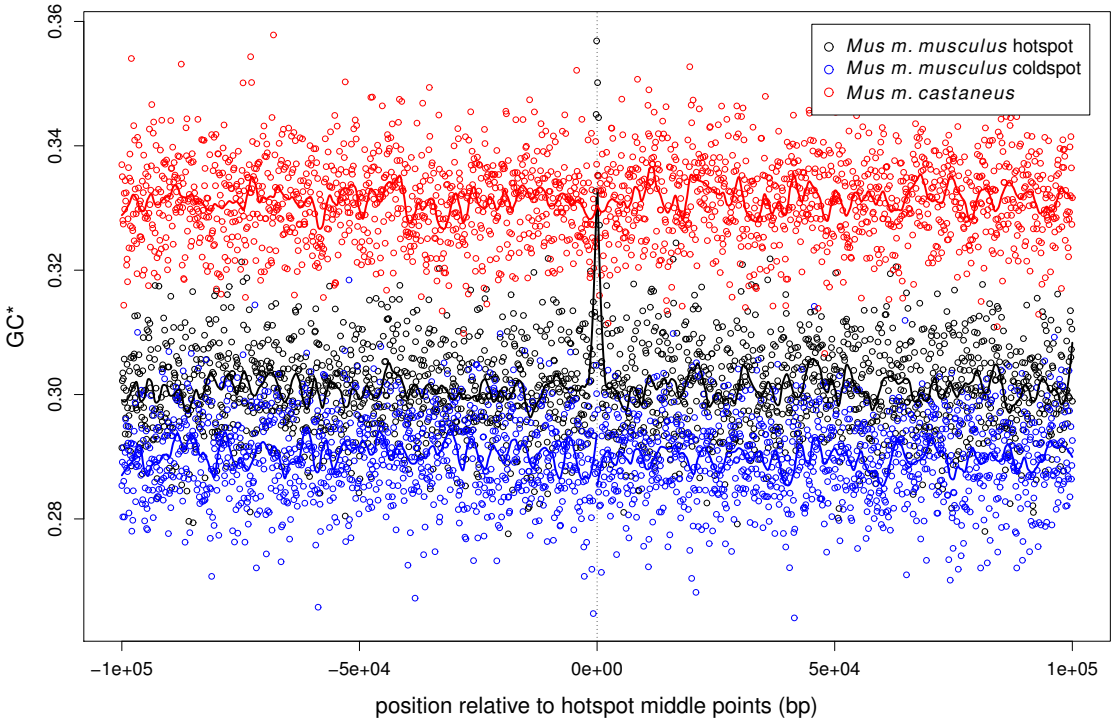


Figure 5.5.: GC* around DSB hotspots middle points in the *Mus m. musculus* lineage (black), the *Mus m. castaneus* lineage (red) and around *Mus m. musculus* DSB coldspots (blue). Lines represent one-sided local regressions computed over 25 neighboring windows.

We controlled if this increase was specific to DSB hotspots inside the *Mus m. musculus* lineage as follows. First, we computed substitution patterns and GC* values in 10,000 randomly chosen regions not overlapping DSB hotspots (hereafter designated as DSB coldspots), using the same method as for DSB hotspots. Results showed no increase of GC* in DSB coldspots (Figures 5.6 and 5.5). We then compared GC* values in DSB hotspots in the *Mus m. musculus* lineage to values in the *Mus m. castaneus* lineage. We did not observe an increase of GC* inside the *Mus m. castaneus* lineage (Figures 5.6 and 5.5), and therefore concluded that the observed increase of GC* is specific to DSB hotspots in the *Mus m. musculus* lineage. This increase is mainly due to an increase of W→S substitution rates (Figures 5.7 and B.6a). This is expected, as recombination increases, the fixation bias favoring G & C alleles is stronger than the fixation bias favoring A & T alleles (see following section).

Furthermore, this increase of GC* cannot be explained by the fact that GC-content is very high close to DSB middle points in *Mus m. musculus* (Figure B.6b): as the *Mus m. musculus* and *Mus m. castaneus* divergence time is only 500,000 years, GC-content profiles are very similar between the two species (Figure B.6b). If the increase of GC* in DSB hotspots was caused solely by an increase of GC-content, high GC* values should be seen in the *Mus m. castaneus* lineage. The absence of such increase rules out GC-content as a potential confounding variable.

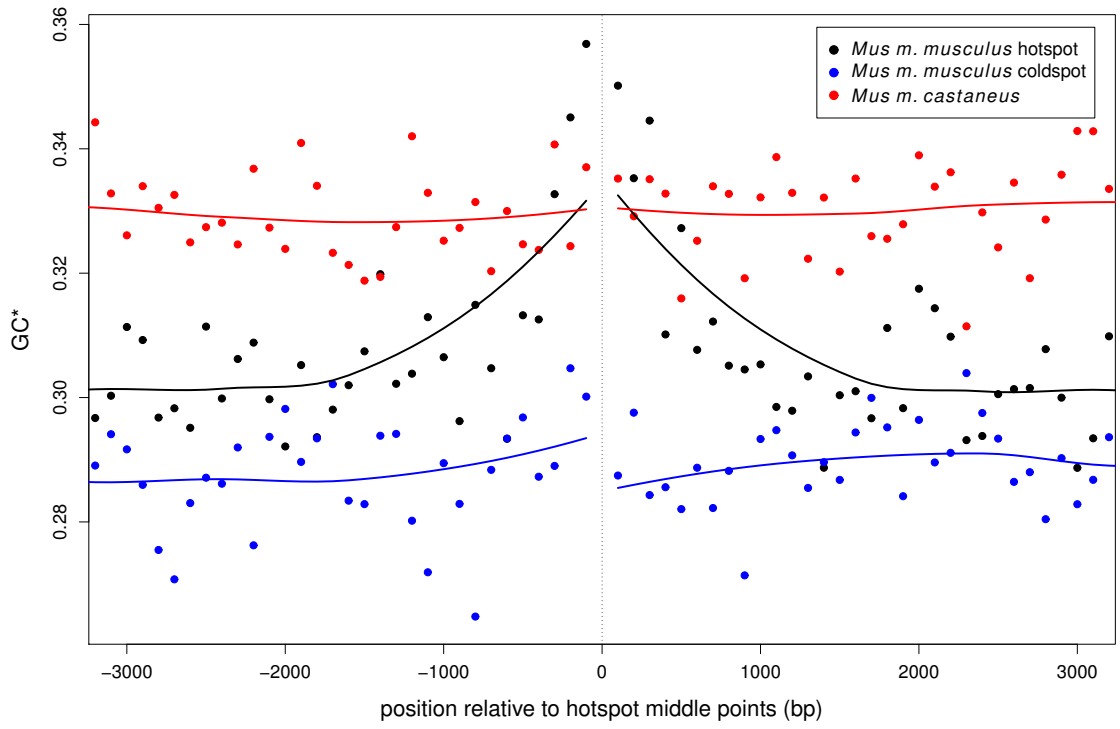


Figure 5.6.: GC* around DSB hotspots' middle points in the *Mus m. musculus* lineage (black), the *Mus m. castaneus* lineage (red) and around *Mus m. musculus* DSB coldspots (blue). Lines represent one-sided local regressions computed over 25 neighboring windows.

It is also noteworthy that GC* values for *Mus m. musculus* are in general lower than for *Mus m. castaneus*. This phenomenon cannot be explained. However, low GC* values were observed in the *Mus m. musculus* lineage compared to its sister species for two different sets of triple alignments (*Mus m. musculus* - *Rattus norvegicus* - *Homo sapiens* or *Mus m. musculus* - *Mus spretus* - *Rattus norvegicus*, data not shown). This suggests that this observation represents a genome-wide elevation of GC* in the *Mus m. musculus* lineage rather than being caused by a technical bias.

5.3.2. gBGC affects W→S substitution rates more than S→W substitution rates

Our results show that the increase of GC* around DSB hotspots' middle points in *Mus m. musculus* is mainly due to an increase of W→S substitution rates rather than a decrease of S→W substitution rates. This can be explained by the following facts. gBGC can be likened to natural selection as it is a fixation bias (Nagylaki, 1983). The fixation probability of an allele in a diploid genome can be therefore written as

$$P_{\text{fixation}} = \frac{1 - e^{-4N_e sf}}{1 - e^{-4N_e s}} \ , \tag{5.1}$$

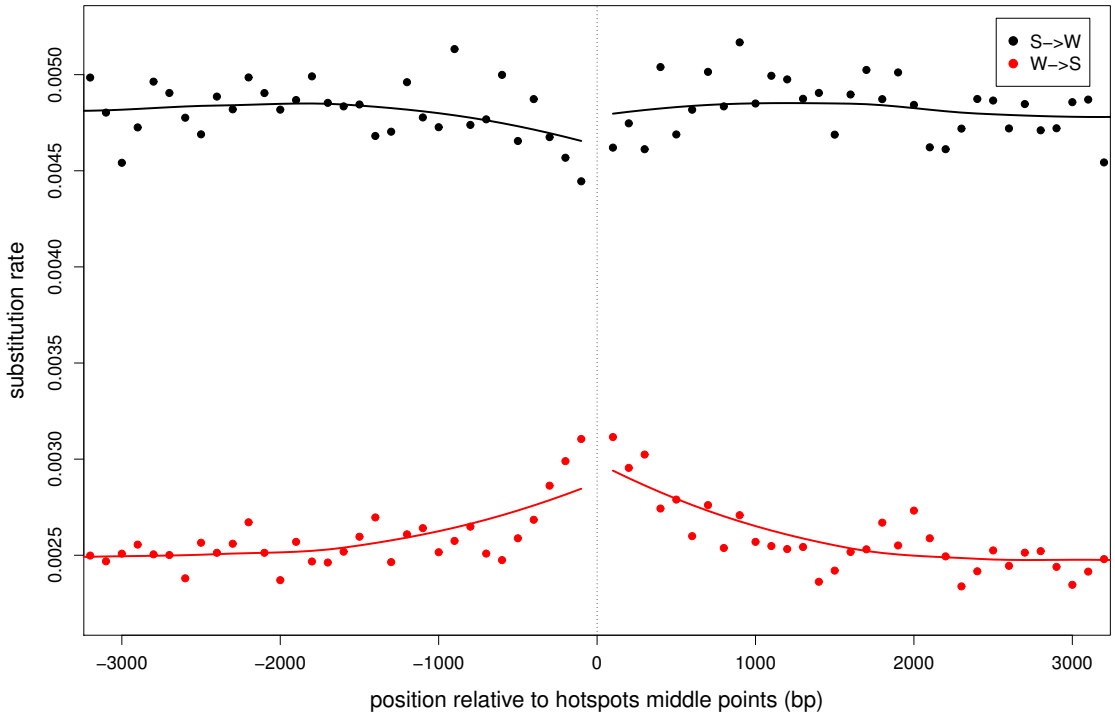


Figure 5.7.: W→S (red) and S→W (black) substitution rates around DSB hotspots' middle points in the *Mus m. musculus* lineage. Lines represent one-sided local regressions computed over 25 neighboring windows.

where s is the strength of gBGC, N_e the effective population size and f the allele frequency in the population (Kimura, 1962). s will be positive for W→S mutations and negative for S→W mutations and can be considered to be proportional to the rate of recombination (Duret and Arndt, 2008). When considering a newly occurring mutation in a diploid genome, its allele frequency f will be $1/2N_e$. Its fixation probability will then be

$$P_{\text{fixation}} = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}} \ . \tag{5.2}$$

Figure B.10 shows a plot of the fixation probabilities for S→W, W→S as well as W→W and S→S mutations. When s increases, the fixation probability of S→W mutations is much closer to the fixation probability of an S→S or W→W mutation ($1/2N_e$) than the fixation probability of W→S mutations. This means that meiotic recombination will affect more W→S substitutions than S→W substitution.

5.3.3. Characteristics of gene conversion tracts

Results show that the region affected by gene conversion in *Mus m. musculus* has an average length of approximately 1.5 kbp, centered around DSB hotspots' middle points. Furthermore, GC* values peak around DSB hotspots' middle points and decrease with distance. We interpret this as regions closer to hotspots' middle points

experiencing more gene conversion than regions more distant to hotspots' middle points and conclude that these results indicate that gene conversion tracts take variable lengths around DSBs. This agrees with previous observations for a handful of recombination hotspots in *Homo sapiens* and *Mus m. musculus* (Guillon and de Massy, 2002; Jeffreys and Neumann, 2002; Jeffreys and May, 2004; Paigen et al., 2008; Webb et al., 2008; Wu et al., 2010).

5.3.4. No evidence for recombination associated strand-specific mutations

A base composition skew has been reported around DSB hotspots' middle points in mouse (Smagulova et al., 2011) (Figure B.6c). As this skew was interpreted as being the result of strand-specific mutations caused by the recombination process, strand asymmetries in substitutions were examined around DSB hotspots' middle points.

For each pair of symmetric rates (2 pairs of transition rates, 4 pairs of transversion rates and one pair of CpG rates), the log of the ratio of the two rates was computed in each window. For example, for A→G and T→C substitution rates, the following value was computed: $\log_2(\frac{\mathrm{A}\to\mathrm{G}}{\mathrm{T}\to\mathrm{C}})$ around DSB hotspots' middle points for both the *Mus m. musculus* and *Mus m. castaneus* lineages.

Results show that strand asymmetries are very weak (Figure B.7). Furthermore, strand asymmetries observed in the *Mus m. musculus* lineage are of the same magnitude as those observed in the *Mus m. castaneus* lineage. This shows that strand-specific mutations are not causing observed base composition skews.

Furthermore, the base composition skews are reverse complement symmetric with respect to DSB hotspots' middle points: higher A frequencies on the 5' end but higher T frequencies on the 3' end of DSB hotspots are observed, as well as higher G frequencies on the 5' end but higher C frequencies on the 3' end (Figure B.6c). Base composition skews are therefore independent of the strand orientation of the hotspots (when computing these skews by randomly choosing the + or − strand for DSB hotspots, the resulting profiles are identical to analyzing only the + strand). As a result, we argue that if they cause the observed base composition skews, strand asymmetries will also be reverse complement symmetric with respect to DSB hotspots' middle points: they will be observable regardless of whether the + or − strand is studied for the hotspots. This makes our results robust to the fact that all DSB hotspots are oriented in the same manner: the chromosome's centromere is located at the 5' end while the telomere is located at the 3' end of DSB hotspots. We therefore conclude that there are no strand-specific mutations around DSBs in *Mus m. musculus*, either caused by meiotic recombination or other processes, such as transcription (Polak and Arndt, 2008) or replication (Chen et al., 2011). Overall, our results suggest a possible association between base composition skews observed around DSB hotspots in *Mus m. musculus* and recombination.

How these skews emerged is still unknown. Nonetheless, there is a bias for DSBs to occur in regions exhibiting such skews. One simple possibility is that regions

exhibiting base composition skews are preferentially recruited as DSB hotspots, either directly or indirectly through chromatin opening.

5.3.5. DSB locations are evolving rapidly

By analyzing substitution patterns around DSB hotspots in *Mus m. musculus* and comparing them with those in corresponding regions in sister species, we can study the evolution of DSBs through time. GC* values increase around DSB hotspots' middle points, which is specific to *Mus m. musculus*. Because DSBs and their subsequent repair will lead to gBGC and an increase of GC* values, the fact that no increase of GC* at the corresponding locations in *Mus m. castaneus* can be observed shows that there is no DSB and recombination occurring at the corresponding locations in *Mus m. castaneus*. DSBs locations therefore are different between these two species. Since the *Mus m. musculus* - *Mus m. castaneus* divergence time is about 500,000 years (Geraldes et al., 2008), it shows that this evolution happens quite fast.

Such an observation is reminiscent of the fact that meiotic recombination hotspots are poorly conserved in primates (Ptak et al., 2004, 2005; Winckler et al., 2005; Coop and Myers, 2007; Jeffreys and Neumann, 2009; Auton et al., 2012). Furthermore, it has been shown that meiotic recombination is controlled in *Mus m. musculus* and *Homo sapiens* by a gene called *Prdm9* (Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010). This gene is under very strong positive selection in metazoans, especially at the DNA binding residues of its zinc finger domains (Oliver et al., 2009). Models have been proposed to link the fast evolution of *Prdm9* with the fast evolution of recombination hotspots and of binding motifs in the human genome (Hochwagen and Marais, 2010; Ponting, 2011). The fast evolution of DSB hotspots locations in *Mus m. musculus* seems to indicate that recombination evolves in a similar way in primates and murids.

