Linking the DNA strand asymmetry to the spatio-temporal replication program : from theory to the analysis of genomic and epigenetic data

Antoine Baker

N° national de thèse: 2011ENSL0700

Thèse

en vue de l'obtention du grade de

Docteur de l'École Normale Supérieure de Lyon - Université de Lyon

Spécialité : Physique

Laboratoire Joliot-Curie
École doctorale de Physique et Astrophysique de Lyon

présentée et soutenue publiquement le 08/12/2011
par Antoine Baker

Linking the DNA strand asymmetry to the spatio-temporal replication program: from theory to the analysis of genomic and epigenetic data

Le programme spatio-temporel de réplication de l'ADN et son impact sur l'asymétrie de composition: d'une modélisation théorique à l'analyse de données génomiques et épigénétiques

---

Directeur de thèse :
Alain Arneodo

Après l'avis de :
Peter Arndt
John Bechhoefer

Devant la commission d'examen formée de :
Peter Arndt, Rapporteur
Alain Arneodo, Directeur
John Bechhoefer, Rapporteur
David Bensimon, Membre
Peter Cook, Membre
Olivier Hyrien, Président
Claude Thermes, Membre

Remerciements

Je remercie tout d'abord Alain Arneodo, qui a dirigé ma thèse et m'a toujours aidé à approfondir et améliorer mes travaux de recherches. Je remercie tout autant Benjamin Audit, auprès duquel j'ai beaucoup appris, et qui a lui aussi directement participé à la direction de ces recherches.

Ce travail de thèse s'inscrit dans une collaboration de longue date entre des physiciens du groupe d'Alain Arneodo, des bio-informaticiens du groupe de Claude Thermes, et des biologistes spécialistes de la réplication du groupe d'Olivier Hyrien. J'ai ainsi eu la chance, au cours de cette thèse, de profiter de la grande clarté et rigueur scientifique de Claude Thermes et d'Olivier Hyrien, sur des domaines, l'évolution moléculaire et la réplication, au sujet desquels j'étais très ignorant avant de commencer cette thèse. Je remercie donc Claude Thermes et son équipe, Yves d'Aubenton-Carafa et Chunlong Chen, ainsi que Olivier Hyrien et son équipe, Guillaume Guilbaud et Aurélien Rappailles, de leur aide et de leurs conseils scientifiques avisés. Je remercie aussi Arach Goldar, qui a rejoint cette collaboration récemment.

Je suis très reconnaissant à Peter Arndt, John Bechhoefer, David Bensimon et Peter Cook de leur intérêt pour ce travail de thèse, et d'avoir accepté de lire et d'évaluer mon manuscrit.

Je remercie tout les membres du Laboratoire Joliot-Curie que j'ai côtoyé ces trois dernières années, et particulièrement les doctorants passés et futurs de l'équipe : Julien, Lamia, Guillaume, et Hanna. Au sein du Laboratoire, je remercie aussi Cédric Vaillant, Benoit Moindrot et Fabien Mongelard, avec qui j'ai eu la chance de collaborer.

Enfin je remercie ma famille pour son soutien tout au long de ma thèse.

Résumé

Deux processus majeures de la vie cellulaire, la transcription et la réplication, nécessitent l'ouverture de la double hélice d'ADN et agissent différemment sur les deux brins, ce qui génère des taux de mutation différents (asymétrie de mutation), et aboutit à des compositions en nucléotides différentes des deux brins (asymétrie de composition). Nous nous proposons de modéliser le programme spatio-temporel de réplication et son impact sur l'évolution des séquences d'ADN. Dans le génome humain, nous montrons que les asymétries de composition et de mutation peuvent être décomposées en deux contributions, l'une associée à la transcription et l'autre à la réplication. Celle associée à la réplication est proportionnelle à la polarité des fourches de réplication, elle-même proportionnelle à la dérivée du "timing" de réplication. La polarité des fourches de réplication délimite, le long des chromosomes humains, des domaines de réplication longs de plusieurs Mpb où le timing de réplication a une forme de U. Ces domaines de réplication sont également observés dans la lignée germinale, où ils sont révélés par une asymétrie de composition en forme de N, indiquant la conservation de ce programme de réplication sur plusieurs centaines de millions d'années. Les bords de ces domaines de réplication sont constituées d'euchromatine, permissive à la transcription et à l'initiation de la réplication. L'analyse de données d'intéraction à longue portée de la chromatine suggère que ces domaines correspondent à des unités structurelles de la chromatine, au coeur d'une organisation hautement parallélisée de la réplication dans le génome humain.

Abstract

Two key cellular processes, namely transcription and replication, require the opening of the DNA double helix and act differently on the two DNA strands, generating different mutational patterns (mutational asymmetry) that may result, after long evolutionary time, in different nucleotide compositions on the two DNA strands (compositional asymmetry). Here, we propose to model the spatio-temporal program of DNA replication and its impact on the DNA sequence evolution. The mutational and compositional asymmetries observed in the human genome are shown to decompose into transcription- and replication-associated components. The replication-associated asymmetry is related to the replication fork polarity, which is also shown to be proportional to the derivative of the mean replication timing. The large-scale variation of the replication fork polarity delineate Mbp scale replication domains where the replication timing is shaped as a U. Such replication domains are also observed in the germline, where they are revealed by a N-shaped compositional asymmetry, which indicates the conservation of this replication program over several hundred million years. The replication domains borders are enriched in open chromatin markers, and correspond to regions permissive to transcription and replication initiation. The analysis of chromatin interaction data suggests that these replication domains correspond to self-interacting chromatin structural units, at the heart of a highly parallelized organization of the replication program in the human genome.

Preface

DNA replication, the basis of genetic inheritance, is of fundamental importance to the cellular life: when the cell fails to regulate its replication program, it strongly affects the genome integrity, which can lead to cell death or cancer. The spatio-temporal replication program, in other words where and when replication initiates and how replication forks propagate, raises several acute questions in today cell biology. How is regulated the spatio-temporal replication program? How much does it change from one cell cycle to another? Is it encoded in the DNA sequence or specified by epigenetic mechanisms? How does it relate with the chromatin tertiary structure?

This thesis will focus on a quite unexpected aspect of the spatio-temporal replication program: how it affects the genome evolution. More precisely we will describe how the mutations generated by the replication process may, during the course of evolution, give rise to a compositional asymmetry, that is a difference of nucleotide composition on the two DNA strands. Indeed, DNA replication is fundamentally a strand-asymmetric process (Fig. 1): the leading strand is replicated continuously while the lagging strand is replicated discontinuously by means of small Okazaki fragments. It has long been proposed that the leading and lagging strands could undergo different mutational patterns, which may in turn generate a compositional asymmetry after long evolutionary time.