Results around DSB hotspots in *Mus m. musculus* show that gene conversion events happen mostly around DSB hotspots' middle points and that recombination is evolving rapidly. In the next section, we investigate substitution patterns in meiotic recombination hotspots in the human lineage.

5.4. Gene conversion and PRDM9 binding sites in human

Since no DSB mapping is currently available for the human genome, one cannot simply apply the same methodology used in *Mus m. musculus* to the human genome. High-resolution genetic maps in human have shown that meiotic recombination is not homogeneously distributed along mammalian chromosomes: it happens mostly in relatively short regions (1 to 2 kbp) that have been called recombination hotspots (Myers et al., 2005; Paigen et al., 2008; Paigen and Petkov, 2010). These hotspots have two characteristics: they live hot and die young (Coop and Myers, 2007). They

show intense recombination rates, with 80% of recombination activity happening in 20% of the genome (Myers et al., 2005). The hotspot locations evolve very rapidly during evolution in murids and primates. They are poorly conserved between human and chimpanzee (Ptak et al., 2005; Winckler et al., 2005; Coop and Myers, 2007; Jeffreys and Neumann, 2009; Auton et al., 2012).

These genetic maps consist of genetic markers scattered along the genome, with a rate of recombination measured in each interval between adjacent markers. Unlike DSB hotspots obtained in *Mus m. musculus* by ChIP-Seq for which information within hotspots is available (such as location of the hotspot peak) (Smagulova et al., 2011), detailed information other than the rate of recombination is unavailable between adjacent markers. Using human meiotic recombination hotspots and pooling them with respect to their middle points is therefore not applicable in human.

It has recently been shown that the *Prdm9* gene controls recombination activity inside recombination hotspots in murid rodents and primates (Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010). In human, this control is mediated by the binding of the PRDM9 protein to a consensus 13 bp motif (CCNCCNTNNCCNC) found in 41% of hotspots (Myers et al., 2008; Baudat et al., 2010; Myers et al., 2010). Also, variations in the zinc finger DNA binding domains of PRDM9 in human populations have been associated with variations in hotspot activity and location (Fledel-Alon et al., 2009; Berg et al., 2010; Kong et al., 2010; Berg et al., 2011; Hinch et al., 2011). This gene is under very strong positive selection in metazoans, especially at the DNA binding residues of the zinc finger domains (Oliver et al., 2009). This extra information about human recombination hotspots was therefore used to investigate the link between PRDM9 and meiotic recombination by looking for traces of gene conversion around PRDM9 binding sites in human meiotic recombination hotspots.

Substitution patterns around PRDM9 binding sites were computed by first detecting PRDM9 binding motifs in human meiotic recombination hotspots and pooling these hotspots using PRDM9 binding sites as a reference position, and then by analyzing substitution patterns in 2,000 non-overlapping windows of 100 bp, centered on binding sites, using human - chimpanzee - gorilla triple alignments. These pooled windows contain a total of around 1,000 Mbp of analyzable sites (sites where all three species share a nucleotide), with an average of over 500 kbp per window. GC* values were computed from substitution patterns to search for regions under strong influence of gBGC and to detect regions experiencing gene conversion (for more details, see the Materials & Methods chapter).

5.4.1. Gene conversion is centered on PRDM9 binding sites in human

In the human lineage, results show an increase of GC* values centered on PRDM9 binding sites (Figure 5.8 and B.8a). Regions with high values of GC* are interpreted as being under the influence of gBGC (Duret and Galtier, 2009). Similarly to the

Mus m. musculus lineage, these regions are inferred to be experiencing high rates of gene conversion.

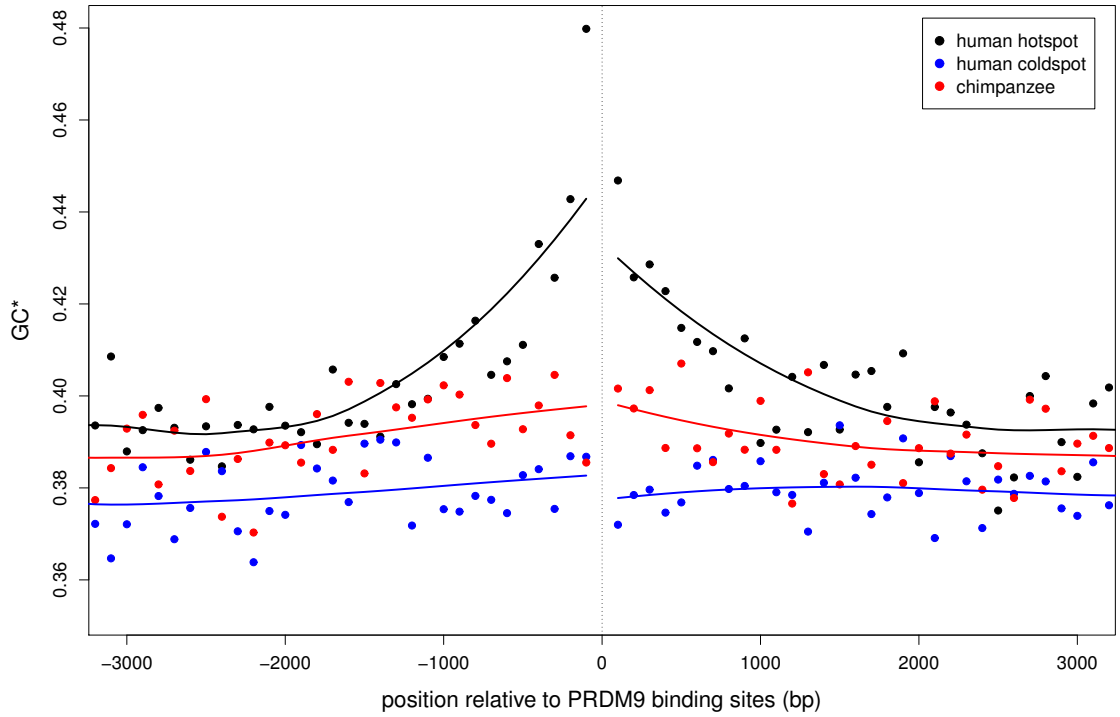


Figure 5.8.: GC* around PRDM9 binding sites in recombination hotspots in the human lineage (black), the chimpanzee lineage (red) and around PRDM9 binding sites in recombination coldspots (blue). Lines represent one-sided local regressions computed over 25 neighboring windows.

We controlled whether this increase was specific to meiotic recombination hotspots in the human lineage as follows. First, substitution patterns were computed around 10,000 randomly chosen PRDM9 binding motifs located outside recombination hotspots (hereafter designated as recombination coldspots), using the same method as for binding motifs inside hotspots. Results showed no increase of GC* in these coldspots (Figure 5.8 and B.8a). Second, substitution rates corresponding to hotspot regions in the human lineage were compared to the chimpanzee lineage. These results showed no increase of GC* in the chimpanzee lineage (Figure 5.8 and B.8a). We therefore conclude that the observed increase of GC* around PRDM9 binding sites in human meiotic recombination hotspots is specific to recombination hotspots in the human lineage. Similarly to what we observe in mouse, this increase of GC* is mainly due to an increase of W→S substitution rates around PRDM9 binding sites (Figure 5.9 and B.8b).

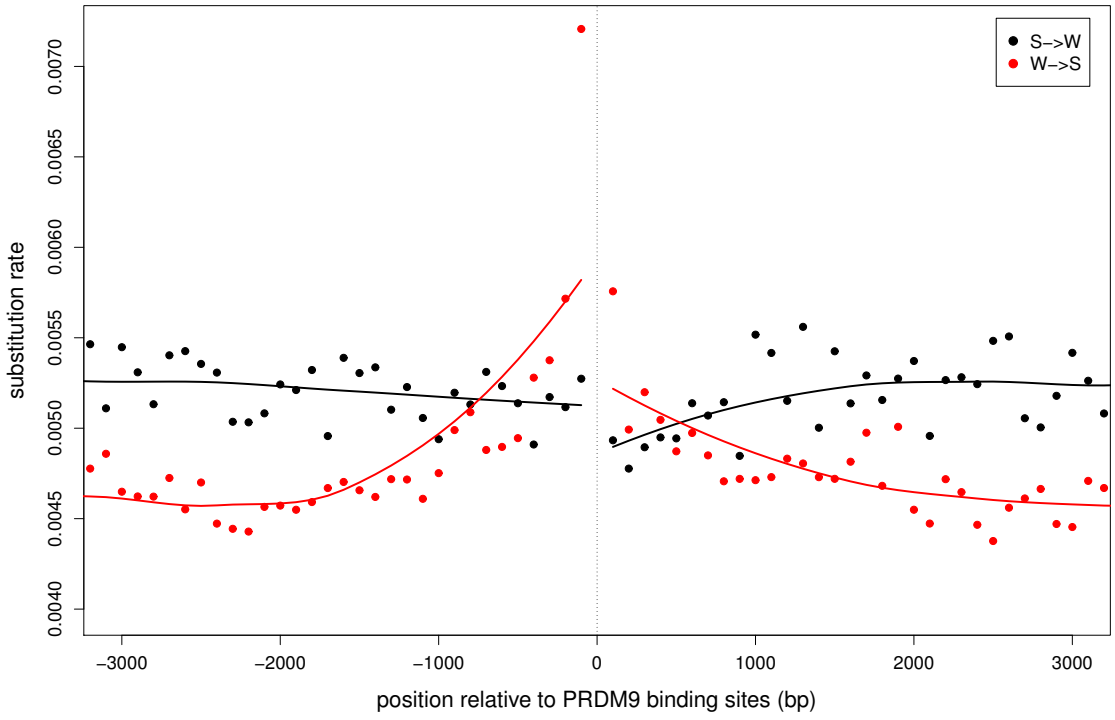


Figure 5.9.: W→S (red) and S→W (black) substitution rates around PRDM9 binding sites in recombination hotspots in the human lineage. Lines represent one-sided local regressions computed over 25 neighboring windows.

Furthermore, we controlled whether the observed increase of GC* was specific to PRDM9 sites in recombination hotspots as follows. The same set of recombination hotspots used above was analyzed, but by pooling them using their middle points as a reference position. Substitution rates in 100 bp windows around this reference position were computed using the same methodology as above. Results showed an increase of GC* around middle points in human recombination hotspots. This increase is, however, less pronounced than that observed around PRDM9 binding sites (Figure 5.10 and B.8d). We therefore conclude that gene conversion tracts are centered on PRDM9 sites.

5.4.2. PRDM9 triggers DSB directly around its binding sites

Our results show that in human recombination hotspots, gene conversion occurs mostly around PRDM9 binding sites, which agrees with recently obtained results (Katzman et al., 2011). Since in mouse, gene conversion occurs in a short region centered on DSB hotspots' middle points, it is inferred that DSBs occur in very close proximity to PRDM9 binding sites. It is therefore concluded that PRDM9 triggers DSB in the proximity of its binding site.

How this control is mediated, however, is still unknown. As PRDM9 binding sites are found all across the human genome, this protein alone cannot control recombination hotspot activity. It is possible that this control is performed through

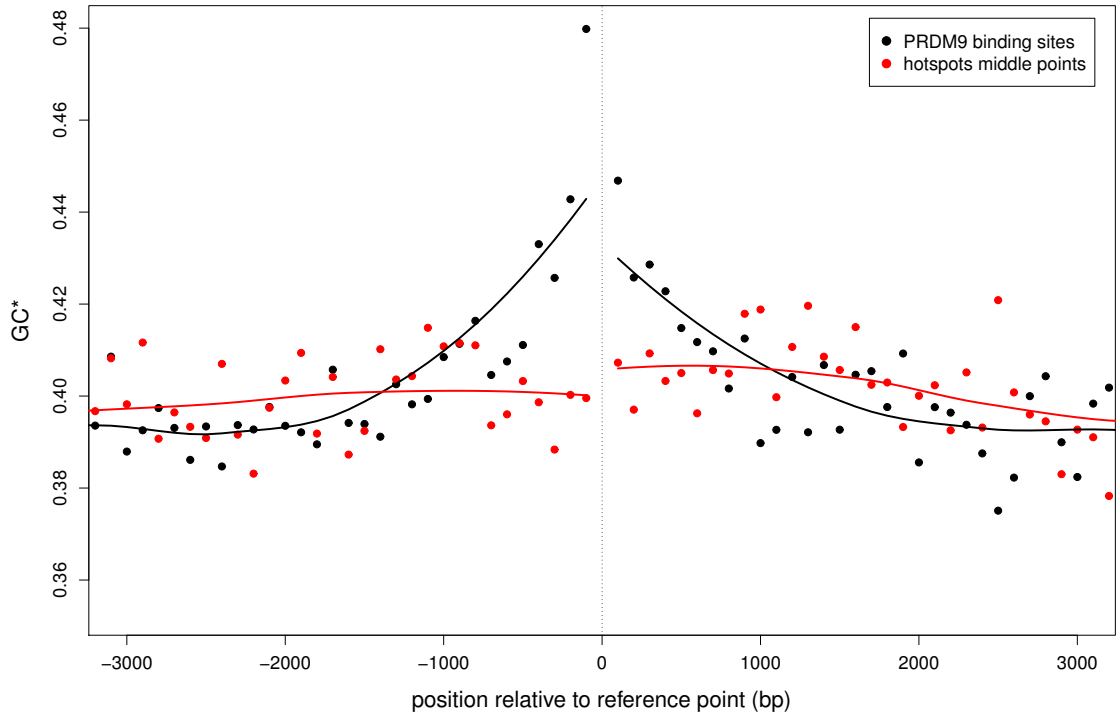


Figure 5.10.: GC* around PRDM9 binding sites (black) or hotspots' middle points (red) in human meiotic recombination hotspots. Lines represent one-sided local regressions computed over 25 neighboring windows.

protein-protein interactions. Another possibility is the recruitment of other proteins through histone modifications. PRDM9 has a histone methyl-transferase domain, and a histone modification (trimethylation of the lysine residue of the histone 3 protein, H3K4me3) has been found to be associated in mouse with recombination hotspots (Buard et al., 2009; Borde et al., 2009; Smagulova et al., 2011) and with PRDM9 binding activity (Grey et al., 2011).

Alternatively, this control could be mediated by the binding of proteins to the DNA in the vicinity of PRMD9 binding sites. To test this hypothesis, we explored over-represented DNA motifs in a 2 kbp region around PRDM9 binding sites using the RSAT peak-motifs tool (Thomas-Chollier et al., 2012). We found one DNA motif that showed a particular over-enrichment close to PRDM9 binding sites in meiotic recombination hotspots (Figure B.9a). However, by repeating this analysis around PRDM9 binding motifs in recombination coldspots, we found exactly the same over-represented DNA binding motifs (Figure B.9b). This indicates that such over-represented motifs are not specific to meiotic recombination hotspots and thus cannot control hotspot activity.

5.4.3. The hotspot conversion paradox

PRDM9 binding sites are directly affected by gene conversion through DSB and subsequent digestion and repair, which means that when binding and subsequent recombination occurs at a locus where there is heterozygosity for the binding motif, this motif will be replaced by gene conversion with an inactive motif, thus disabling PRDM9 binding and causing the recombination activity to disappear (although gBGC can slow down this process for reasons explained below). Indeed, it has been recently shown that there is a drive against the consensus binding motif specific to human inside meiotic recombination hotspots (Myers et al., 2010). This drive will cause a selective pressure favoring new binding specificities of the zinc finger domains of the PRDM9 protein (Hochwagen and Marais, 2010; Ponting, 2011). This has been put forward to explain the very fast evolution of hotspot structure between human and chimpanzee (Ptak et al., 2005; Winckler et al., 2005; Coop and Myers, 2007; Jeffreys and Neumann, 2009), the very fast evolution of the *Prdm9* gene in metazoans (Oliver et al., 2009) and the variation in hotspot structure observed between human populations (Fledel-Alon et al., 2009; Berg et al., 2010; Kong et al., 2010; Berg et al., 2011; Hinch et al., 2011). This entire process has been modeled in the hotspot conversion paradox model, the fact that an allele promoting recombination activity will promote its own disruption (Boulton et al., 1997; Jeffreys and Neumann, 2002; Pineda-Krch and Redfield, 2005). Overall, our results agree with that model.

It is possible that gBGC helps maintain recombination activity as it promotes the fixation of G and C alleles: the consensus PRDM9 binding motif in human is indeed C-rich, and the fact that a mismatch involving a C will be repaired often back to a C by gBGC is expected. This effect remains relatively small for the following reasons. First, the consensus motif is C-rich, not GC-rich, therefore a C→G substitution in the motif will result in its inactivation. Second, as the biased mismatch repair is biased towards both G and C bases, it is then possible that mismatches will be repaired into G, which will also inactivate the motif. Finally, gBGC is not expected to be 100% efficient: not all mismatches will be repaired into G and C. Inversely, if the PRDM9 binding motif was AT-rich, gBGC might actually speed up the turnover of recombination hotspots for reasons mentioned above.

5.4.4. Differences in hotspot structures between human and chimpanzee

Recently, meiotic recombination hotspots have been mapped in the chimpanzee genome using a sequencing-based technique (Auton et al., 2012), providing a unique opportunity to compare the recombination process between two closely related species at a small scale. This study shows that although meiotic recombination works in hotspots in both genomes, the hotspot structure is much more complex in the chimpanzee genome: hotspots in chimpanzee are not as well defined as in human. This is

explained by the fact that *Prdm9* alleles are much more variable in chimpanzee than in human and that contrary to human, no consensus binding motif can be found in chimpanzee. Such differences are surprising and cannot be explained by positive selection being more efficient in chimpanzee compared to human, given that both species' effective population sizes are of similar orders of magnitude (Yu et al., 2003; Keightley et al., 2005). It is possible that the increased diversity of *Prdm9* alleles in chimpanzee is due to the fact that these alleles are very recent and thus harder to identify. It is also possible that the PRM9 protein is less predominant in hotspot determinism in chimpanzee compared to human.

5.4.5. Contrasting results in human and mouse lineages

Our results in the human lineage show that the region affected by gene conversion has a length of approximately 2 kbp centered on PRDM9 binding sites in human meiotic recombination hotspots. This is a little longer than what we observe in mouse DSB hotspots. Differences in gene conversion tract length have already been raised as a potential explanation to differences in large-scale dynamics of GC-content evolution between primates and murid rodents (Clément and Arndt, 2011). These different results can be explained as follows. It is possible that DSBs occur directly at PRDM9 binding sites and that gene conversion tracts are indeed longer in human compared to mouse. However, it is also possible that DSBs do not occur directly at PRDM9 binding sites, and that gene conversion tract have the same length in both species. Recent results in the mouse genome indicate that PRDM9 binding sites are normally distributed around DSB hotspots' middle points (Smagulova et al., 2011), which supports the second hypothesis. However, without large-scale information about DSB in human, whether human and mouse gene conversion tract lengths differ remains an open question.

5.4.6. Comparison of different hotspots datasets

To study biased gene conversion inside meiotic recombination hotspots in the human genome, a hotspot dataset was generated from a high density genetic map available (hereafter designated as Clement hotspots, International HapMap Consortium et al., 2007). The same analysis done previously was performed using already published hotspots (hereafter designated as Myers hotspots, Myers et al., 2005).

Following the same methodology as Necşulea et al. (2011), hotspots locations were first downloaded, their coordinates were then converted to the *hg19* version of the human genome. More than 34,000 recombination hotspots were obtained. First, crossover rates were computed inside each recombination hotspot by computing the weighted average of crossover rates of chromosomal regions that overlap the hotspots. Only about half of the Myers hotspots have a crossover rate of at least 10 cM/Mb, our criteria to define recombination hotspots in genetic maps. Then, Myers hotspots were divided into two equally sized groups, a highly recombining hotspots

group and a lowly recombining hotspots group and applied the same procedure as for Clement hotspots to analyze substitution patterns around PRDM9 binding sites in Myers hotspots (see the Materials & Methods chapter for more details). This procedure was done in both groups separately as well as for all Myers hotspots grouped together.

When analyzing substitution patterns around PRDM9 binding sites, we observe that on average Myers hotspots have lower GC* values compared to Clement hotspots (Figure 5.11 and B.8e). The analysis of lowly and highly recombining Myers hotspots reveal interesting features. Whereas highly recombining hotspots have GC* values comparable to that of Clement hotspots, lowly recombining hotspots have GC* values comparable to what is observed in the chimpanzee lineage (Figure 5.8 and B.8a). As what differentiates both groups of hotspots is recombination rates, we interpret differences in substitution patterns as gBGC having less influence in lowly recombining hotspots compared to highly recombining hotspots. We therefore argue that the absolute crossover rate primarily determines how strong gBGC will be in a particular region.

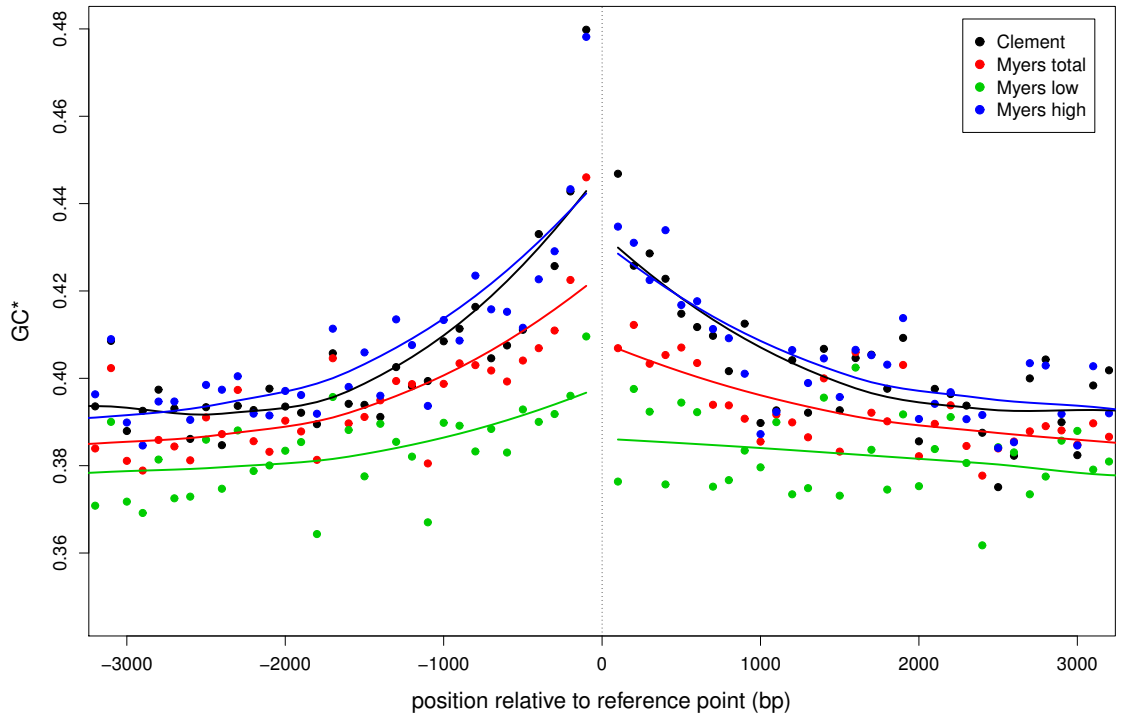


Figure 5.11.: GC* around PRDM9 binding sites in Clement hotspots (black), all Myers hotspots (red), Myers hotspots with low crossover rates (green) and Myers hotspots with high crossover rates (blue). Lines represent one-sided local regressions computed over 25 neighboring windows.

A number of studies focusing on possible effects of meiotic recombination and gBGC on genome evolution, particularly in and around recombination hotspots, use the distance of a particular region to a hotspot as a proxy measure of recombination

(Berglund et al., 2009; Necşulea et al., 2011). As the absolute crossover rate is shown to best predicts the influence of gBGC in a particular genomic region, we argue that using the distance to the nearest hotspots will be less informative than the crossover rate in the region of interest.

5.5. Conclusion

By investigating the link between double strand breaks and substitution patterns across the *Mus m. musculus* genome, we were able to show that DSB better predicts GC-content evolution than crossover rates, possibly due to the influence of non-crossover events on substitution patterns. Also, we found independent changes in substitution patterns in different mouse lineages. Furthermore, by analyzing substitution patterns in the vicinity of double strand breaks in *Mus m. musculus*, we were able to show that gene conversion events occur mostly around double strand breaks hotspots' middle points. Finally, we show that gene conversion is triggered in the proximity of PRDM9 binding sites in human. Overall, these results give a precise picture of how recombination works in both *Mus m. musculus* and human genome.

A. Appendix A

	Sex-averaged	**LDT**
	ρ	ρ
GC-content	0.420***	-0.378***
GC*	0.681***	-0.532***

Table A.1.: Spearman rho correlation coefficients between substitution rates, crossover rates and LDT in human. ***p-value $< 10^{-10}$

	Sex-averaged	**Male-specific**	**Female-specific**	**LDT**
	ρ	ρ	ρ	ρ
GC-content	0.182***	0.269***	0.096*	-0.273***
GC*	0.190***	0.282***	0.097*	-0.334***

Table A.2.: Spearman rho correlation coefficients between substitution rates, crossover rates and LDT in mouse. *p-value < 0.05; ***p-value $< 10^{-10}$

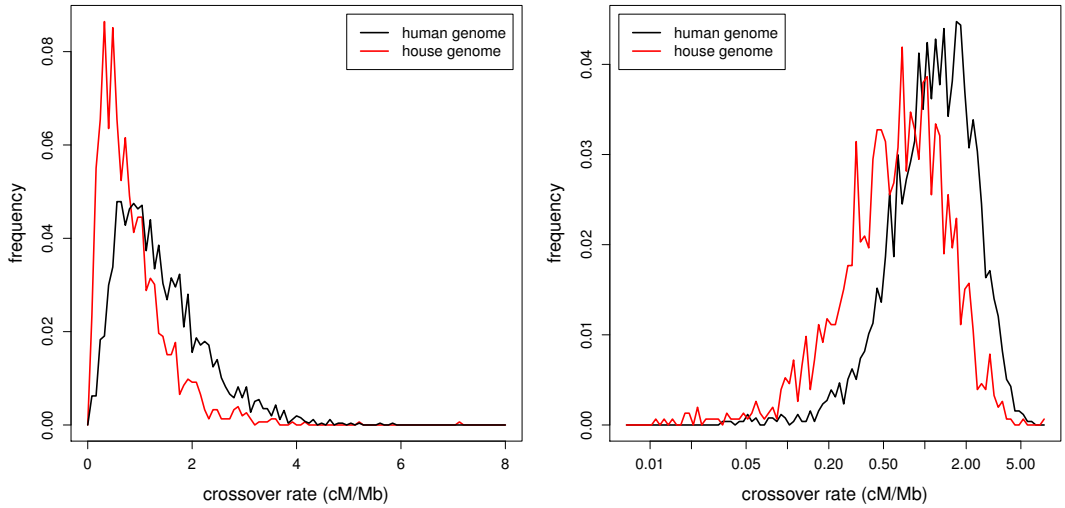


Figure A.1.: Distribution of CO on a normal scale (left panel) and on a logarithmic (right panel) for the human (black) and mouse (red) genomes.

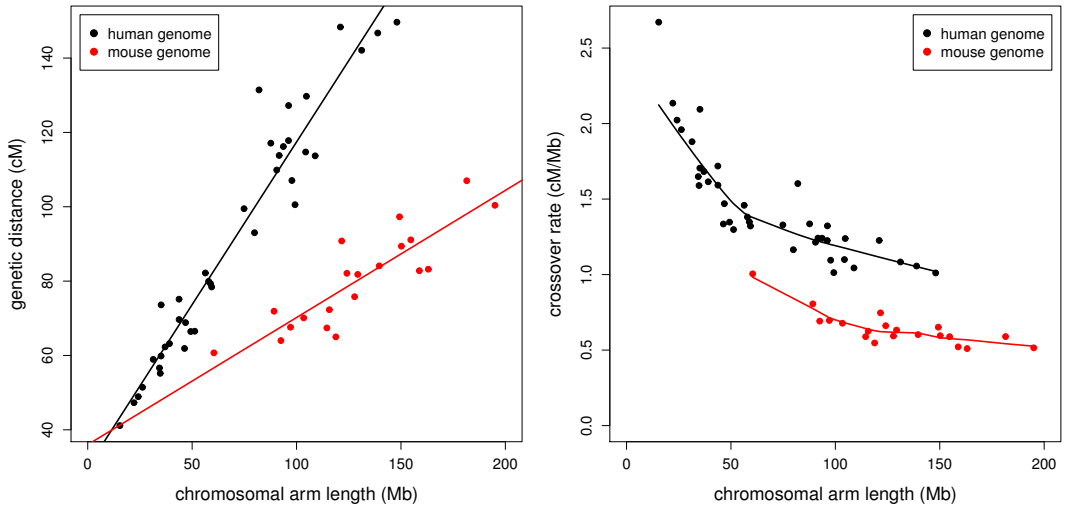


Figure A.2.: Relationship between chromosomal arm length and genetic distance (left panel) and crossover rates (right panel) for the human (black) and mouse (red) genomes.

	GC	LCO	LDT	RepTime	Exons	SINEs	LINEs	LTRs	CpGodds
	R	R	R	R	R	R	R	R	R
GC*	0.508***	0.631***	-0.604***	-0.226***	0.294***	0.338***	-0.427***	-0.309***	0.516***
W→S	-0.028	0.526***	-0.530***	0.351***	-0.321***	-0.272***	0.006	0.195***	0.070*
S→W	-0.123**	0.190***	-0.300***	0.445***	-0.447***	-0.376***	0.103**	0.354***	-0.012
W→W	-0.412***	0.119**	-0.120**	0.493***	-0.573***	-0.495***	0.288***	0.388***	-0.274***
S→S	-0.126**	0.283***	-0.353***	0.329***	-0.346***	-0.271***	0.070*	0.196***	0.010
CpG Rate	-0.678***	-0.140***	0.198***	0.634***	-0.696***	-0.699***	0.587***	0.583***	-0.668***
Total Rate	0.070*	0.443***	-0.521***	0.344***	-0.316***	-0.236***	-0.058*	0.206***	0.160***

Table A.3.: Pearson correlation coefficients between genomic features and substitution rates in human. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$

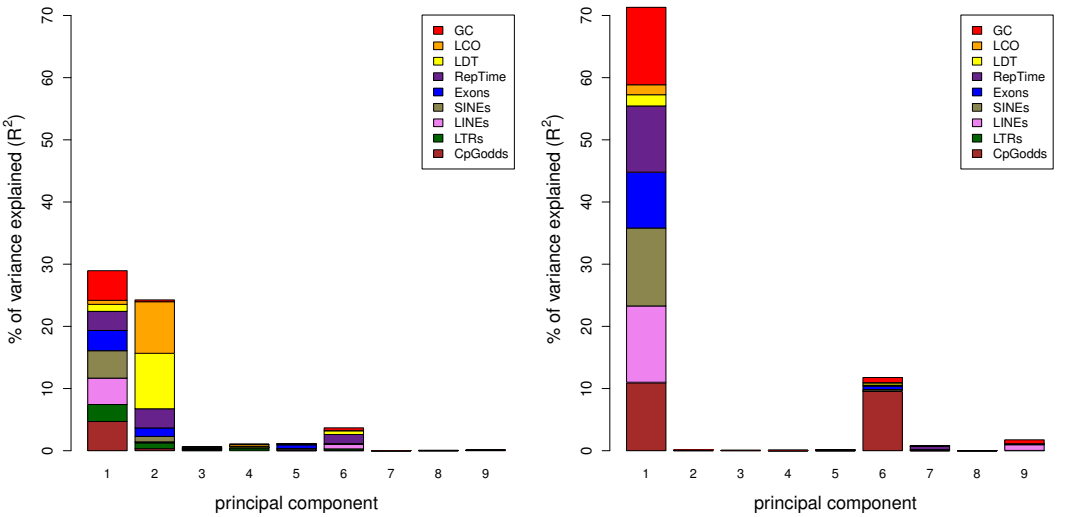


Figure A.5.: Principal component regression for GC* in the human (left panel) and the mouse (right panel) lineages. The height of each bar represents how much of the variable's variance is explained by the corresponding component. Each colored area is proportional to the relative importance of the corresponding feature inside a component.