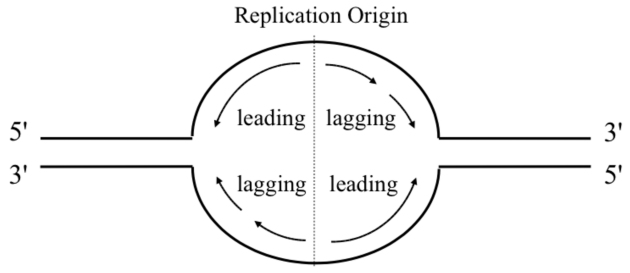

Figure 1: **Replication is a strand-asymmetric process.**

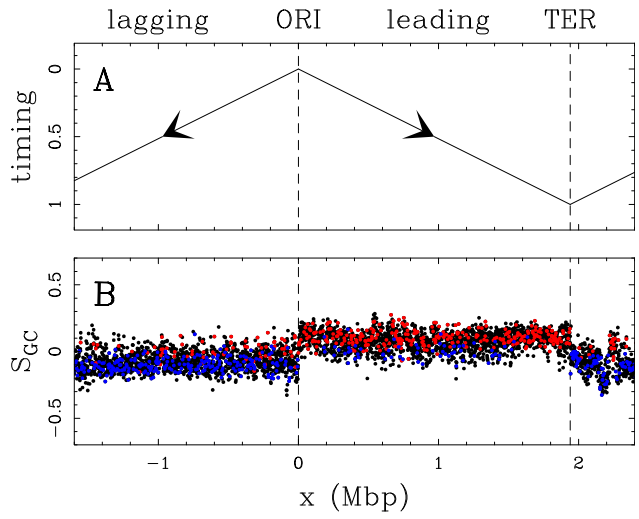

Figure 2: **Comparing GC skew $S_{GC} = \frac{G-C}{G+C}$ and replication timing in Bacillus subtilis genome.** (A) Schematic representation of the replicon model: divergent bidirectional progression of the two replication forks from the replication origin (ORI) to the replication terminus (TER). The replication timing is indicated from early, 0 to late, 1. (B) S_{GC} calculated in 1 kbp windows along the genomic sequence of *Bacillus subtilis*. Black points correspond to intergenic regions, red (resp. blue) points correspond to (+) (resp. (−)) genes, which coding sequences are on the published (resp. complementary) strand.

Compositional asymmetry in bacteria

A clear relationship between replication and compositional asymmetry was first established in prokaryotic genomes by Lobry (1996a). In bacteria, the spatio-temporal replication program is particularly simple. Most prokaryotes follow the replicon model depicted in Fig. 2A: the replication origin is defined by a consensus sequence, replication therefore always initiates at the same genomic locus (ORI), two divergent forks then replicate the DNA until they meet at the replication terminus (TER). As shown in Fig. 2B for *Bacillus subtilis*, many prokaryotic genomes are divided into two halves: one presents an excess of guanine over cytosine, and the other one, on the opposite, an excess of cytosine over guanine. The GC skew, defined as $S_{GC} = \frac{G-C}{G+C}$, is thus positive on one half of the genome and negative on the other. Remarkably, the GC skew profile is tightly related to the spatio-temporal replication program: the leading strand has positive GC skew whereas the lagging strand has negative GC skew.

Compositional asymmetry in the human genome

By contrast, the spatio-temporal replication program in eukaryotes is much more complex. Several initiation sites are used each cell cycle, and they fire at different times during the S phase. Furthermore, the genomic positions and firing times of the initiation sites change from one cell cycle to another. Does the relationship observed

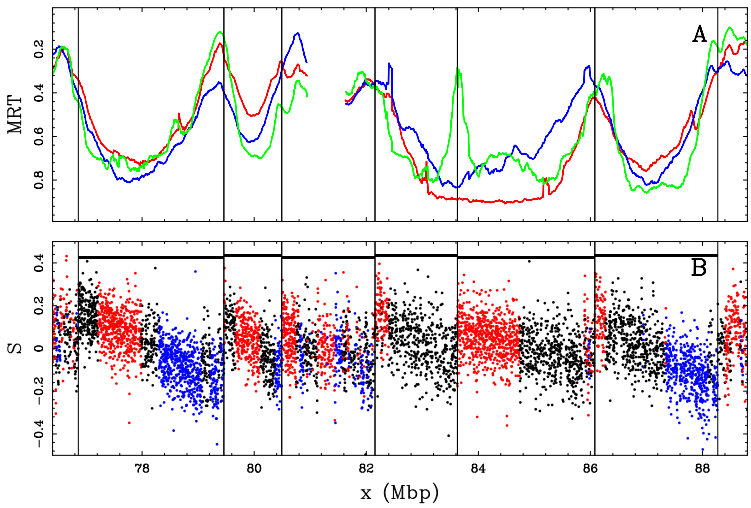

Figure 3: **Comparing compositional skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ and mean replication timing (MRT) in the human genome**. (A) MRT profiles along a 11.4 Mbp long fragment of human chromosome 10, from early, 0 to late, 1 for BG02 embryonic stem cell (green), K562 erythroid (red) and GM06990 lymphoblastoid (blue) cell lines. Replication timing data was retrieved from (Hansen et al. 2010). (B) S calculated in 1 kbp windows of repeat-masked sequence. The colors correspond to intergenic (black), (+) genes (red) and (−) genes (blue). Six skew N-domains (horizontal black bars) were detected in this genomic region.

between the compositional asymmetry and the replication program in bacteria generalize to eukaryotic genomes?

We observe in Fig. 3 a clear relationship between the compositional asymmetry and the replication timing in the human genome: a N-shaped compositional skew $S = \frac{G-C}{G+C} + \frac{T-A}{T+A}$ profile remarkably corresponds to a U-shaped replication timing profile. Previous work has led to the objective delineation of N-shaped skew domains in the human genome (Touchon et al. 2005; Brodie of Brodie et al. 2005; Nicolay 2006). Those genomic domains, that were called N-domains, were shown to exhibit a very peculiar gene organization and chromatin state (Huvet et al. 2007; Audit et al. 2009; Zaghloul 2009). Based on the analogy with the bacterial case (the upward jump of the GC skew colocalizes with the ORI in Fig. 2) , the N-domains borders (upward jumps of the skew) were proposed to be replication origins, evolutionary conserved and active in the germline (Touchon et al. 2005; Brodie of Brodie et al. 2005). However in our current perspective we know that the skew profile observed in N-domains is not a trivial extension of the replicon model in bacteria, with replication origins located at the N-domains borders. For instance, as N-domains have $\sim$ 1 Mbp characteristic size, this model would imply that large $\sim$ 1 Mbp replicons are produced in the 30% of the human genome covered by N-domains, in conflict with the typical observed replicon size ($\sim$ 100 kbp).

In this thesis, we will try to answer the following questions. How do the mutational asymmetries generated by the replication process relate to the replication program? How does a genome submitted to a mutational asymmetry evolve? On which time scales were the skew N-domains generated? Which characteristics of the replication program do explain the N-shaped skew profile? Why do skew N-domains coincide with U-shaped replication timing domains? As compositional asymmetry is also associated to transcription in the human genome (Touchon et al. 2003), is it possible to disentangle in the skew profile the contributions associated to transcription and replication?

Replication fork polarity

In eukaryotes, due to the multiple initiation sites and the inherent stochasticity of the replication program, a locus can be replicated by a right-moving fork in some cell cycles and by a left-moving fork in other cell cycles. Therefore, in contrast to bacteria, genomic regions cannot be unambiguously assigned as leading or lagging strands. To study replication-associated strand asymmetry in eukaryotes we need to consider the replication fork polarity, defined as the difference of proportions of right-moving and left-moving forks replicating a locus. In this thesis, we demonstrate that **the compositional asymmetry and the replication timing are both related to the replication fork polarity**. We further argue that it provides an unifying mechanism which explains why the replicon model in bacteria (Fig. 2A) results in the crenel-like skew profile (Fig. 2B), and why the U-shaped replication timing profile observed in the human genome (Fig. 3A) results in the N-shaped skew profile (Fig. 3B).