	GC *R*	LCO *R*	LDT *R*	RepTime *R*	Exons *R*	SINEs *R*	LINEs *R*	LTRs *R*	CpGodds *R*
GC*	0.757***	0.275***	-0.305***	-0.637***	0.625***	0.734***	-0.733***	-0.053*	0.909***
W→S	0.219***	0.196***	-0.234***	-0.174***	0.132**	0.174***	-0.268***	0.048	0.603***
S→W	-0.680***	-0.180***	0.181***	0.630***	-0.628***	-0.719***	0.636***	0.097*	-0.614***
W→W	-0.699***	-0.175***	0.135**	0.510***	-0.608***	-0.690***	0.594***	0.057*	-0.516***
S→S	-0.685***	-0.140**	0.134**	0.538***	-0.623***	-0.685***	0.571***	0.067*	-0.421***
CpG Rate	-0.594***	-0.172***	0.190***	0.568***	-0.583***	-0.609***	0.555***	0.050*	-0.582***
Total Rate	-0.433***	-0.040	0.016	0.438***	-0.469***	-0.527***	0.382***	0.124**	-0.201***

Table A.4.: Pearson correlation coefficients between genomic features and substitution rates in mouse. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$

	GC *R*	LCO *R*	LDT *R*	RepTime *R*	Exons *R*	SINEs *R*	LINEs *R*	LTRs *R*	CpGodds *R*
GC	*NA*	0.354***	-0.453***	-0.543***	0.631***	0.718***	-0.744***	-0.510***	0.817***
LCO	0.354***	*NA*	-0.474***	-0.006	0.103**	0.165***	-0.318***	-0.046*	0.265***
LDT	-0.453***	-0.474***	*NA*	0.012	-0.114**	-0.177***	0.373***	0.140***	-0.536***
RepTime	-0.543***	-0.006	0.012	*NA*	-0.605***	-0.718***	0.515***	0.463***	-0.555***
Exons	0.631***	0.103**	-0.114**	-0.605***	*NA*	0.662***	-0.477***	-0.469***	0.569***
SINEs	0.718***	0.165***	-0.177***	-0.718***	0.662***	*NA*	-0.681***	-0.559***	0.698***
LINEs	-0.744***	-0.318***	0.373***	0.515***	-0.477***	-0.681***	*NA*	0.565***	-0.747***
LTRs	-0.510***	-0.046*	0.140***	0.463***	-0.469***	-0.559***	0.565***	*NA*	-0.527***
CpGodds	0.817***	0.265***	-0.536***	-0.555***	0.569***	0.698***	-0.747***	-0.527***	*NA*

Table A.5.: Pearson correlation coefficients between genomic features in human. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$

	GC *R*	LCO *R*	LDT *R*	RepTime *R*	Exons *R*	SINEs *R*	LINEs *R*	LTRs *R*	CpGodds *R*
GC	*NA*	0.240***	-0.304***	-0.721***	0.657***	0.798***	-0.879***	-0.017	0.706***
LogDis	0.240***	*NA*	-0.285***	-0.158**	0.146**	0.209***	-0.259***	-0.044	0.277***
LDT	-0.304***	-0.285***	*NA*	0.131**	-0.071*	-0.197***	0.330***	0.114**	-0.331***
RepTime	-0.721***	-0.158**	0.131**	*NA*	-0.659***	-0.804***	0.679***	-0.052*	-0.658***
Exons	0.657***	0.146**	-0.071*	-0.659***	*NA*	0.711***	-0.572***	0.021	0.610***
SINEs	0.798***	0.209***	-0.197***	-0.804***	0.711***	*NA*	-0.808***	-0.042	0.720***
LINEs	-0.879***	-0.259***	0.330***	0.679***	-0.572***	-0.808***	*NA*	0.203***	-0.743***
LTRs	-0.017	-0.044	0.114**	-0.052*	0.021	-0.042	0.203***	*NA*	-0.074*
CpGodds	0.706***	0.277***	-0.331***	-0.658***	0.610***	0.720***	-0.743***	-0.074*	*NA*

Table A.6.: Pearson correlation coefficients between genomic features in mouse. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$

	GC	LCO	LDT	RepTime	Exons	SINEs	LINEs	LTRs	CpGodds	R^2
GC*	0.014	0.463	-0.321	-0.104	0.055	-0.031	0.102	-0.168	0.130	0.597
W→S	-0.188	0.424	-0.325	0.106	-0.219	-0.226	0.025	0.039	0.318	0.598
S→W	-0.213	0.067	-0.032	0.199	-0.330	-0.246	-0.073	0.217	0.260	0.581
W→W	-0.351	0.166	-0.092	0.133	-0.350	-0.153	-0.050	0.118	0.325	0.458
S→S	-0.267	0.224	-0.223	0.126	-0.256	-0.134	0.016	0.048	0.395	0.354
CpG Rate	-0.134	-0.012	-0.038	0.142	-0.279	-0.106	-0.018	0.177	-0.184	0.658
Total Rate	0.045	0.280	-0.217	0.198	-0.312	-0.258	-0.014	0.144	0.465	0.570

Table A.7.: Slopes of linear regressions between substitution rates and genomic features in the human lineage.

	GC	LCO	LDT	RepTime	Exons	SINEs	LINEs	LTRs	CpGodds	R^2
GC*	0.328	0.017	0.012	0.101	0.012	0.130	0.176	-0.017	0.771	0.860
W→S	-0.122	0.035	-0.003	0.106	-0.169	-0.305	-0.006	0.125	1.076	0.541
S→W	-0.515	-0.006	-0.018	-0.161	-0.131	-0.388	-0.282	0.091	-0.367	0.791
W→W	-0.672	-0.022	-0.051	-0.271	-0.166	-0.539	-0.343	0.088	0.010	0.596
S→S	-0.645	-0.004	0.008	-0.129	-0.246	-0.530	-0.307	0.113	0.264	0.608
CpG Rate	-0.199	-0.006	0.033	0.101	-0.222	-0.136	-0.092	0.052	-0.193	0.456
Total Rate	-0.236	0.020	-0.004	0.083	-0.241	-0.553	-0.158	0.174	0.454	0.409

Table A.8.: Slopes of linear regressions between substitution rates and genomic features in the mouse lineage.

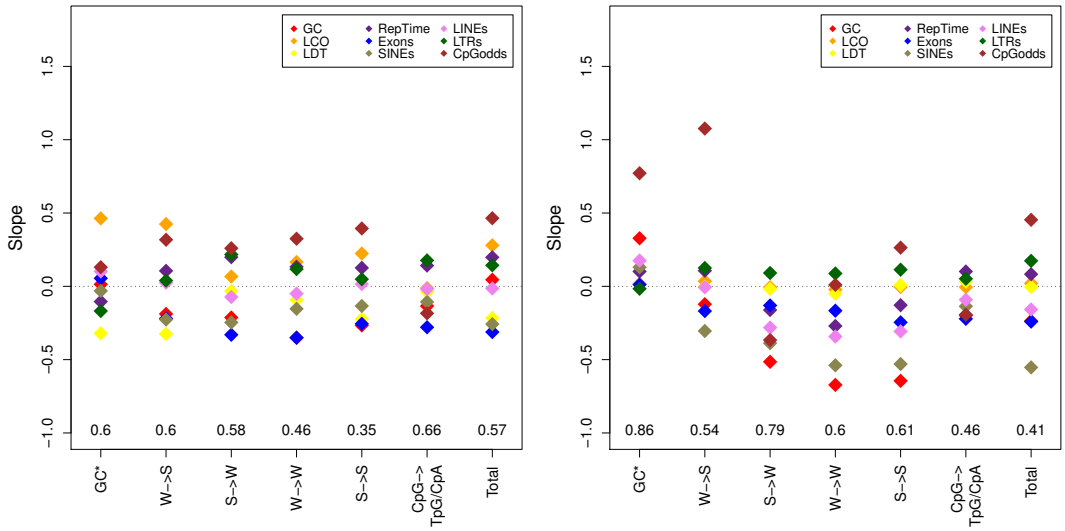


Figure A.3.: Slopes of linear regressions between substitution patterns and genomic features in the human (left panel) and mouse (right panel) lineages. Total R^2 values for each linear model is indicated below.

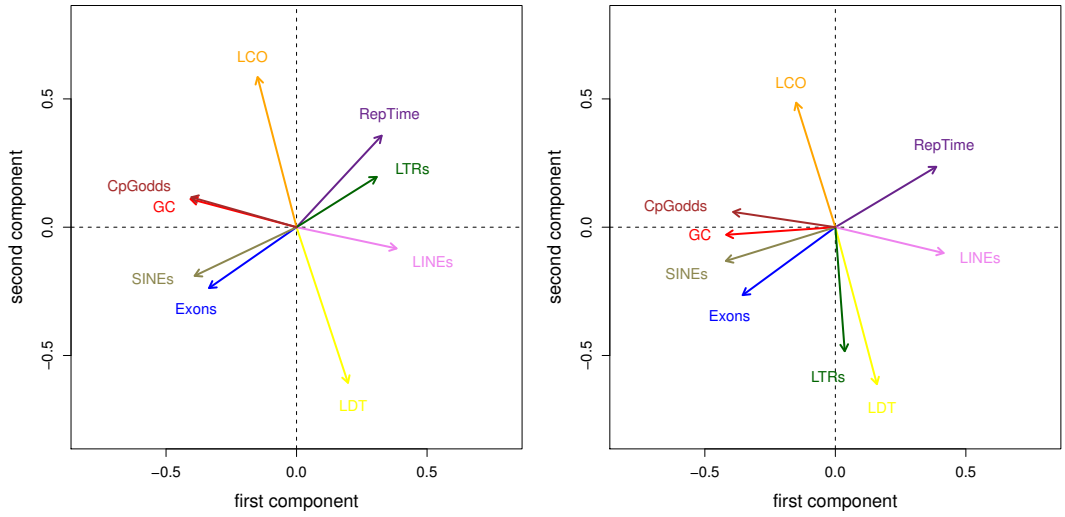


Figure A.4.: Entries of the eigenvectors of the first two principal components in the human (left) and mouse (right) lineages. Two independent projections were performed in the human and mouse genomes. Entries of each eigenvector were normalized such as $\sum \text{entries}^2 = 1$.

	PC1 R^2	PC2 R^2	PC3 R^2	PC4 R^2	PC5 R^2	PC6 R^2	PC7 R^2	PC8 R^2	PC9 R^2	Total R^2
GC*	0.128^{***}	0.261^{***}	0.034^{***}	0.003^{*}	0.001	0.001	0.007^{*}	0.031^{***}	0.005^{*}	0.471^{***}
W→S	0.031^{***}	0.408^{***}	0.065^{***}	0.001	0.007^{*}	0.002^{*}	0.014^{**}	0.013^{***}	0.017^{**}	0.558^{***}
S→W	0.617^{***}	0.010^{**}	0.009^{**}	0.001	0.015^{**}	*NA*	0.003^{*}	0.004^{*}	0.009^{**}	0.669^{***}
W→W	0.344^{***}	0.077^{***}	0.021^{***}	0.003^{*}	0.023^{***}	0.001	0.006^{*}	0.001	0.022^{***}	0.497^{***}
S→S	0.081	0.231^{***}	0.034^{***}	*NA*	0.008^{*}	0.004^{*}	0.001	*NA*	0.037^{***}	0.396^{***}
CpG Rate	0.064^{***}	0.008^{*}	0.009^{**}	0.002^{*}	0.008^{*}	0.003^{*}	0.001	0.019^{***}	0.008^{*}	0.123^{***}
Total Rate	0.009^{**}	0.454^{***}	0.043^{***}	0.011^{**}	0.002^{*}	0.016^{**}	*NA*	0.052^{***}	0.001	0.589^{***}

Table A.9.: Results of principal component regression on substitution patterns in the *HCM* branch. *p-value < 0.05; **p-value $< 10^{-5}$; ***p-value $< 10^{-10}$
NA: $R^2 < 0.001$

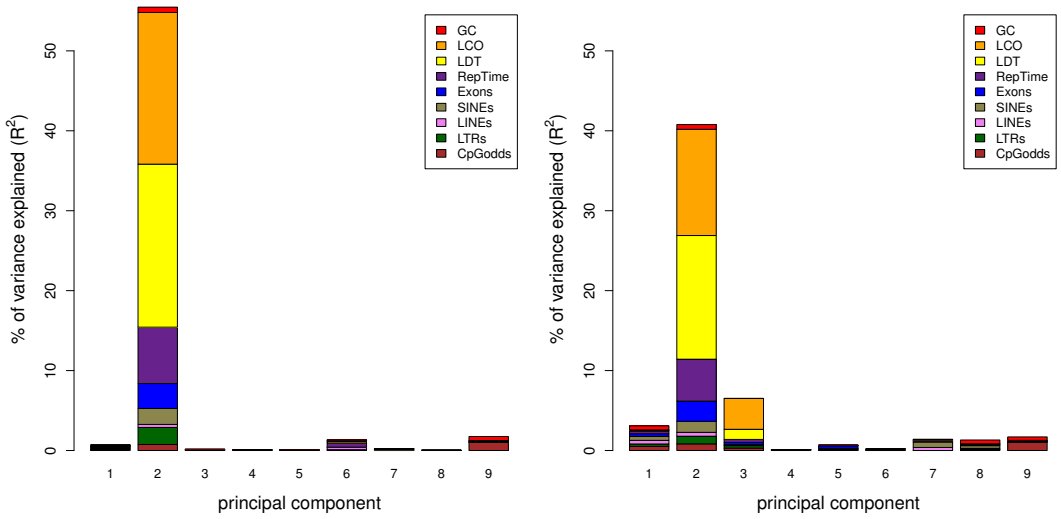


Figure A.6.: Principal component regression for W→S substitution rates in the *HC* (left panel) and *HCM* (right panel) branches. The height of each bar represents how much of the variable's variance the corresponding component explains. Each coloured area is proportional to the relative importance of the corresponding feature inside a component.

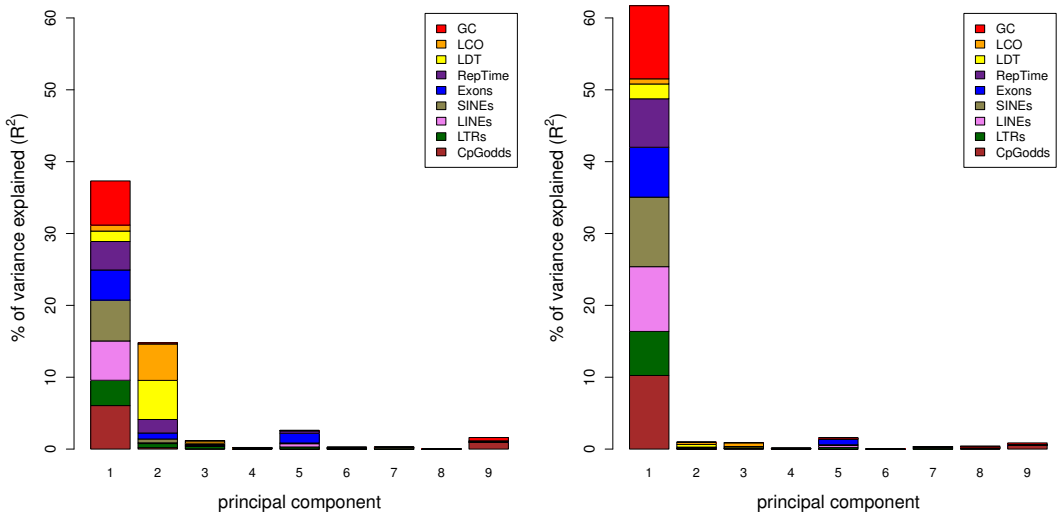


Figure A.7.: Principal component regression for S→W substitution rates in the *HC* (left panel) and *HCM* (right panel) branches. The height of each bar represents how much of the variable's variance the corresponding component explains. Each coloured area is proportional to the relative importance of the corresponding feature inside a component.

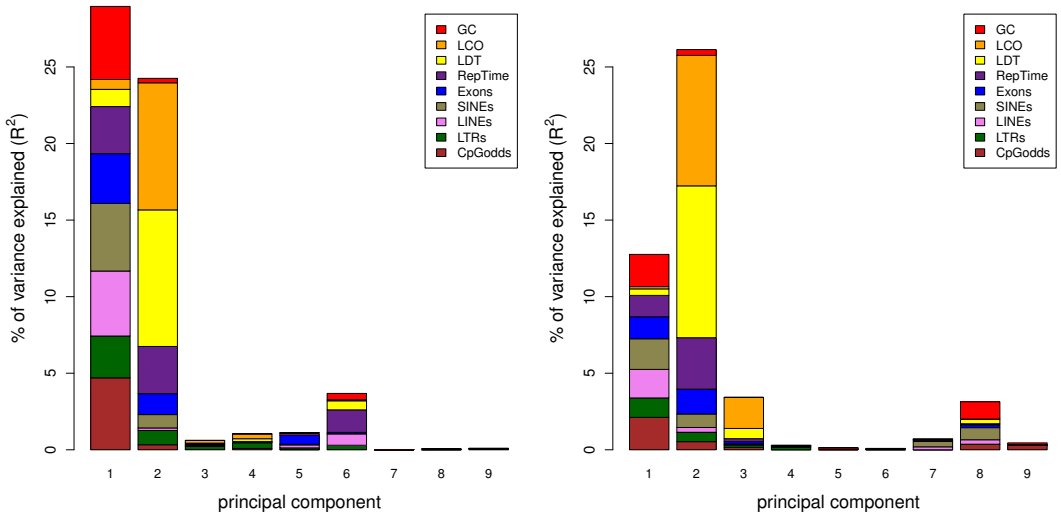


Figure A.8.: Principal component regression for GC* values in the human *HC* (left panel) and *HCM* (right panel) branches. The height of each bar represents how much of the variable's variance the corresponding component explains. Each coloured area is proportional to the relative importance of the corresponding feature inside a component.

B. Appendix B

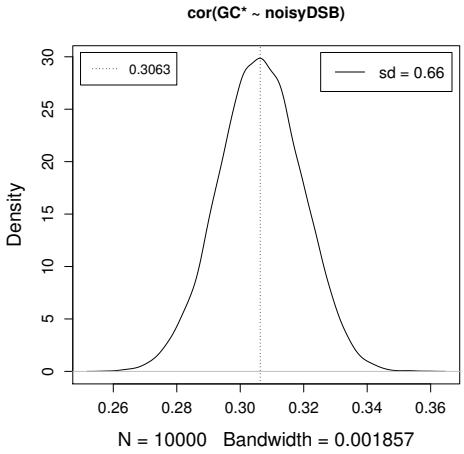
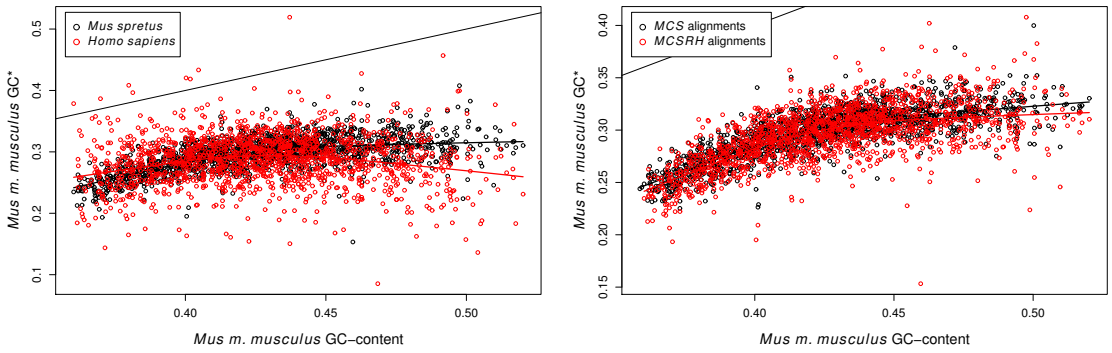


Figure B.1.: Density plot of correlation coefficients between GC* values and noisy DSB hotspot density in the *Mus m. musculus* lineage in *Mus m. musculus* - *Mus m castaneus* - *Mus spretus* alignments. The vertical dotted line indicates the correlation coefficient between GC* and male crossover rates in the same lineage.



(a) (b)

Figure B.2.: GC* values plotted against GC-content values for the *Mus m. musculus* lineage. Substitution patterns were computed by comparing: **(a)** *Mus m. musculus* to *Mus m. castaneus* sequences and using either *Mus spretus* (black) or *Homo sapiens* (red) as an outgroup or **(b)**: *Mus m. musculus* to *Mus m. castaneus* sequences and using *Mus spretus* as an outgroup in *MCS* alignments (black) or *MCSRH* alignments (red). Curves represent LOWESS local regressions. The straight line represents the $x = y$ relationship.

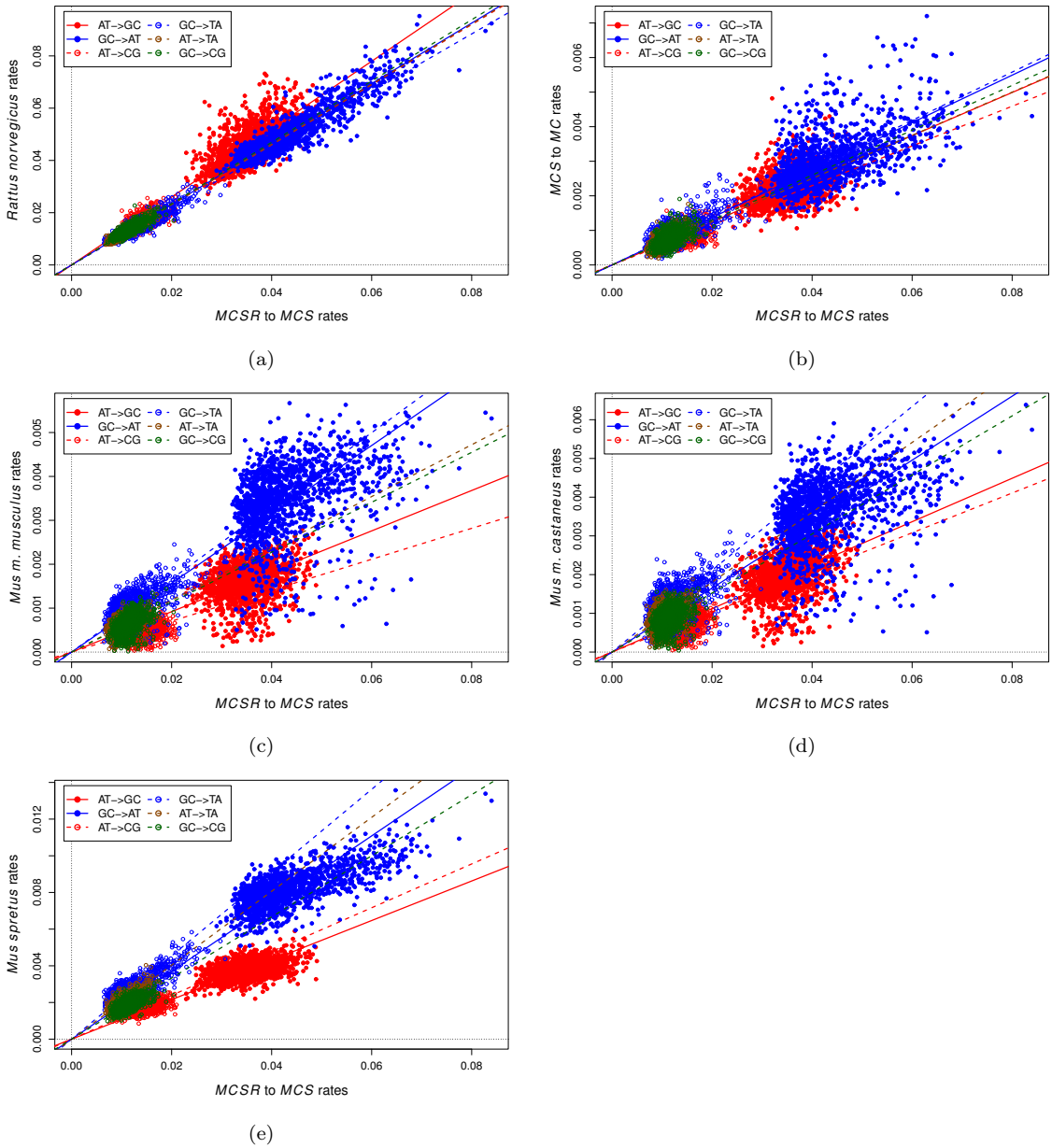
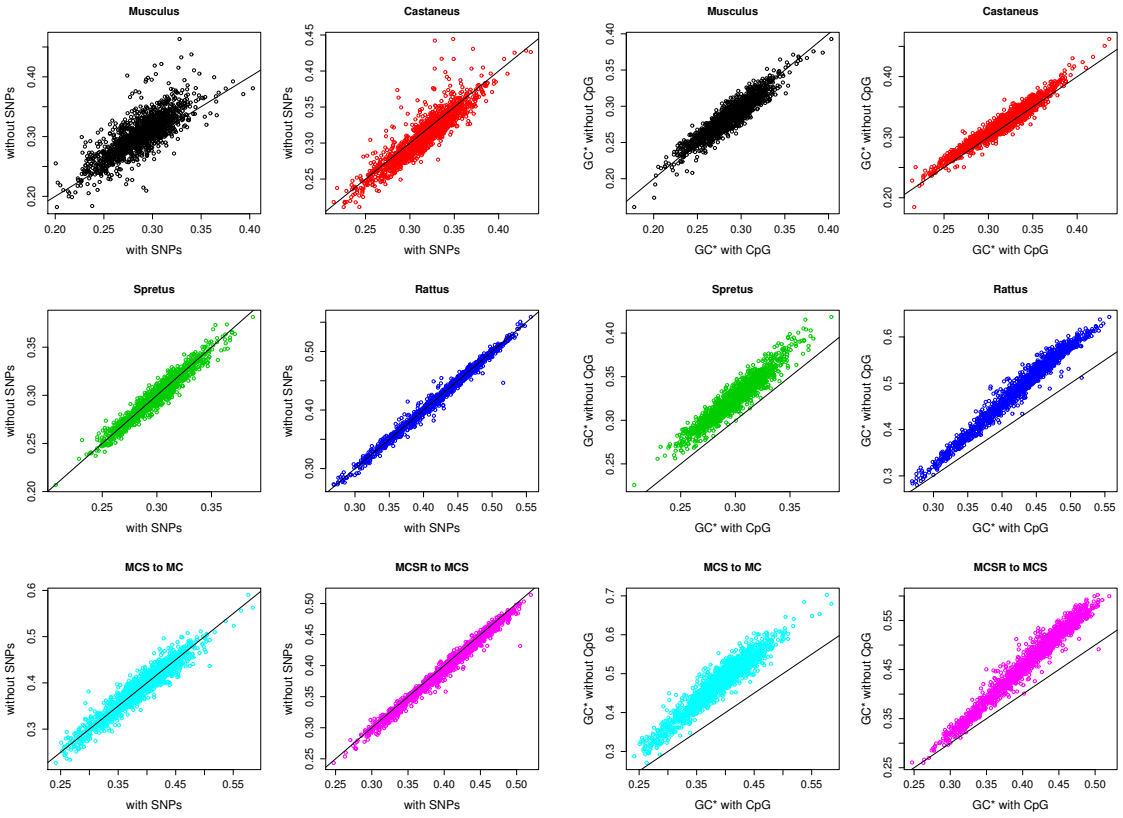
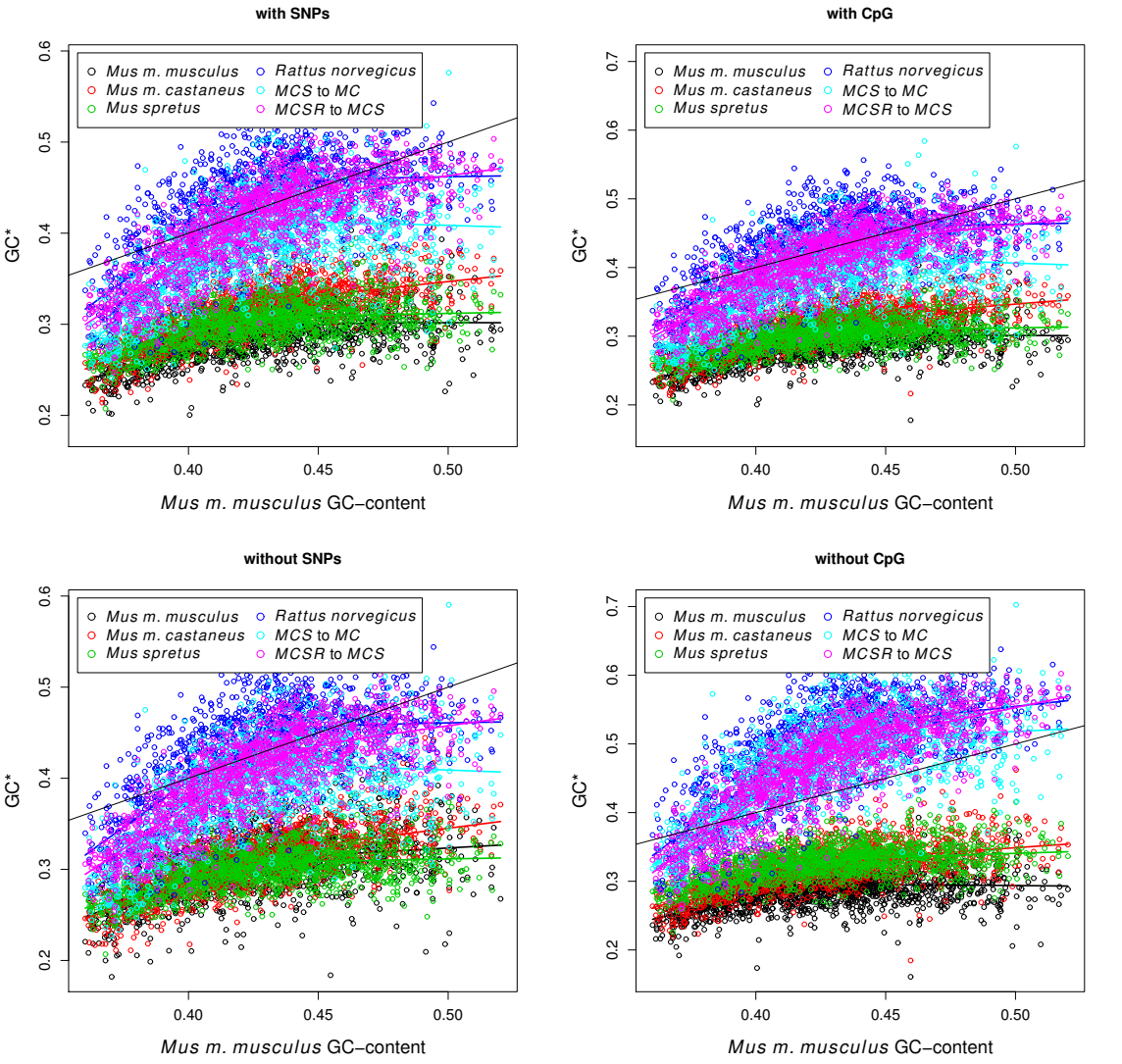


Figure B.3.: Substitution rates for: **(a)** the *Rattus norvegicus* lineage, **(b)** the *MCS* to *MC* branch, **(c)** the *Mus m. musculus* lineage, **(d)** the *Mus m. castaneus* lineage and **(e)** the *Mus spretus* lineage plotted against rates for the *MCSR* to *MCS* branch. Lines represent linear regression between rates in the two branches, with a forced intercept of 0.



(a) (b)

Figure B.4.: **(a)**: GC* values computed from alignments where SNPs in *Mus m. musculus* were masked plotted against original GC* values for each branch studied in *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* - *Rattus norvegicus* - *Homo sapiens* multiple alignments. **(b)**: GC* values computed without taking CpG hypermutability into account plotted against original GC* values for each branch studied in *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* - *Rattus norvegicus* - *Homo sapiens* multiple alignments. Lines represent the $x = y$ relationship.



(a) (b)

Figure B.5.: Original GC* values for each branch studied in *Mus m. musculus* - *Mus m. castaneus* - *Mus spretus* - *Rattus norvegicus* - *Homo sapiens* multiple alignments (left panel) and (right panel) GC* values computed **(a)**: after masking SNPs in *Mus m. musculus* and **(b)**: without taking the CpG hypermutability into account. All GC* values are plotted against *Mus m. musculus* GC-content.