Outline

The manuscript is organized in five Chapters. Chapters I and II deal with the mathematical aspects of the DNA composition evolution and of the spatio-temporal DNA replication program. Our modeling will lead to three main theoretical propositions (i) the compositional asymmetry decomposes into transcription and replication associated contributions, (ii) the compositional asymmetry generated by the replication process is proportional to the replication fork polarity, and (iii) the replication fork polarity is proportional to the derivative of the mean replication timing. In Chapter III, the analysis of substitution rates in the human genome will support the decomposition of the compositional asymmetry into transcription and replication associated components. Chapter IV focuses on the detection of N-domains in the human and mouse genomes, with the specific goal to extract in the skew profile the contributions associated to replication and transcription. In Chapter V, we will objectively delineate U-shaped replication timing domains in several human cell lines. We will also study their properties in term of chromatin state and long-range chromatin interactions. Finally, we will present the conclusions and perspectives of this work.

Contents

Chapter I

Evolution of compositional skews: a theoretical approach

Replication and transcription are asymmetric processes with respect to the two DNA strands. How do the asymmetries due to replication and transcription reflect on the substitution rates and, after long evolutionary time, on the composition? These questions received a lot of attention experimentally, but not so much formally. In this chapter, we present a mathematical formalism to describe the establishment and maintenance of the strand asymmetry. Only Sections I.3 and I.4 are mathematically involved. As some readers may wish to skip the mathematical part, we sum up at the end of the chapter the key arguments and the main results.

I.1 Introduction: the two DNA strands, transcription and replication

We briefly recall a fundamental property of DNA (Alberts et al. 2008), namely the double-helix structure and the base-pairing of the two DNA strands. We expect approximately the same substitution rates and the same composition on the two DNA strands (strand symmetry) (Sueoka 1995). However two key processes of the cell, transcription and replication, are generally thought to generate strand asymmetries (Francino and Ochman 1997; Frank and Lobry 1999). In order to model replication and transcription related strand asymmetries, we introduce the gene orientation and the replication fork polarity.

I.1.1 Double-helix structure of DNA

Generalities

A deoxyribonucleic acid (DNA) molecule consists of two polynucleotides chains, called DNA strands. The two DNA strands are held together by hydrogen bonds, and the resulting double-stranded DNA has a double-helix structure (Fig. 1). The nucleotide chain is composed of a backbone of alternating sugars and phosphates,

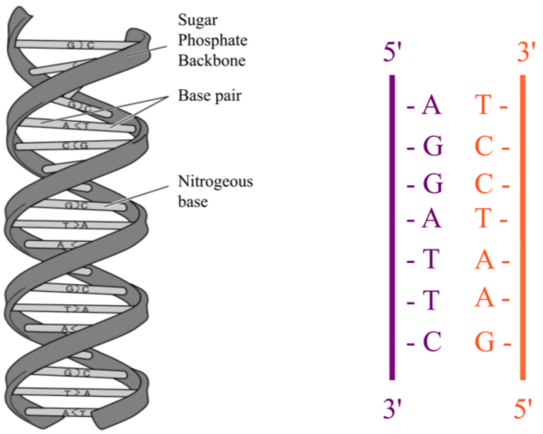

Figure 1: **Schematic view of the DNA double helix structure.** Left image courtesy of the National Human Genome Research Institute (www.genome.gov).

and the nucleotides differ by their base: Thymine, Adenine, Guanine or Cytosine, generally abbreviated as T, A, G, C. A guanine G on one strand is always linked to a cytosine C on the other strand by three hydrogen bonds (strong coupling). A thymine T on one strand is always linked to an adenine A on the other strand by two hydrogen bonds (weak coupling). The polynucleotide chain has also an orientation. A sugar in the backbone has its 5' phosphate group attached to the 3' hydroxyl group of the next sugar. This gives a polarity to the DNA strand, one end (referred to as the 5' end) has a 5' phosphate group, and the other (referred to as the 3' end) a 3' hydroxyl group. In the DNA double-helix, the polarities of the two strands run anti-parallel to each other (Fig. 1). The nucleotide sequence of one DNA strand is conventionally read in the $5' \to 3'$ direction. The polarity of the DNA strand has great biological importance. For instance the DNA polymerase always synthesizes the newly replicated strand in the $5' \to 3'$ direction. Similarly, the RNA polymerase always synthesizes the messenger RNA in the $5' \to 3'$ direction.

Reverse complementarity

Due to base-pairing and anti-parallel orientation, the nucleotide sequence on the two DNA strands are related by **reverse complementarity**. The nucleotide sequence on one strand is entirely determined by the nucleotide sequence of the other strand, *e.g.* on Fig. 1 the sequence of the purple strand reads $5' - AGGATTC - 3'$, and the sequence on the orange complementary strand reads $3' - TCCTAAG - 5'$ *i.e.* $5' - GAATCCT - 3'$. To determine the DNA sequence, we have to choose arbitrary one of the two strands, for instance the published strand. The sequence is read in the $5' \to 3'$ direction on this **reference strand**. The sequence on the **complementary strand** is then given by reverse complementarity.

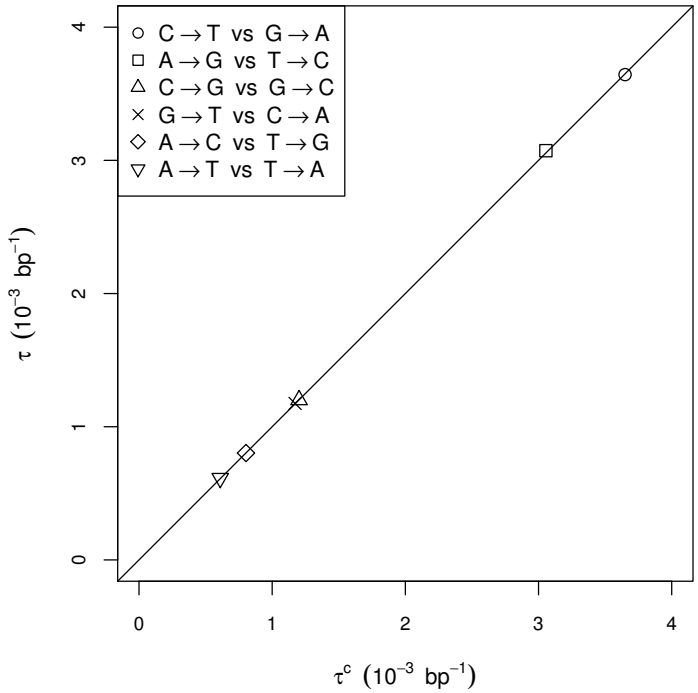

Figure 2: **PR1 in the human genome.** Genome wide substitution rate τ versus the reverse complementary substitution rate τ^c.

I.1.2 Strand symmetry

Parity rule type 1

We expect that the two DNA strands experience on average the same mutational and repair mechanisms. The substitution rates should therefore be approximately equal on the two DNA strands. A substitution (*e.g.* $G \to T$) on one strand always corresponds to the reverse complementarity substitution (*e.g.* $(G \to T)^c = C \to A$) on the complementary strand. Therefore we expect complementary substitutions to have approximately equal rates, when computed on a given strand (*e.g.* $G \to T \sim C \to A$). This symmetry law is known as Parity rule type 1 (PR1) (Sueoka 1995). As shown in Fig. 2 for the human genome, PR1 is very well verified at the genome scale. Although substitution rates can vary over a large range of values (the transition $C \to T$ is three fold higher than the transversion $C \to G$ in Fig. 2), reverse complementary substitution rates are nearly equal.