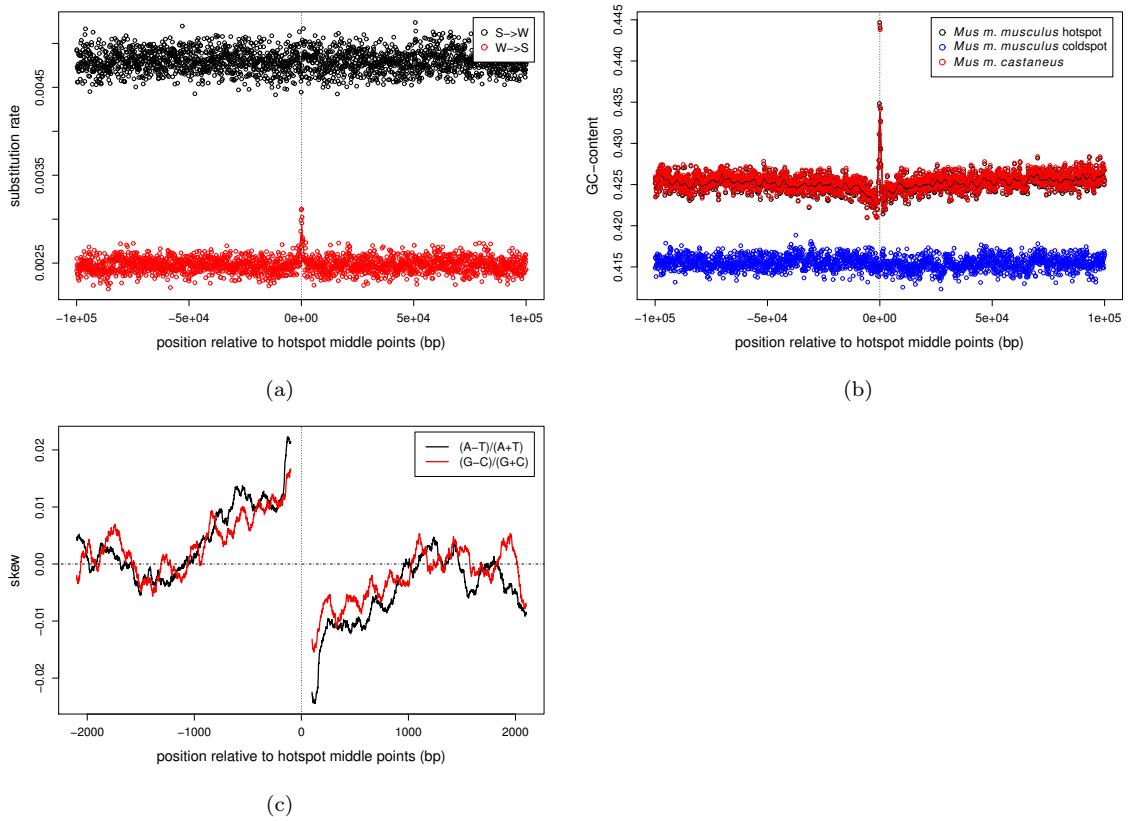


Figure B.6.: **(a)**: W→S (red) and S→W (black) substitution rates around DSB hotspots middle points in the *Mus m. musculus* lineage. Lines represent one-sided local regressions computed over 25 neighboring windows. **(b)**: GC-content around DSB hotspots middle points in *Mus m. musculus* (black), *Mus m. castaneus* (red) and around DSB coldspots in *Mus m. musculus* (blue). Lines represent one-sided local regressions computed over 25 neighboring windows. **(c)**: AT skews (black) and GC skews (red) around DSB hotspots middle points in *Mus m. musculus*. AT and GC skews were computed in 2000 windows upstream and 2000 windows downstream of length 100 bp which overlap for 99 bp.

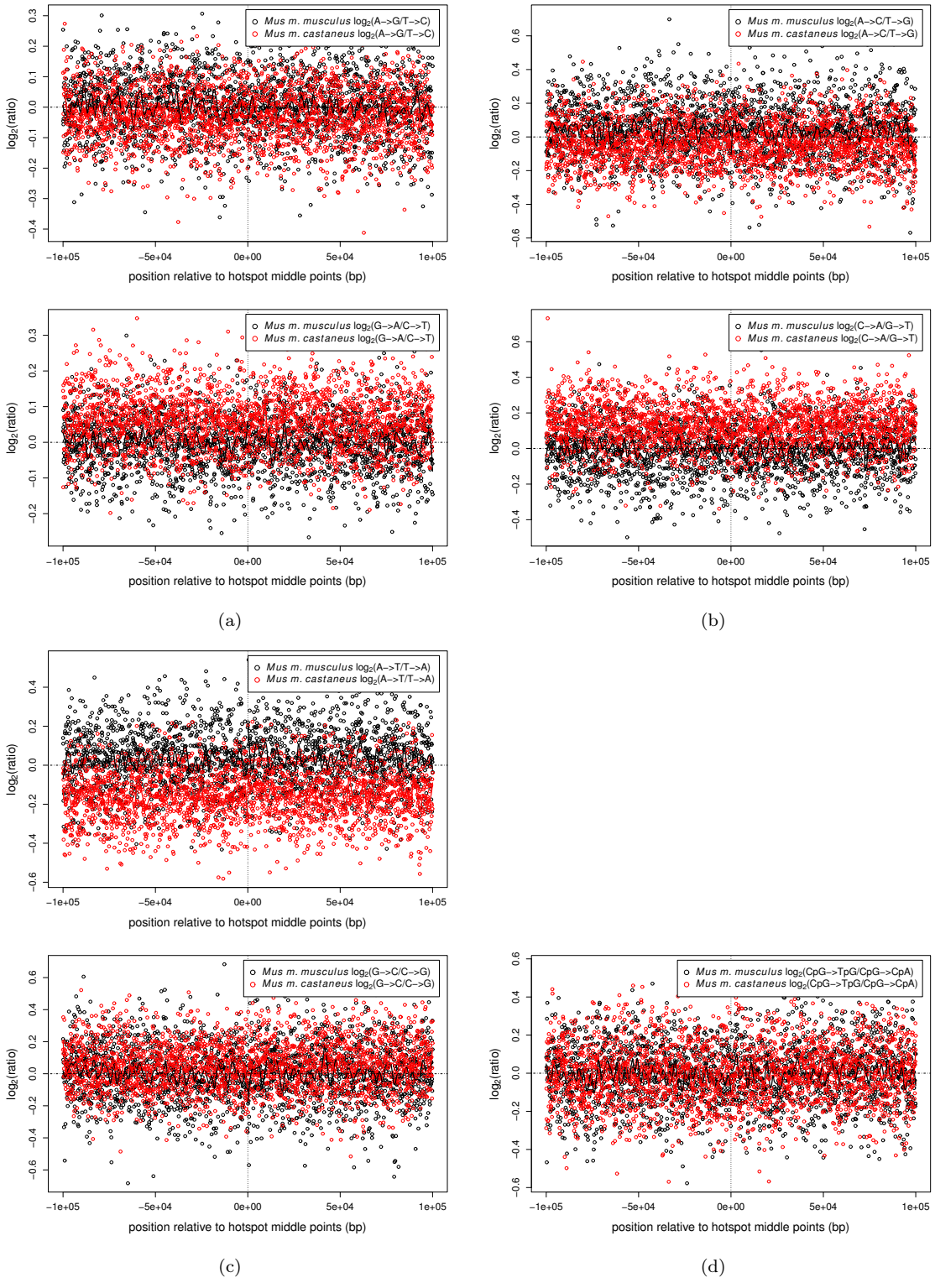


Figure B.7.: Strand asymmetries for transversion rates around DSB hotspots middle points in the *Mus m. musculus* lineage (black) and the *Mus m. castaneus* lineage (red) for **(A)** transition rates, **(b)** & **(c)** transversion rates and **(d)** CpG rates. Lines represent one-sided local regressions computed over 25 neighboring windows.

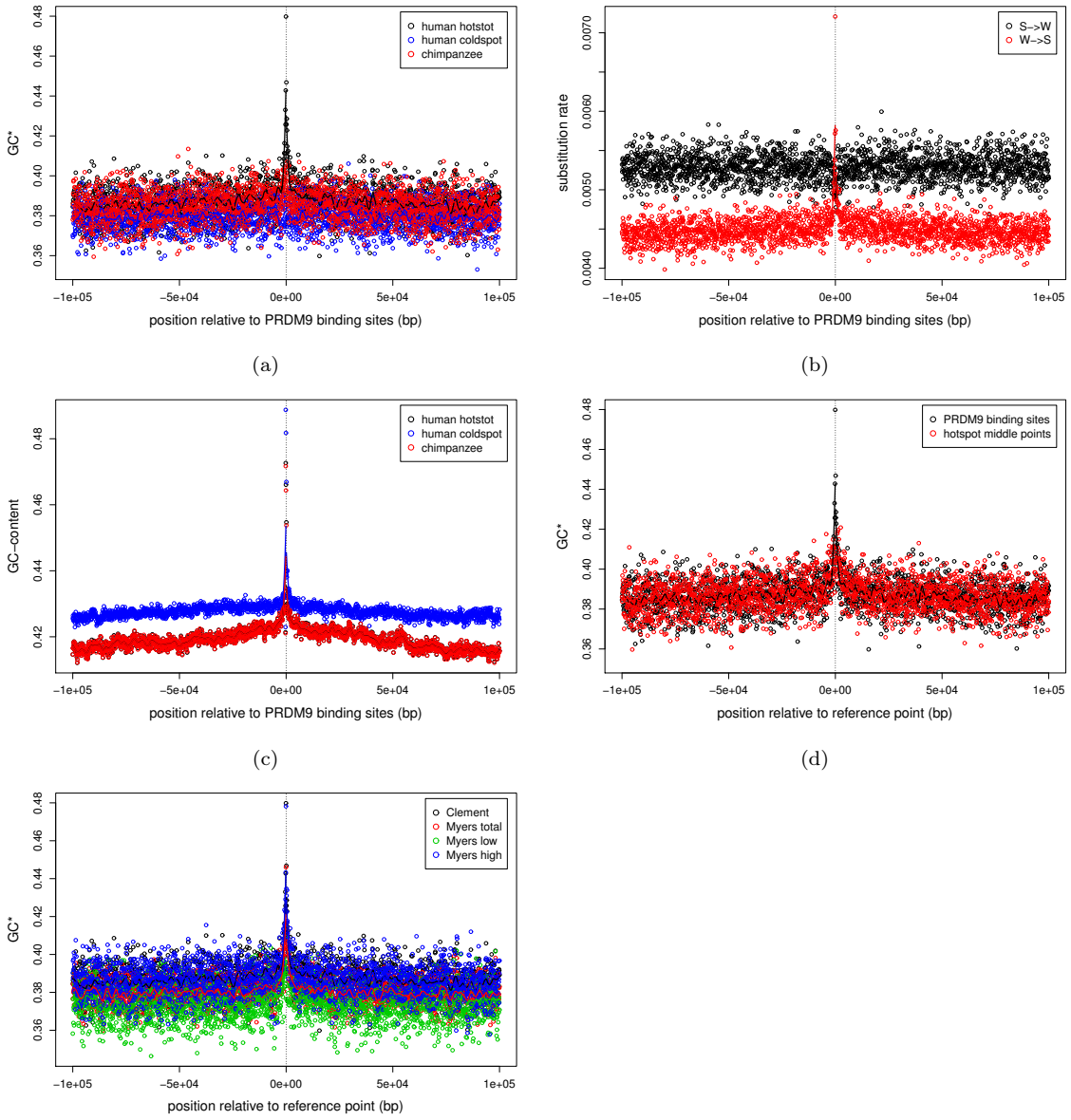
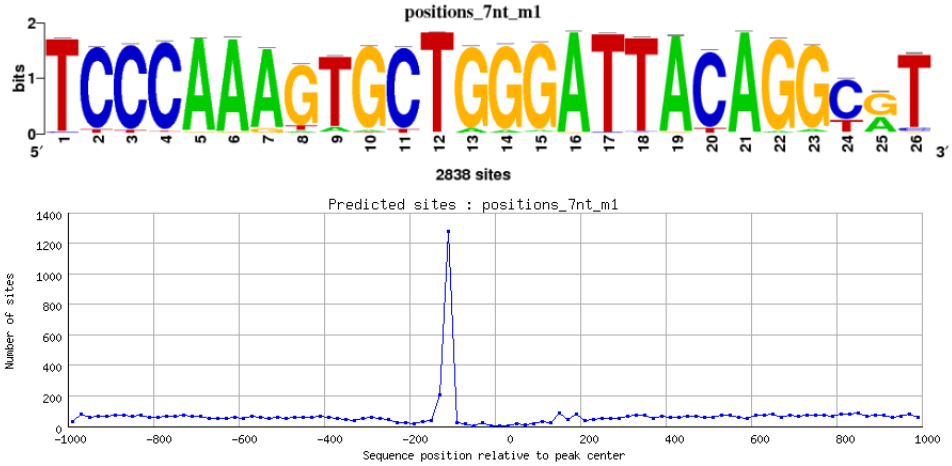
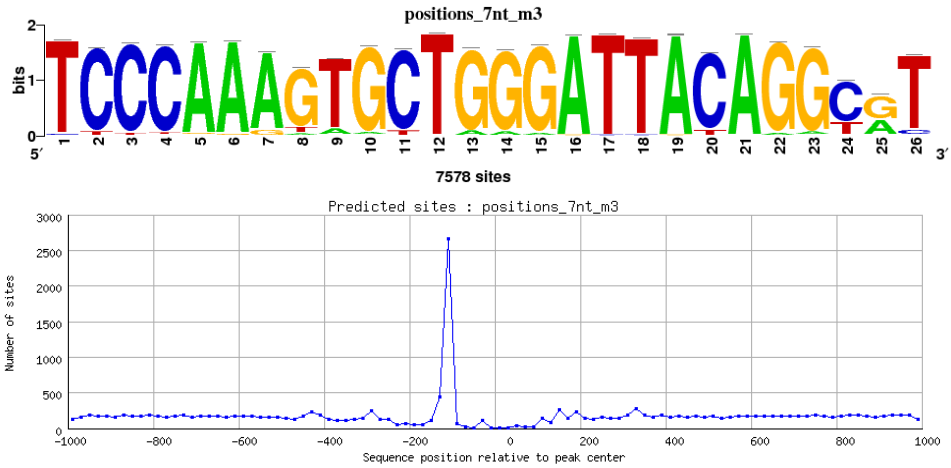


Figure B.8.: **(a)**: GC* around PRDM9 binding sites in recombination hotspots in the human lineage (black), the chimpanzee lineage (red) and around PRDM9 binding sites in recombination coldspots (blue). **(b)**: W→S (red) and S→W (black) substitution rates around PRDM9 binding sites in recombination hotspots in the human lineage. Lines represent one-sided local regressions computed over 25 neighboring windows. **(c)**: GC-content around DSB hotspots middle points in the human lineage (black), the chimpanzee lineage (red) and around PRDM9 binding sites in recombination coldspots (blue). **(d)**: GC* around PRDM9 binding sites (black) or hotspots middle points (red) in human meiotic recombination hotspots. **(e)**: GC* around PRDM9 binding sites in home-made hotspots (black), all Myers hotspots (red), Myers hotspots with low crossover rates (green) and Myers hotspots with high crossover rates (blue).



(a)



(b)

Figure B.9.: Sequence logo (top panel) and relative frequency around PRDM9 binding site (here designated as peak center) of over represented DNA sequence motif in **(a)**: meiotic recombination hotspots and **(b)**: meiotic recombination coldspots.

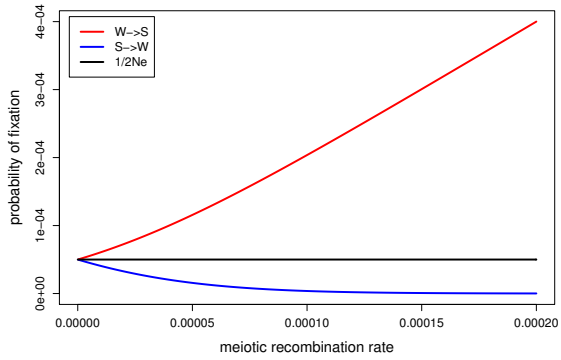


Figure B.10.: Fixation probabilities of W→S (red) and S→W (blue) mutations as a function of gBGC strength (s). The black line represents the fixation probability of S→S and W→W mutations = $1/2N_e$ where $N_e = 10^4$ in this example.

Bibliography

Abrusán, G., H.-J. Krambeck, T. Junier, J. Giordano, and P. E. Warburton, 2008: Biased distributions and decay of long interspersed nuclear elements in the chicken genome. *Genetics*, **178 (1)**, 573–81.

Alföldi, J., F. D. Palma, M. Grabherr, C. Williams, L. Kong, et al., 2011: The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477 (7366)**, 587–91.