Parity rule type 2

If the substitution rates are nearly equal on the two DNA strands, we expect in turn the compositions of the two DNA strands to be nearly equal. Therefore we expect complementary nucleotides to have approximately equal frequencies, when

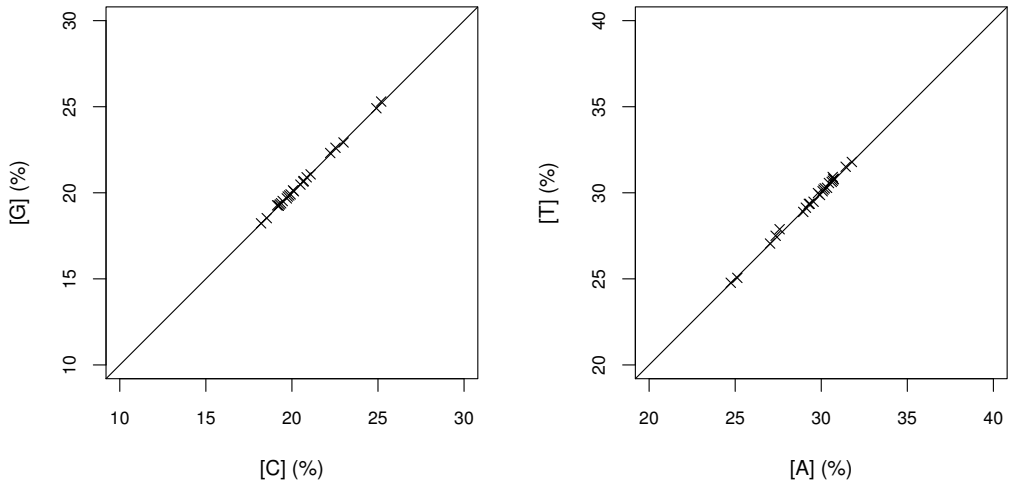

Figure 3: **PR2 in the human genome.** $[G]$ versus $[C]$ and $[T]$ versus $[A]$ for the 22 human autosomes. Reproduced from (Sueoka 1995; Lobry 1995), where PR2 plots are shown for organims over the whole life tree.

computed on a given strand: $[G] \sim [C]$ and $[T] \sim [A]$. This second symmetry law is known as Parity rule type 2 (PR2) (Rudner et al. 1968; Sueoka 1995). As shown in Fig. 3 for the human genome, PR2 is very well verified at the chromosomal scale. Although the GC content ($\theta_{GC} = [G] + [C]$) can vary over a large range of values (from 36% to 50% in Fig. 3), the frequencies of complementary nucleotides are nearly equal. PR2 formally derives from PR1: under symmetrical substitution rates (PR1), the DNA composition should verify PR2 (Lobry 1995; Lobry and Lobry 1999).

☛ *PR1 and PR2 are approximate symmetries, they are well verified on the chromosomal scale, but we can observe systematic deviations at finer scales. The breaking of PR1 and PR2 symmetries, i.e. strand asymmetry, has generally been associated to two key processes of the cell, namely transcription and replication.*

I.1.3 Transcription

Transcription is a strand asymmetric process

During the transcription of a gene (Fig. 4), the RNA polymerase synthesizes a messenger RNA similar to the coding sequence (with the replacement of thymines T by uracils U). The RNA polymerase synthesizes the messenger RNA in the $5' \to 3'$ direction by base-pairing using the other strand as a template. The RNA polymerase therefore progresses in the $3' \to 5'$ direction on the template (or transcribed) strand (Alberts et al. 2008). The coding strand and the transcribed strand could undergo different mutational and repair events that generate strand asymmetry (Francino and Ochman 1997; Frank and Lobry 1999; Francino and Ochman 2001; Green et al.

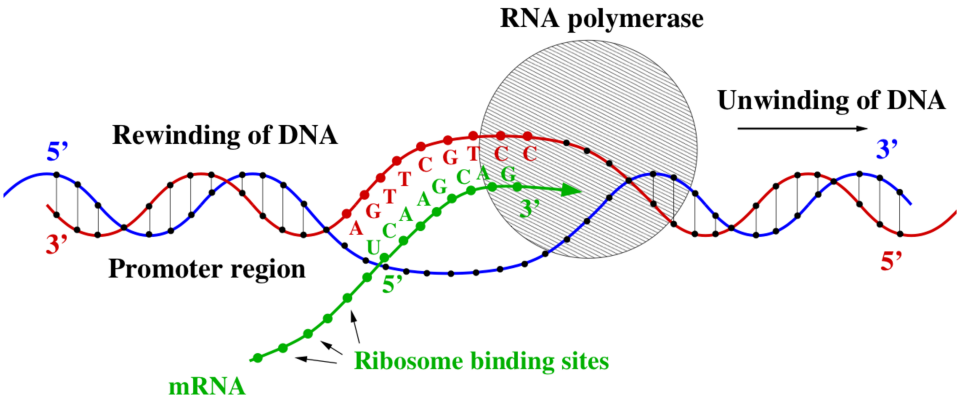

Figure 4: **Schematic view of transcription.**

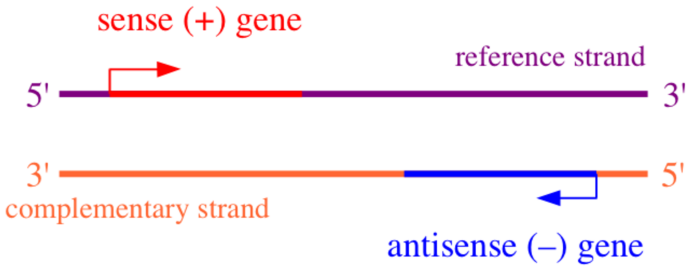

Figure 5: **Definition of sense** (+) **and antisense** (−) **gene.**

2003; Polak and Arndt 2008; Mugal et al. 2009) . During transcription, the coding strand is transiently in single-stranded state (ssDNA), while the transcribed strand is protected by the RNA polymerase. The coding strand is possibly more exposed to mutagenic lesions (Beletskii and Bhagwat 1996, 1998; Francino and Ochman 1997, 2001; Polak and Arndt 2008; Mugal et al. 2009) than the transcribed strand. It has also been proposed that repair mechanisms (Francino et al. 1996; Francino and Ochman 1997; Green et al. 2003; Mugal et al. 2009) could generate strand asymmetries. A mechanism known as transcription-coupled repair (TCR) (Alberts et al. 2008), associated with the passage of the RNA polymerase, preferentially repairs towards the coding strand (Svejstrup 2002). Strand asymmetry associated to transcription has been observed across the whole life tree: in several bacterial strains (Francino et al. 1996; Francino and Ochman 2001), in human (Green et al. 2003; Touchon et al. 2003; Polak and Arndt 2008; Mugal et al. 2009) and in many other eukaryotes (Touchon et al. 2004).

Definition of gene orientation

Transcription generates strand asymmetries by discriminating a coding and a transcribed strand. We further need to define the strand asymmetry as seen by the reference strand, where the DNA sequence is computed. A gene is defined as **sense** $(+)$ **gene** if its coding sequence is on the reference strand, and as **antisense** $(-)$ **gene** if its coding sequence is on the complementary strand (Fig. 5). For a sense $(+)$ gene the reference strand is the coding strand and the complementary strand is the transcribed strand. For an antisense $(-)$ gene we have the opposite situation. Therefore the **gene orientation** $(\pm)$ is a crucial parameter of transcription-associated strand asymmetry. Another crucial parameter is the **transcription rate** (hereafter noted as α), which reflects how many times the gene has been transcribed during a cell cycle. The more the gene is transcribed, the stronger we expect the strand asymmetries to be.

☛ *Gene orientation and transcription rate are the natural parameters to describe the strand asymmetry due to transcription.*

I.1.4 Replication

Replication is a strand asymmetric process

When a cell divides, the genome of the mother cell is duplicated and transmitted to the two daughter cells. The DNA replication is semi-conservative (Fig. 6): each daughter cell inherits a DNA strand of the mother cell, which serves as a template for the DNA polymerase to synthesize the complementary strand (Alberts et al. 2008). During the S-phase (phase of the cell cycle where the genome is duplicated), replication initiates at loci called **replication origins**. At a replication origin (Fig. 6) the DNA double helix is opened, and two divergent replication forks replicate the DNA on each side of the replication origin, creating a "replication bubble". Each replication fork is composed of two DNA polymerases that replicate separately the two parental strands. The DNA polymerases always synthesize the new strand in the $5' \to 3'$ direction progressing on the parental strand in the $3' \to 5'$ direction. Due to the anti-parallel polarities of the parental strands, one strand is synthesized continuously (the **leading strand**) and the other discontinuously (the **lagging strand**). The parental strand oriented in the $3' \to 5'$ direction as seen by the replication fork (the leading strand template) is replicated continuously by the DNA polymerase, producing continuously the synthesized leading strand in the $5' \to 3'$ direction. On the parental strand oriented in the $5' \to 3'$ direction as seen by the replication fork (the lagging strand template), the DNA polymerase synthesizes discontinuously small nascent strands, called Okazaki fragments, in the $5' \to 3'$ direction, progressing in the $3' \to 5'$ direction on the parental strand, opposite to the global replication fork movement. The leading/lagging strands usually refer to the newly synthesized leading/lagging strands. But they can also refer to the parental strands, in that case the leading (resp. lagging) strand is understood as the lagging (resp. leading) strand template.

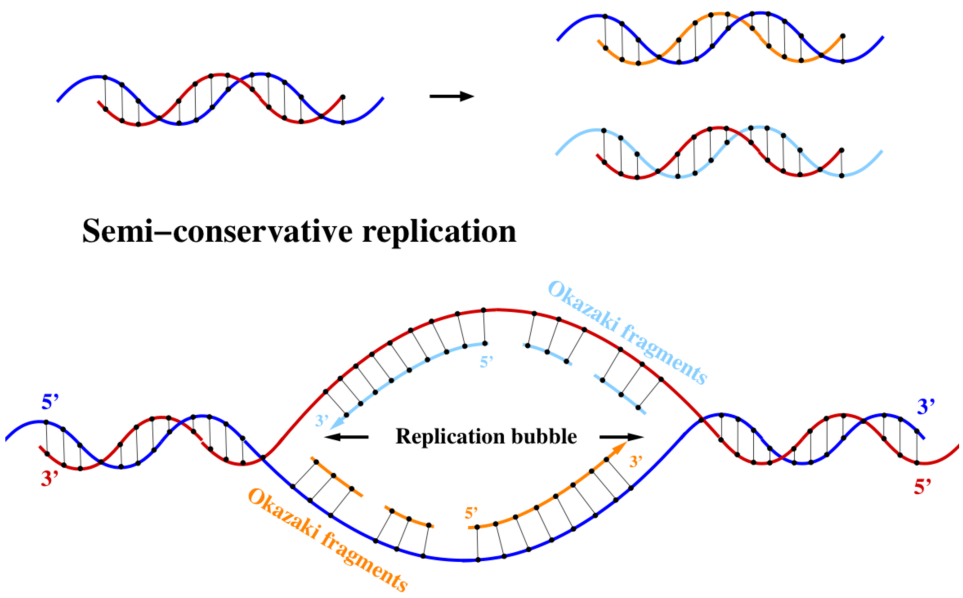

Figure 6: **Schematic view of replication.**

Replication could induce strand asymmetries by several means (Francino and Ochman 1997; Frank and Lobry 1999; Rocha et al. 1999, 2006; Polak and Arndt 2009). For instance the leading strand, when it serves as a template for the lagging synthesis of the complementary strand, is transiently in ssDNA, where it could be more exposed to mutagenic lesions (Francino and Ochman 1997; Frank and Lobry 1999). In eukaryotes, the leading and lagging strands are presumably synthesized by two distinct DNA polymerases (Kunkel and Burgers 2008). Strand asymmetries could result from the different error spectra of the two DNA polymerases (Polak and Arndt 2009). Strand asymmetry associated to replication was observed across the whole life tree: in several bacterial strains (Lobry 1996a,b; Mrázek and Karlin 1998; Rocha et al. 1999; Tillier and Collins 2000; Rocha et al. 2006), in viruses (Mrázek and Karlin 1998; Grigoriev 1999), in yeast (Gierlik et al. 2000), in mitochondria (Reyes et al. 1998), and in human (Brodie of Brodie et al. 2005; Touchon et al. 2005; Polak and Arndt 2009; Chen et al. 2011).

Definition of replication fork polarity

Replication generates strand asymmetries by discriminating a leading and a lagging strand. We further need to define the strand asymmetry as seen by the reference strand, where the DNA sequence is computed. A replication fork is defined as **sense** $(+)$ **fork** if it "moves" in the $5' \to 3'$ direction seen from the reference strand, and as **antisense** $(-)$ **fork** if it "moves" in the opposite $3' \to 5'$ direction (Fig. 7). In other words, a sense $(+)$ fork comes from a replication origin that fired upstream ($5'$ direction of the reference strand), whereas an antisense $(-)$ fork comes from a