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman, 2009: A high density linkage map of the bovine genome. *BMC Genet*, **10**, 18.

Arndt, P. F., C. B. Burge, and T. Hwa, 2003a: DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol*, **10 (3-4)**, 313–22.

Arndt, P. F. and T. Hwa, 2005: Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, **21 (10)**, 2322–8.

Arndt, P. F., T. Hwa, and D. A. Petrov, 2005: Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol*, **60 (6)**, 748–63.

Arndt, P. F., D. A. Petrov, and T. Hwa, 2003b: Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol Biol Evol*, **20 (11)**, 1887–96.

Auton, A., A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel, et al., 2012: A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science*.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, et al., 2010: PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, **327 (5967)**, 836–40.

Baudat, F. and B. de Massy, 2007: Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res*, **15 (5)**, 565–77.

Belle, E. M. S., L. Duret, N. Galtier, and A. Eyre-Walker, 2004: The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. *J Mol Evol*, **58 (6)**, 653–60.

Berg, I. L., R. Neumann, K.-W. G. Lam, S. Sarbajna, L. Odenthal-Hesse, et al., 2010: PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet*, **42 (10)**, 859–63.

Berg, I. L., R. Neumann, S. Sarbajna, L. Odenthal-Hesse, N. J. Butler, et al., 2011: Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci USA*, **108 (30)**, 12 378–83.

Berglund, J., K. S. Pollard, and M. T. Webster, 2009: Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*, **7 (1)**, e26.

Bernardi, G., 2000: Isochores and the evolutionary genomics of vertebrates. *Gene*, **241 (1)**, 3–17.

Bernardi, G., 2007: The neoselectionist theory of genome evolution. *Proc Natl Acad Sci USA*, **104 (20)**, 8385–90.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, et al., 1985: The mosaic genome of warm-blooded vertebrates. *Science*, **228 (4702)**, 953–8.

Bill, C. A., W. A. Duran, N. R. Miselis, and J. A. Nickoloff, 1998: Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. *Genetics*, **149 (4)**, 1935–43.

Bird, A. P., 1978: Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol*, **118 (1)**, 49–60.

Borde, V., N. Robine, W. Lin, S. Bonfils, V. Géli, et al., 2009: Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J*, **28 (2)**, 99–111.

Boulton, A., R. S. Myers, and R. J. Redfield, 1997: The hotspot conversion paradox and the evolution of meiotic recombination. *Proc Natl Acad Sci USA*, **94 (15)**, 8058–63.

Brown, T. C. and J. Jiricny, 1988: Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell*, **54 (5)**, 705–11.

Buard, J., P. Barthès, C. Grey, and B. de Massy, 2009: Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *EMBO J*, **28 (17)**, 2616–24.

Capra, J. A. and K. S. Pollard, 2011: Substitution patterns are gc-biased in divergent sequences across the metazoans. *Genome Biol Evol*, **3**, 516–27.

Chen, C.-L., L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, et al., 2011: Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol*, **28 (8)**, 2327–37.

Chen, C.-L., A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, et al., 2010: Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res*, **20 (4)**, 447–57.

Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, 2007: Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*, **8 (10)**, 762–75.

Clément, Y. and P. F. Arndt, 2011: Substitution patterns are under different influences in primates and rodents. *Genome Biol Evol*, **3**, 236–45.

Coop, G. and S. R. Myers, 2007: Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet*, **3 (3)**, e35.

Coop, G. and M. Przeworski, 2007: An evolutionary view of human recombination. *Nat Rev Genet*, **8 (1)**, 23–34.

Cortadas, J., G. Macaya, and G. Bernardi, 1977: An analysis of the bovine genome by density gradient centrifugation: fractionation in Cs2SO4/3,6-bis(acetatomercurimethyl)dioxane density gradient. *Eur J Biochem*, **76 (1)**, 13–9.

Costantini, M. and G. Bernardi, 2008: Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci USA*, **105 (9)**, 3433–7.

Cox, A., C. L. Ackert-Bicknell, B. L. Dumont, Y. Ding, J. T. Bell, et al., 2009: A new standard genetic map for the laboratory mouse. *Genetics*, **182 (4)**, 1335–44.

Cuny, G., P. Soriano, G. Macaya, and G. Bernardi, 1981: The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem*, **115 (2)**, 227–33.

de Massy, B., 2003: Distribution of meiotic recombination sites. *Trends Genet*, **19 (9)**, 514–22.

de Villena, F. P.-M. and C. Sapienza, 2001: Recombination is proportional to the number of chromosome arms in mammals. *Mamm Genome*, **12 (4)**, 318–22.

Drummond, D. A., A. Raval, and C. O. Wilke, 2006: A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*, **23 (2)**, 327–37.

Dumont, B. L., M. A. White, B. Steffy, T. Wiltshire, and B. A. Payseur, 2011: Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res*, **21 (1)**, 114–25.

Duret, L., 2006: The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. *J Mol Evol*, **62 (6)**, 803–6.

Duret, L. and P. F. Arndt, 2008: The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, **4 (5)**, e1000 071.

Duret, L. and N. Galtier, 2009: Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, **10**, 285–311.

Duret, L., D. Mouchiroud, and C. Gautier, 1995: Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol*, **40 (3)**, 308–17.

Duret, L., M. Semon, G. Piganeau, D. Mouchiroud, and N. Galtier, 2002: Vanishing GC-rich isochores in mammalian genomes. *Genetics*, **162 (4)**, 1837–47.

Escobar, J. S., S. Glémin, and N. Galtier, 2011: Gc-biased gene conversion impacts ribosomal dna evolution in vertebrates, angiosperms, and other eukaryotes. *Mol Biol Evol*, **28 (9)**, 2561–75.

Eyre-Walker, A., 1993: Recombination and mammalian genome evolution. *Proc Biol Sci*, **252 (1335)**, 237–43.

Eyre-Walker, A., 1998: Problems with parsimony in sequences of biased base composition. *J Mol Evol*, **47 (6)**, 686–90.

Eyre-Walker, A., 1999: Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, **152 (2)**, 675–83.

Eyre-Walker, A. and L. D. Hurst, 2001: The evolution of isochores. *Nat Rev Genet*, **2 (7)**, 549–55.

Fabre, P.-H., A. Rodrigues, and E. J. P. Douzery, 2009: Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol*, **53 (3)**, 808–25.

Federico, C., S. Saccone, and G. Bernardi, 1998: The gene-richest bands of human chromosomes replicate at the onset of the S-phase. *Cytogenet Cell Genet*, **80 (1-4)**, 83–8.

Felsenstein, J., 1978: Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, **27 (4)**, 401.

Felsenstein, J., 1981: Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17 (6)**, 368–76.

Felsenstein, J., 2005: *Theoretical Evolutionary Genetics*. Joseph Felsenstein.

Filipski, J., 1988: Why the rate of silent codon substitutions is variable within a vertebrate's genome. *J Theor Biol*, **134 (2)**, 159–64.

Filipski, J., J. P. Thiery, and G. Bernardi, 1973: An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. *J Mol Biol*, **80 (1)**, 177–97.

Fledel-Alon, A., D. J. Wilson, K. Broman, X. Wen, C. Ober, et al., 2009: Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet*, **5 (9)**, e1000 658.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, et al., 2011: Ensembl 2011. *Nucleic Acids Res*, **39 (Database issue)**, D800–6.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, et al., 2012: Ensembl 2012. *Nucleic Acids Res*, **40 (Database issue)**, D84–90.

Franklin, R. E. and R. G. Gosling, 1953: Molecular configuration in sodium thymonucleate. *Nature*, **171 (4356)**, 740–1.

Fujita, M. K., S. V. Edwards, and C. P. Ponting, 2011: The Anolis lizard genome: an amniote genome without isochores. *Genome Biol Evol*, **3**, 974–84.

Fullerton, S. M., A. B. Carvalho, and A. G. Clark, 2001: Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol*, **18 (6)**, 1139–42.

Galtier, N., E. Bazin, and N. Bierne, 2006: Gc-biased segregation of noncoding polymorphisms in drosophila. *Genetics*, **172 (1)**, 221–8.

Galtier, N. and L. Duret, 2007: Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet*, **23 (6)**, 273–7.

Galtier, N., L. Duret, S. Glémin, and V. Ranwez, 2009: GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*, **25 (1)**, 1–5.

Galtier, N. and D. Mouchiroud, 1998: Isochore evolution in mammals: a human-like ancestral structure. *Genetics*, **150 (4)**, 1577–84.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret, 2001: GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, **159 (2)**, 907–11.

Geraldes, A., P. Basset, B. Gibson, K. L. Smith, B. Harr, et al., 2008: Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol Ecol*, **17 (24)**, 5349–63.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, et al., 2000: Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci USA*, **97 (21)**, 11 383–90.

Giannelli, F., T. Anagnostopoulos, and P. M. Green, 1999: Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. *Am J Hum Genet*, **65 (6)**, 1580–7.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, et al., 2004: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428 (6982)**, 493–521.

Grey, C., P. Barthès, G. C.-L. Friec, F. Langa, F. Baudat, et al., 2011: Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol*, **9 (10)**, e1001 176.

Groenen, M. A. M., P. Wahlberg, M. Foglio, H. H. Cheng, H.-J. Megens, et al., 2009: A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res*, **19 (3)**, 510–9.

Guillon, H. and B. de Massy, 2002: An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat Genet*, **32 (2)**, 296–9.

Hamada, K., T. Horiike, S. Kanaya, H. Nakamura, H. Ota, et al., 2002: Changes in body temperature pattern in vertebrates do not influence the codon usages of alpha-globin genes. *Genes Genet Syst*, **77 (3)**, 197–207.

Hamada, K., T. Horiike, H. Ota, K. Mizuno, and T. Shinozawa, 2003: Presence of isochore structures in reptile genomes suggested by the relationship between GC contents of intron regions and those of coding regions. *Genes Genet Syst*, **78 (2)**, 195–8.

Harr, B., 2006: Genomic islands of differentiation between house mouse subspecies. *Genome Res*, **16 (6)**, 730–7.

Harrison, R. J. and B. Charlesworth, 2011: Biased gene conversion affects patterns of codon usage and amino acid usage in the saccharomyces sensu stricto group of yeasts. *Mol Biol Evol*, **28 (1)**, 117–29.

Hedges, S. B., J. Dudley, and S. Kumar, 2006: TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22 (23)**, 2971–2.

Hinch, A. G., A. Tandon, N. Patterson, Y. Song, N. Rohland, et al., 2011: The landscape of recombination in African Americans. *Nature*, **476 (7359)**, 170–5.

Hiratani, I., T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, et al., 2008: Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol*, **6 (10)**, e245.

Hochwagen, A. and G. A. B. Marais, 2010: Meiosis: a PRDM9 guide to the hotspots of recombination. *Curr Biol*, **20 (6)**, R271–4.

Hodgkinson, A., E. Ladoukakis, and A. Eyre-Walker, 2009: Cryptic variation in the human mutation rate. *PLoS Biol*, **7 (2)**, e1000 027.

Hubbard, T. J. P., B. L. Aken, S. Ayling, B. Ballester, K. Beal, et al., 2009: Ensembl 2009. *Nucleic Acids Res*, **37 (Database issue)**, D690–7.

Huchon, D., P. Chevret, U. Jordan, C. W. Kilpatrick, V. Ranwez, et al., 2007: Multiple molecular evidences for a living mammalian fossil. *Proc Natl Acad Sci USA*, **104 (18)**, 7495–9.

Hughes, S., O. Clay, and G. Bernardi, 2002: Compositional patterns in reptilian genomes. *Gene*, **295 (2)**, 323–9.

Hughes, S., D. Zelus, and D. Mouchiroud, 1999: Warm-blooded isochore structure in Nile crocodile and turtle. *Mol Biol Evol*, **16 (11)**, 1521–7.

Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, et al., 2007: Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460.

Ideraabdullah, F. Y., E. de la Casa-Esperón, T. A. Bell, D. A. Detwiler, T. Magnuson, et al., 2004: Genetic and haplotype diversity among wild-derived mouse inbred strains. *Genome Res*, **14 (10A)**, 1880–7.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, et al., 2007: A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449 (7164)**, 851–61.

Janes, D. E., C. L. Organ, M. K. Fujita, A. M. Shedlock, and S. V. Edwards, 2010: Genome evolution in Reptilia, the sister group of mammals. *Annu Rev Genomics Hum Genet*, **11**, 239–64.

Jeffreys, A. J. and C. A. May, 2004: Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, **36 (2)**, 151–6.

Jeffreys, A. J. and R. Neumann, 2002: Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*, **31 (3)**, 267–71.

Jeffreys, A. J. and R. Neumann, 2009: The rise and fall of a human recombination hot spot. *Nat Genet*, **41 (5)**, 625–9.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, et al., 2004: Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res*, **14 (4)**, 528–38.

John, J. A. S., E. L. Braun, S. R. Isberg, L. G. Miles, A. Y. Chong, et al., 2012: Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol*, **13 (1)**, 415.

Jukes, T. H. and C. R. Cantor, 1969: *Evolution of Protein Molecules*. New York: Academic Press.

Kaback, D. B., 1996: Chromosome-size dependent control of meiotic recombination in humans. *Nat Genet*, **13 (1)**, 20–1.

Katzman, S., J. A. Capra, D. Haussler, and K. S. Pollard, 2011: Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol*, **3**, 614–26.

Keane, T. M., L. Goodstadt, P. Danecek, M. A. White, K. Wong, et al., 2011: Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477 (7364)**, 289–94.

Keightley, P. D., M. J. Lercher, and A. Eyre-Walker, 2005: Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol*, **3 (2)**, e42.

Khelifi, A., J. Meunier, L. Duret, and D. Mouchiroud, 2006: GC content evolution of the human and mouse genomes: insights from the study of processed pseudogenes in regions of different recombination rates. *J Mol Evol*, **62 (6)**, 745–52.

Kimura, M., 1962: On the probability of fixation of mutant genes in a population. *Genetics*, **47**, 713–9.

Kimura, M., 1968: *The neutral theory of molecular evolution*. Cambridge University Press.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, et al., 2002: A high-resolution recombination map of the human genome. *Nat Genet*, **31 (3)**, 241–7.

Kong, A., G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, et al., 2010: Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, **467 (7319)**, 1099–103.

Kostka, D., M. J. Hubisz, A. Siepel, and K. S. Pollard, 2012: The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol*, **29 (3)**, 1047–57.

Kumar, S. and S. B. Hedges, 2011: TimeTree2: species divergence times on the iPhone. *Bioinformatics*, **27 (14)**, 2023–4.

Kvikstad, E. M. and K. D. Makova, 2010: The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res*, **20 (5)**, 600–13.

Kvikstad, E. M., S. Tyekucheva, F. Chiaromonte, and K. D. Makova, 2007: A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol*, **3 (9)**, 1772–82.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al., 2001: Initial sequencing and analysis of the human genome. *Nature*, **409 (6822)**, 860–921.

Li, W. and J. Freudenberg, 2009: Two-parameter characterization of chromosome-scale recombination rate. *Genome Res*, **19 (12)**, 2300–7.

Lynch, M., 2007: *The Origins of Genome Architecture*. Sinauer Associates Inc.

Macaya, G., J. Cortadas, and G. Bernardi, 1978: An analysis of the bovine genome by density-gradient centrifugation. Preparation of the dG+dC-rich DNA components. *Eur J Biochem*, **84 (1)**, 179–88.

Macaya, G., J. P. Thiery, and G. Bernardi, 1976: An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol*, **108 (1)**, 237–54.

Marais, G., 2003: Biased gene conversion: implications for genome and sex evolution. *Trends Genet*, **19 (6)**, 330–8.

Meselson, M., F. W. Stahl, and J. Vinograd, 1957: Equilibrium sedimentation of macromolecules in density gradients. *Proc Natl Acad Sci USA*, **43 (7)**, 581–8.

Meunier, J. and L. Duret, 2004: Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*, **21 (6)**, 984–90.

Mevik, B. and R. Wehrens, 2007: The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, **18 (2)**, 1–24.

Montoya-Burgos, J. I., P. Boursot, and N. Galtier, 2003: Recombination explains isochores in mammalian genomes. *Trends Genet*, **19 (3)**, 128–30.

Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, et al., 2002: Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420 (6915)**, 520–62.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005: A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310 (5746)**, 321–4.

Myers, S., R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, et al., 2010: Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327 (5967)**, 876–9.

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean, 2008: A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, **40 (9)**, 1124–9.

Nagylaki, T., 1983: Evolution of a finite population under gene conversion. *Proc Natl Acad Sci USA*, **80 (20)**, 6278–81.

Necşulea, A., A. Popa, D. N. Cooper, P. D. Stenson, D. Mouchiroud, et al., 2011: Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat*, **32 (2)**, 198–206.

Oliver, P. L., L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, et al., 2009: Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet*, **5 (12)**, e1000 753.

Paigen, K. and P. Petkov, 2010: Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*, **11 (3)**, 221–33.

Paigen, K., J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov, et al., 2008: The recombinational anatomy of a mouse chromosome. *PLoS Genet*, **4 (7)**, e1000 119.

Parvanov, E. D., P. M. Petkov, and K. Paigen, 2010: Prdm9 controls activation of mammalian recombination hotspots. *Science*, **327 (5967)**, 835.

Pereira, S. L. and A. J. Baker, 2006: A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol Biol Evol*, **23 (9)**, 1731–40.

Petronczki, M., M. F. Siomos, and K. Nasmyth, 2003: Un ménage à quatre: the molecular biology of chromosome segregation in meiosis. *Cell*, **112 (4)**, 423–40.

Pineda-Krch, M. and R. J. Redfield, 2005: Persistence and loss of meiotic recombination hotspots. *Genetics*, **169 (4)**, 2319–33.

Pink, C. J. and L. D. Hurst, 2010: Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol Biol Evol*, **27 (5)**, 1077–86.

Pink, C. J. and L. D. Hurst, 2011: Late replicating domains are highly recombining in females but have low male recombination rates: implications for isochore evolution. *PLoS ONE*, **6 (9)**, e24 480.

Polak, P. and P. F. Arndt, 2008: Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res*, **18 (8)**, 1216–23.

Pollard, K. S., S. R. Salama, B. King, A. D. Kern, T. Dreszer, et al., 2006a: Forces shaping the fastest evolving regions in the human genome. *PLoS Genet*, **2 (10)**, e168.

Pollard, K. S., S. R. Salama, N. Lambert, M.-A. Lambot, S. Coppens, et al., 2006b: An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443 (7108)**, 167–72.

Ponting, C. P., 2011: What are the genomic drivers of the rapid evolution of PRDM9? *Trends Genet*, **27 (5)**, 165–71.

Popa, A., P. Samollow, C. Gautier, and D. Mouchiroud, 2012: The sex-specific impact of meiotic recombination on nucleotide composition. *Genome Biol Evol*.

Poux, C., P. Chevret, D. Huchon, W. W. de Jong, and E. J. P. Douzery, 2006: Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. *Syst Biol*, **55 (2)**, 228–44.

Pozzoli, U., G. Menozzi, M. Fumagalli, M. Cereda, G. P. Comi, et al., 2008: Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol*, **8**, 99.

Ptak, S. E., D. A. Hinds, K. Koehler, B. Nickel, N. Patil, et al., 2005: Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, **37 (4)**, 429–34.

Ptak, S. E., A. D. Roeder, M. Stephens, Y. Gilad, S. Pääbo, et al., 2004: Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol*, **2 (6)**, e155.

Romiguier, J., V. Ranwez, E. J. P. Douzery, and N. Galtier, 2010: Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, **20 (8)**, 1001–9.

Ryba, T., I. Hiratani, J. Lu, M. Itoh, M. Kulik, et al., 2010: Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, **20 (6)**, 761–70.

Salcedo, T., A. Geraldes, and M. W. Nachman, 2007: Nucleotide variation in wild and inbred mice. *Genetics*, **177 (4)**, 2277–91.

Shifman, S., J. T. Bell, R. R. Copley, M. S. Taylor, R. W. Williams, et al., 2006: A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol*, **4 (12)**, e395.

Sigurdsson, M. I., A. V. Smith, H. T. Bjornsson, and J. J. Jonsson, 2009: HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res*, **19 (4)**, 581–9.

Smagulova, F., I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, et al., 2011: Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, **472 (7343)**, 375–8.

Smit, A. F., 1999: Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev*, **9 (6)**, 657–63.

Smith, K. N. and A. Nicolas, 1998: Recombination at work for meiosis. *Curr Opin Genet Dev*, **8 (2)**, 200–11.

Smith, N. G. C. and A. Eyre-Walker, 2002: The compositional evolution of the murid genome. *J Mol Evol*, **55 (2)**, 197–201.

Spencer, C. C. A., P. Deloukas, S. Hunt, J. Mullikin, S. Myers, et al., 2006: The influence of recombination on human genetic diversity. *PLoS Genet*, **2 (9)**, e148.

Squartini, F. and P. F. Arndt, 2008: Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol Biol Evol*, **25 (12)**, 2525–35.

Thiery, J. P., G. Macaya, and G. Bernardi, 1976: An analysis of eukaryotic genomes by density gradient centrifugation. *J Mol Biol*, **108 (1)**, 219–35.

Thomas-Chollier, M., C. Herrmann, M. Defrance, O. Sand, D. Thieffry, et al., 2012: RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*, **40 (4)**, e31.

Touchon, M., S. Nicolay, B. Audit, E.-B. B. of Brodie, Y. d'Aubenton Carafa, et al., 2005: Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci USA*, **102 (28)**, 9836–41.

Tyekucheva, S., K. D. Makova, J. E. Karro, R. C. Hardison, W. Miller, et al., 2008: Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol*, **9 (4)**, R76.

Veyrunes, F., J. Britton-Davidian, T. J. Robinson, E. Calvet, C. Denys, et al., 2005: Molecular phylogeny of the African pygmy mice, subgenus Nannomys (Rodentia, Murinae, Mus): implications for chromosomal evolution. *Mol Phylogenet Evol*, **36 (2)**, 358–69.

Walser, J.-C. and A. V. Furano, 2010: The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Res*, **20 (7)**, 875–82.

Walser, J.-C., L. Ponger, and A. V. Furano, 2008: CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Res*, **18 (9)**, 1403–14.

Watanabe, Y., A. Fujiyama, Y. Ichiba, M. Hattori, T. Yada, et al., 2002: Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum Mol Genet*, **11 (1)**, 13–21.

Watson, J. D. and F. H. Crick, 1953: Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171 (4356)**, 737–8.

Webb, A. J., I. L. Berg, and A. Jeffreys, 2008: Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci USA*, **105 (30)**, 10 471–6.

Webster, M. T., E. Axelsson, and H. Ellegren, 2006: Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol*, **23 (6)**, 1203–16.

Webster, M. T., N. G. C. Smith, L. Hultin-Rosenberg, P. F. Arndt, and H. Ellegren, 2005: Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol*, **22 (6)**, 1468–74.

Wilkins, M. H. F., A. R. Stokes, and H. R. Wilson, 1953: Molecular structure of deoxypentose nucleic acids. *Nature*, **171 (4356)**, 738–40.

Winckler, W., S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, et al., 2005: Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, **308 (5718)**, 107–11.

Wolfe, K. H., P. M. Sharp, and W. H. Li, 1989: Mutation rates differ among regions of the mammalian genome. *Nature*, **337 (6204)**, 283–5.

Wong, A. K., A. L. Ruhe, B. L. Dumont, K. R. Robertson, G. Guerrero, et al., 2010: A comprehensive linkage map of the dog genome. *Genetics*, **184 (2)**, 595–605.

Wu, Z. K., I. V. Getun, and P. R. J. Bois, 2010: Anatomy of mouse recombination hot spots. *Nucleic Acids Res*, **38 (7)**, 2346–54.

Yang, S., A. F. Smit, S. Schwartz, F. Chiaromonte, K. M. Roskin, et al., 2004: Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res*, **14 (4)**, 517–27.

Yu, N., M. I. Jensen-Seaman, L. Chemnick, J. R. Kidd, A. S. Deinard, et al., 2003: Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, **164 (4)**, 1511–8.

Notations and Abbreviations

A Adenine

T Thymine

G Guanine

C Cytosine

DNA Deoxyribonucleic acid

SINEs Short interspersed elements

LINEs Long interspersed elements

gBGC GC-biased gene conversion

SNPs Single nucleotide polymorphisms

DSB Double strand break

GC* Equilibrium GC-content

CpGodds Observed CpG dinucleotide normalized by expected CpG dinucleotide frequency

LTR Long terminal repeats

CO Crossover rates

LCO Logarithm of crossover rates

LDT Logarithm of the distance to telomeres

RepTime Replication-timing

RCVE Relative contribution to variability explained

PCA Principal component analysis

PCR Principal component regression

PC Principal component

ChIP-seq Chromatine immuno-precipitation followed by sequencing

ILS Incomplete lineage sorting

Indels Insertion or deletion

TE Transposable elements

Summary

As one of the most basic properties of genomic sequences, base composition has been extensively studied for years. It is traditionally summarized by the GC-content, the frequency of G and C bases in the sequence of interest. One striking feature of mammalian genomes is the fact that GC-content is not homogeneous along chromosomes: one can observe large-scale variations of the GC-content. These variations have been called isochores and are linked to a number of genomic features such as gene density or replication-timing. While different hypotheses have been put forward over the years to explain these GC-content variations, GC-biased gene conversion has been identified as a major force influencing GC-content evolution. This process is neutral and works as follows. During meiotic recombination, double strand breaks are repaired by gene conversion, the copy and paste of one DNA fragment in another. Mismatches can occur during this copy step, which repair mechanism is biased towards G and C. As a result, the fixation of G and C alleles is going to be favored over that of A and T alleles. In this thesis, we investigated base composition variations and GC-content evolution in mammalian genomes.

We first estimated GC-content variations for random DNA sequences and compared them to that of mammalian genomic sequences and found that base composition is more variable than expected by chance in these genomes. We then analyzed GC-content variations along the genome of several organisms and were able to find major differences between groups of organisms, for example rodents' genomes have a much less variable base composition than primates' genomes.

We then investigated substitution patterns and GC-content evolution across mouse and human genomes using a comparative approach. We found that GC-biased gene conversion is active in the mouse genome but that GC-content is evolving differently in the human and mouse genomes. Furthermore, we investigated substitution patterns and how much different genomic features influence them. We found that, while meiotic recombination through GC-biased gene conversion is the major feature influencing A or T $\rightarrow$ G or C substitution rates in the human genome, the CpG dinucleotide content best predicts these substitution rates in the mouse genome, showing that GC-biased gene conversion is active but weak in this genome and that substitution patterns are under different influences in the human and mouse genomes.

The recent discovery that the *Prdm9* gene controls meiotic recombination in mammals as well as its binding motif in human meiotic recombination hotspots in human and the publication of double strand breaks hotspots in mouse enabled the study of the influence of meiotic recombination on substitution patterns at a fine-scale and

derive characteristics of meiotic recombination hotspots. Also, the publication of several mouse subspecies' genomes allowed the study of substitution patterns at short timescales as well as in more mouse lineages than was previously possible. We found that double strand break hotspots are a better proxy measure of meiotic recombination than crossover rates, which means that the influence of GC-biased gene conversion in mammalian genomes could be underestimated. Furthermore, when analyzing substitution patterns in several mouse lineages, we also found that GC-content evolution is complex and that at least two recent independent shifts in substitution patterns occurred in these lineages. The study of substitution patterns in meiotic recombination hotspots revealed that gene conversion is centered around double strand break hotspots middle points in mouse and around PRDM9 binding sites in human, affecting a region of approximately 1.5 kbp. Finally, we show that hotspots locations are evolving rapidly in mouse, mirroring observations in human.

Zusammenfassung

Genomische Sequenzen können durch ihre prozentuale Zusammensetzung aus den vier Basen Adenin (A), Guanin (G), Thymin (T) und Cytosin (C) beschrieben werden. Diese Zusammensetzung wurde in den vergangenen Jahren ausführlich untersucht. Meist wird sie im GC-Gehalt zusammengefasst, dem Anteil an G- und C-Basen an der Gesamtsequenz. Bei Säugetieren ist interessanterweise zu beobachten, dass der GC-Gehalt entlang der Chromosomen nicht konstant ist, vielmehr bestehen große Variationen. Diese Abweichungen werden als *Isochores* bezeichnet. Sie sind mit einer Reihe genomischer Eigenschaften wie Gen-Dichte und Zeitpunkt der Replikation assoziiert. Es gibt verschiedene Ansätze, die Abweichungen im GC-Gehalt zu erklären. Es hat sich herausgestellt, dass *GC-biased gene conversion* (gBGC) einen großen Einfluss hat. gBGC ist ein neutraler Prozess der folgendermaßen abläuft: Während der meiotischen Rekombination werden Doppelstrangbrüche mittels *gene conversion* repariert, d.h. es wird ein Genfragment in die jeweils andere Sequenz des Doppelstrangs kopiert. Bei diesem Schritt kann es zu Fehlpaarungen kommen, deren Reparatur überdurchschnittlich häufig mit G- und C-Basen erfolgt. Das führt dazu, dass bevorzugt G- und C-Allele fixiert werden im Gegensatz zu A- und T-Allelen. Diese Doktorarbeit beschäftigt sich mit Variationen der Basenzusammensetzung sowie der Evolution des GC-Gehalts in Säugetieren.

Zunächst haben wir Variationen im GC-Gehalt von Säugetieren mit denen von zufällig erzeugten DNA-Sequenzen verglichen und beobachtet, dass die Variationen in diesen Genomen größer sind als bei Zufallssequenzen erwartet. Anschließend untersuchten wir die Genome mehrerer Organismen, wobei wir große Unterschiede zwischen den verschiedenen Gruppen feststellen konnten, z.B. ist die Basenzusammensetzung in den Genomen der Nagetiere wesentlich weniger variabel als die von Primaten.

Mit einem vergleichenden Ansatz untersuchten wir dann Substitutions-Muster und GC-Gehalt Evolution der Genome von Maus und Mensch. Unsere Ergebnisse zeigen, dass gBGC im Genom der Maus von Bedeutung ist und dass die Evolution des GC-Gehaltes in beiden Genomen verschieden ist. Außerdem haben wir geprüft, inwieweit verschiedene genomische Eigenschaften die Substitutions - Muster beeinflussen. Wir haben herausgefunden, dass die Substitutionsraten A oder T $\rightarrow$ G oder C im menschlichen Genom hauptsächlich durch die Meiose (mittels gBGC) beeinflusst werden, während diese Substitutionsraten sich im Genom der Maus am besten durch den CpG Dinukleotid-Gehalt vorhersagen lassen. Daraus schließen wir, dass gBGC im Genom der Maus zwar aktiv, aber schwach ist und dass die Einflüsse auf die Substitutions-Muster in den Genomen von Maus und Mensch verschieden sind.

Kürzlich wurde entdeckt, dass das *Prdm9*-Gen die meiotische Rekombination in Säugetieren kontrolliert und das Bindungsmotiv in den Rekombinations-Hotspots im menschlichen Genom wurde ermittelt. Außerdem wurden die Hotspots für Doppelstrangbrüche im Genom der Maus publiziert. Das zusammen ermöglichte unsere detaillierte Studie über den Einfluss von meiotischer Rekombination auf Substitutions-Muster und die Ableitung von Charakteristika meiotischer Rekombinations-Hotspots. Des weiteren wurden die Genome mehrerer Unterarten der Maus publiziert, die wir zur Untersuchung der Substitutions-Muster in kürzeren Zeiträumen und in mehr Mausarten als bisher möglich genutzt haben. Unsere Studie zeigt, dass Hotspots für Doppelstrangbrüche meiotische Rekombination besser vorhersagen als Crossover-Raten. Daraus schlussfolgern wir, dass der Einfluss von gBGC in Säugetiergenomen unterschätzt sein könnte. Bei unserer Untersuchung der Substitutions-Muster in verschiedenen Mausarten konnten wir feststellen, dass die Evolution des GC-Gehaltes komplex ist und es in diesen Linien mindestens zwei unabhängige Verschiebungen der Substitutions-Muster gegeben haben muss. Die Studie über die Substitutions-Muster in Hotspots meiotischer Rekombination zeigte, dass *gene conversion* in der Maus um die Mittelpunkte der Hotspots von Doppelstrangbrüchen zentriert ist, während *gene conversion* beim Menschen um die PRMD9 Bindungsstellen, in einer Region von etwa 1,5 kbp, zentriert ist. Abschließend zeigen wir, dass die Positionen der Hotspots in der Maus, ebenso wie im Menschen, schnell evolvieren.

Curriculum vitæ

For reasons of data protection, the Curriculum vitæ is not published in the online version

Erklärung zur Urheberschaft

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Yves Clément Berlin, im August 2012