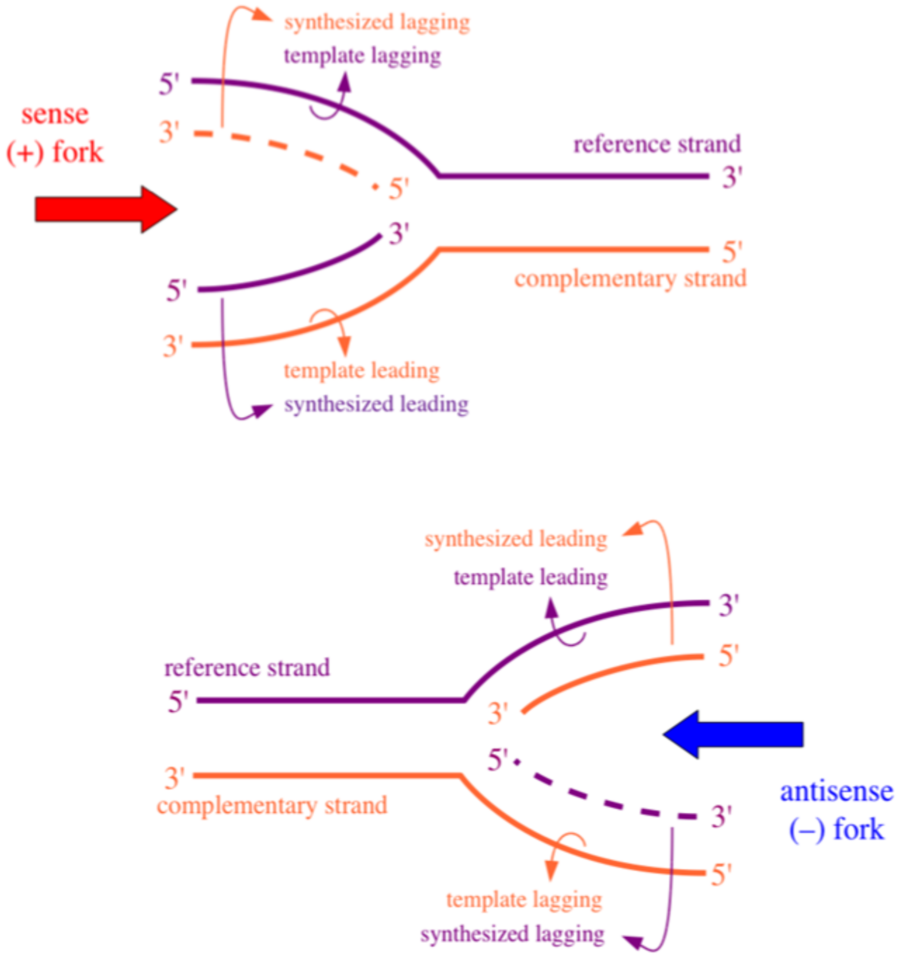

Figure 7: **Definition of sense (+) and antisense (−) replication forks.**

replication origin that fired downstream ($3'$ direction of the reference strand). For a sense (+) fork (Fig. 7), the reference strand is either the synthesized leading strand or the lagging strand template, whereas the complementary strand is either the synthesized lagging strand or the leading strand template. For an antisense (−) fork (Fig. 7) we have the opposite situation. During the S-phase, each locus is replicated once and only once, and it is either replicated by a sense or an antisense fork. Over cell cycles, the locus x will be replicated by a proportion $p_{(\pm)}(x)$ of ($\pm$) forks. As the proportions of sense and antisense forks always sum up to one, only the difference of proportions is relevant. This difference defines the **replication fork polarity**:

$$p(x) = p_{(+)}(x) - p_{(-)}(x). \tag{1}$$

We define the replication fork polarity for a locus x, but it can be equally defined for a genomic region. When the replication fork polarity $p = +1$ (resp. $p = -1$), the genomic region only undergoes leading (resp. lagging) strand synthesis, hence the strand asymmetry due to replication is maximal in such regions. Between these two extreme cases, the replication fork polarity can take values in the whole interval $[-1, 1]$. When the replication fork polarity $p = 0$, there is as many leading and lagging strand synthesis, and consequently no strand asymmetry due to replication in these regions.

☛ *Replication fork polarity is the natural parameter to describe the strand asymmetry due to replication.*

I.1.5 From mutations to substitutions

Mutations, if they occur in germ line cells, can be transmitted from an individual to its descendants. These mutations, at the population level, can have their frequencies increased or decreased over time, and ultimately reach fixation (when the mutation is present in all individuals of the species) or disappear. A mutation that reaches fixation is called a **substitution**. Natural selection and random genetic drift are two competing forces that determine fixation or disappearance of a mutation (Graur and Li 1999). Random genetic drift corresponds to the stochastic variation of a mutation frequency, due to the random sampling of alleles (Graur and Li 1999). Under random genetic drift alone, the fixation probability is the same for all mutations, and the substitution rate observed at the population level directly reflects the mutation rate at the individual level (Kimura 1968; Graur and Li 1999) (neutral molecular evolution). On the opposite, natural selection affects the probability of fixation of a mutation, the fixation probability of an advantageous mutation is increased (positive selection), on the opposite the fixation probability of a deleterious mutation is decreased (purifying selection) (Graur and Li 1999). The fixation probability can also be affected by neutral processes such as biased gene conversion (Galtier et al. 2001; Duret and Arndt 2008; Duret and Galtier 2009): gene conversion, a common event during meiosis recombination, is biased towards the fixation of GC rich alleles.

When selection plays no role at a locus, the site is said to evolve neutrally.

☛ *In the following discussion on substitutional asymmetry, we will only consider neutral sites.*

I.2 Substitutional asymmetry

As transcription and replication are strand asymmetric processes, the two DNA strands could experience different mutational events, which would result in different substitution rates. We propose here to model the impact of replication fork polarity, gene orientation and transcription rate on substitution rates. The model we propose is the simplest model that takes into account the basic symmetries of the problem. We show that most molecular mechanims proposed so far to explain strand asymmetry are particular cases of this minimal model.

I.2.1 Minimal model

The model has to respect strand exchange symmetry

The arbitrary choice of the reference strand leads, as we shall see, to strong symmetrical properties of the substitution rates. A substitution on a given strand, *e.g.* $G \to T$, always corresponds to the reverse complementary substitution on the other strand, *e.g.* $(G \to T)^c = C \to A$. The substitution rates are computed on a given reference strand, and the definitions of gene orientation and replication fork polarity are made relatively to this reference strand. Note that calling a gene (or a replication fork) sense or antisense is just related to our arbitrary choice of the reference strand. An antisense gene (or fork) for the reference strand is a sense gene (or fork) for the complementary strand and vice-versa. Let us consider a substitution rate τ (*e.g.* $T \to C$) for a locus located in a sense gene and having a replication fork polarity p. Computed on the complementary strand, the reverse complementary substitution rate τ^c (*e.g.* $A \to G$) has the same value, and seen from the complementary strand, the locus is located in an antisense gene and it has a replication fork polarity $-p$. Therefore substitution rates have to respect what we call **strand exchange symmetry**:

$$\tau[\xi, p, \alpha, (+)] = \tau^c[\xi, -p, \alpha, (-)], \tag{2}$$

where the "ξ dependence" is here to remind us that substitution rates depend on many other variables, but that these variables do not discriminate the two strands (*e.g.* replication timing, distance to telomeres, recombination rate). Indeed, it is much more convenient to study strand asymmetry using the symmetrical part $\tau^s = [\tau + \tau^c]/2$ and asymmetrical part $\tau^a = [\tau - \tau^c]/2$ of substitution rates. The symmetrical part corresponds to the average of a substitution rate on the two DNA strands, while the asymmetrical part measures the **substitutional asymmetry**

between the two DNA strands. The symmetrical part is invariant under strand exchange symmetry whereas the asymmetrical part changes sign:

$$\tau^s[\xi, p, \alpha, (+)] = \tau^s[\xi, -p, \alpha, (-)], \tag{3}$$
$$\tau^a[\xi, p, \alpha, (+)] = -\tau^a[\xi, -p, \alpha, (-)]. \tag{4}$$

We propose here to take the simplest model, with the minimal set of parameters, that respects strand exchange symmetry (Eq. (2)). More sophisticated models could be proposed, but they would still have to verify the general relation Eq. (2).

Substitution rates in genic and intergenic regions

Hereafter, we will forget about the "ξ dependence" to focus only on the effect of gene orientation ($\pm$), transcription rate α and replication fork polarity p on substitution rates. Therefore these rates have to be understood either as secretly depending on the ξ parameters, or as averaged over the ξ parameters. In our minimal model, substitution rates in intergenic regions are given by:

$$\tau_{\text{intergenic}}[p] = \tau_0^s + p_{(+)}\tau_R + p_{(-)}\tau_R^c. \tag{5}$$

The different coefficients can be interpreted as follows. Mutational events associated with the passage of a sense (+) replication fork give rise to a substitution rate τ_R. Due to strand exchange symmetry (Eq. (2)), the passage of an antisense (−) fork contributes by the reverse complementary substitution rate τ_R^c. We assume that mutational events not associated to the passage of replication forks affect equally the two DNA strands. Thus they give rise to a symmetrical substitution rate τ_0^s that is equal on the two DNA strands, in other words τ_0^s satisfies PR1. In genic regions, we propose to model the net effect of transcription by:

$$\tau_{\text{genic }(+)}[p, \alpha] = \tau_{\text{intergenic}}[p] + \tau_T[\alpha], \tag{6}$$
$$\tau_{\text{genic }(-)}[p, \alpha] = \tau_{\text{intergenic}}[p] + \tau_T^c[\alpha]. \tag{7}$$

The reverse complementary coefficient τ_T^c appears in antisense gene due to strand exchange symmetry (Eq. (2)). If $\tau_T[\alpha]$ is interpreted as a substitution rate resulting from additional mutational events associated to transcription, then it has to be positive. If this coefficient also takes into account repair mechanisms associated to transcription, then it can be either way positive or negative. This coefficient should depend on the transcription rate α. We expect the effect of transcription to be stronger if the gene is more transcribed; in other words $\tau_T[\alpha]$ should increase in magnitude with α. For weakly expressed genes ($\alpha \to 0$), we expect to recover the intergenic case ($\tau_T[\alpha] \to 0$). The main assumption of our model is that **transcription and replication contribute separately to substitution rates**. In this model we also **neglect non-coding transcription**. Recent studies have shown that most genomic DNA, including intergenic regions, is transcribed (The ENCODE Project Consortium 2007), producing non-coding transcripts (Cheng et al. 2005; Core et al.

2008; He et al. 2008; Preker et al. 2008; Seila et al. 2008). Non-coding transcripts could generate strand asymmetries in intergenic regions not associated to replication (Necsulea et al. 2009). In our model non-coding transcripts are not taken into account, and we will always assume that substitutional asymmetry in intergenic regions is solely due to replication.

The substitutional asymmetry is decomposed into transcription- and replication-associated components

For the model defined by Eqs. (5) to (7), the symmetrical part of the substitution rates depends neither on the replication fork polarity nor on the gene orientation. It depends only on the transcription rate α:

$$\tau^s[p, \alpha, (\pm)] = \tau_0^s + \tau_R^s + \tau_T^s[\alpha], \tag{8}$$

where $\alpha = 0$ ($\tau_T[0] = 0$) corresponds to the intergenic case. The asymmetrical part depends on the replication fork polarity p, the gene orientation ($\pm$), and the transcription rate α:

$$\boxed{\tau^a[p, \alpha, (\pm)] = p\tau_R^a \pm \tau_T^a[\alpha]}, \tag{9}$$

where $\alpha = 0$ ($\tau_T[0] = 0$) corresponds to the intergenic case.

☛ *In our minimal model, substitutional asymmetry can be decomposed into transcription and replication associated components. The replication-associated substitutional asymmetry is proportional to the replication fork polarity, the transcription-associated one increases in magnitude with transcription rate and changes sign with gene orientation.*

I.2.2 Examples of molecular mechanisms

To illustrate the relevance of our minimal model, in this subsection we review more concrete biological processes. These molecular mechanisms are of interest to us, as they were proposed to explain strand asymmetry in the human genome. We will show that they all reduce to the decomposition of the substitution rates given in Eqs. (8) and (9), except for the last example where a new term will appear in the symmetrical part (Eq. (8)).

Remark. In the following I will confound substitutions with mutations to make the discussion easier to follow. The relationship between substitution and mutation rate is direct (Kimura 1968) if there are no (neutral or selective) fixation bias. To my knowledge, no concrete neutral fixation bias were proposed to generate strand asymmetry, but a fixation bias could surely modulate the strength of the substitutional asymmetry.

Misinsertions induced by the DNA polymerases

In eukaryotes, the leading and lagging strands are presumably synthesized by two distinct DNA polymerases. This is demonstrated at least in yeast, where pol ϵ is used for the leading strand synthesis and pol δ for the lagging strand synthesis (Kunkel and Burgers 2008). In the human genome, the substitutional asymmetries associated to replication were proposed to result from the different error spectra of the two DNA polymerases (Polak and Arndt 2009; Chen et al. 2011). The error spectra of the human DNA polymerases are currently unknown, it is therefore difficult to infer the sign of the asymmetries and thus to check this hypothesis (Polak and Arndt 2009).

Modelling the misinsertions. For clarity let us call pol ϵ (resp. pol δ) the leading (resp. lagging) polymerase as in yeast. For a nucleotide $i \in \{T, A, G, C\}$ we denote by $i^c \in \{A, T, C, G\}$ the complementary nucleotide. For nucleotides $i, j \in \{T, A, G, C\}$, we denote by Σ_{ji} (resp. Δ_{ji}) the misinsertion rate of a j instead of i, in other words a j misinserted in front of i^c, by the ϵ (resp. δ) polymerase. For simplicity we assume that the mispaired base $j : i^c$ will persist until the next replication round, where the mispaired base has a priori 50% chance to result in the $i \to j$ substitution. For a sense fork, the reference strand is the leading strand (synthesized by pol ϵ), while the complementary strand is the lagging strand (synthesized by pol δ). For a antisense fork the role of the complementary and reference strands are exchanged. Sense and antisense forks contribute to the $i \to j$ substitution by the following pathways:

$$(+) \quad \text{fork} \quad \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{\Sigma_{ji}} \begin{matrix} j \\ i^c \end{matrix} \xrightarrow{\frac{1}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1}{2}} \begin{matrix} i \\ j^c \end{matrix} \xleftarrow{\Delta^c_{ji}} \begin{matrix} i \\ i^c \end{matrix} \tag{10}$$

$$(-) \quad \text{fork} \quad \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{\Delta_{ji}} \begin{matrix} j \\ i^c \end{matrix} \xrightarrow{\frac{1}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1}{2}} \begin{matrix} i \\ j^c \end{matrix} \xleftarrow{\Sigma^c_{ji}} \begin{matrix} i \\ i^c \end{matrix} \tag{11}$$

where the upper strand is the reference strand, $\Delta^c_{ji} = \Delta_{j^c i^c}$ and $\Sigma^c_{ji} = \Sigma_{j^c i^c}$. Therefore the misinsertion process contributes to the $i \to j$ substitution by:

$$p_{(+)} \frac{(\Sigma + \Delta^c)_{ji}}{2} + p_{(-)} \frac{(\Sigma^c + \Delta)_{ji}}{2}. \tag{12}$$

We recover our minimal model Eq. (5) with:

$$\tau_R = \frac{\Sigma + \Delta^c}{2}. \tag{13}$$

The substitutional asymmetry associated to replication is then given by:

$$p\tau^a_R \quad \text{with} \quad \tau^a_R = \frac{\Sigma^a - \Delta^a}{2}, \tag{14}$$

in agreement with Eq. (9).

Cytosine deamination in single-stranded DNA

Cytosine can spontaneously deaminates into uracil. After two replication rounds the uracil can become a thymine ($U : G \to U : A \to T : A$) and the cytosine deamination results in a proper $C \to T$ substitution. The cytosine deamination is much more frequent (140 fold) in single-stranded DNA (ssDNA) than in double-stranded DNA (Frederico et al. 1990). The leading strand, when it serves as a template for the lagging synthesis of the complementary strand, is transiently in ssDNA and could undergo an excess of cytosine deamination. The process was proposed to explain strand asymmetry associated to replication in bacteria (Frank and Lobry 1999), in mitochondria (Reyes et al. 1998), and recently in human (Polak and Arndt 2009; Chen et al. 2011). Similarly the coding strand is transiently in ssDNA during transcription and could undergo an excess of cytosine deamination (Beletskii and Bhagwat 1996, 1998). This process was proposed to explain the strand asymmetry observed in *E. coli* genes (Francino and Ochman 2001), and in human genes (Mugal et al. 2009; Polak and Arndt 2008).

Modelling the cytosine deamination. Let us call $(C \to U)_{ssDNA}$ the rate of cytosine deamination into uracil in ssDNA. For simplicity we assume that the uracil has a 25% chance to result, after two replication rounds ($U : G \overset{1/2}{\to} U : A \overset{1/2}{\to} T : A$), in a thymine. We called t_R (resp. t_T) the time lapse in ssDNA for the leading (resp. coding) strand during replication (resp. transcription). For a gene with a transcription rate α, the time lapse in ssDNA over the cell cycle is thus αt_T. The cytosine deamination in ssDNA contributes to the $C \to T$ substitution rate by:

$$p_{(+)}\frac{t_R}{4}(C \to U)_{ssDNA} + \begin{cases} \alpha\frac{t_T}{4}(C \to U)_{ssDNA} & (+) \quad \text{gene} \\ 0 & (-) \quad \text{gene} \end{cases} \tag{15}$$

and to the reverse complementary $G \to A$ substitution rate by:

$$p_{(-)}\frac{t_R}{4}(C \to U)_{ssDNA} + \begin{cases} 0 & (+) \quad \text{gene} \\ \alpha\frac{t_T}{4}(C \to U)_{ssDNA} & (-) \quad \text{gene} \end{cases} \tag{16}$$

We again recover our minimal model (Eqs. (5) to (7)) with:

$$(C \to T)_R = \frac{t_R}{4}(C \to U)_{ssDNA} \qquad \text{and} \quad (G \to A)_R = 0, \tag{17}$$

$$(C \to T)_T[\alpha] = \alpha\frac{t_T}{4}(C \to U)_{ssDNA} \qquad \text{and} \quad (G \to A)_T[\alpha] = 0. \tag{18}$$

The $(C \to T)^a$ asymmetry then follows the decomposition predicted by Eq. (9):

$$(C \to T)^a = p(C \to T)^a_R \pm \alpha(C \to T)^a_T, \tag{19}$$

with
$$(C \to T)^a_R = \frac{t_R}{8}(C \to U)_{ssDNA} > 0, \tag{20}$$

and
$$(C \to T)^a_T = \frac{t_T}{8}(C \to U)_{ssDNA} > 0. \tag{21}$$

The cytosine deamination theory predicts positive asymmetries for both replication (Frank and Lobry 1999; Polak and Arndt 2009) and transcription (Francino and Ochman 2001; Mugal et al. 2009). Importantly the cytosine deamination also affects the symmetrical part of the substitution rates (Eq. (8)):

$$(C \to T)^s = (C \to T)^s_0 + (C \to T)^s_R + \alpha(C \to T)^s_T, \tag{22}$$

with
$$(C \to T)^s_R = \frac{t_R}{8}(C \to U)_{ssDNA} > 0, \tag{23}$$

and
$$(C \to T)^s_T = \frac{t_T}{8}(C \to U)_{ssDNA} > 0. \tag{24}$$

Thus in the cytosine deamination theory, symmetrical substitution rates are higher in genic regions than in intergenic regions.

Transcription induced mutations

Along with the cytosine deamination, other types of mutagenic reactions can be considered. Mugal et al. (2009) proposed, for the human genome, to take into account the deamination of cytosine, the deamination of adenine, the oxidative stress of guanine, and the loss of a purine ($Y = A$ or G), which would result respectively after two replication rounds in the $C \to T$, $A \to G$, $G \to T$ and $Y \to T$ substitutions. During transcription, the coding strand is possibly more exposed to those mutagenic reactions (Beletskii and Bhagwat 1996, 1998). These mutagenic reactions lead to an increase of $C \to T$, $A \to G$, $G \to T$ and $A \to T$ substitution rates on the coding strand compared to the flanking intergenic regions, as observed in the human genome by Mugal et al (2009). Under transcription induced mutations alone, those substitution rates should be the same in the transcribed strand and in the flanking intergenic region. These processes lead to the asymmetries (Mugal et al. 2009):

$$(C \to T)^a_T, (A \to G)^a_T, (G \to T)^a_T, (A \to T)^a_T > 0. \tag{25}$$

Importantly they also imply that symmetrical substitution rates are higher in genic regions than in intergenic regions:

$$(C \to T)^s_T, (A \to G)^s_T, (G \to T)^s_T, (A \to T)^s_T > 0. \tag{26}$$

Transcription-coupled repair

Transcription-coupled repair (TCR), see (Svejstrup 2002) for review, has also been proposed to generate strand asymmetries (Francino et al. 1996; Green et al. 2003; Mugal et al. 2009). TCR is triggered by the stalling of RNA polymerase II due to DNA damage on the transcribed strand, and then repair is achieved using the coding strand as a template (Svejstrup 2002). TCR can therefore reduce rates of mutagenic reactions on the transcribed strands. Mugal et al (2009) proposed that the $C \to T$, $A \to G$, $G \to T$ and $A \to T$ substitution rates were lower in the transcribed strand than in the flanking intergenic region due to TCR, or other repair mechanisms. As in transcription induced mutations it leads to the asymmetries:

$$(C \to T)^a, (A \to G)^a_T, (G \to T)^a_T, (A \to T)^a_T > 0. \tag{27}$$

On the contrary to the transcription induced mutations, symmetrical substitution rates are lower in genic regions than in intergenic regions:

$$(C \to T)^s_T, (A \to G)^s_T, (G \to T)^s_T, (A \to T)^s_T < 0. \tag{28}$$

Likely, both transcription induced mutations and repair impact on substitution rates. The sign of the symmetrical part will depend of the relative strength of the two competing processes.

TCR acting on misinserted bases

Green et al. (2003) proposed that TCR could act on misinserted bases during the previous replication round. After the stalling of the RNA polymerase II, TCR can detect mispaired base in the vicinity through the MutSα mismatch repair complex, and will resolve the mismatch using the coding strand as a template (Svejstrup 2002). In non-transcribed regions, a misinserted base can presumably persist until the next replication round, where it will have a 50% chance to result in a substitution. In transcribed regions however, if TCR resolves the mismatch, it results in a substitution if and only if the misinserted base was on the coding strand.

Modelling TCR acting on a misinserted base. Let us call $P_{TCR}[\alpha]$ the probability that TCR detects and repairs the mismatch. This probability is likely close to 0 for weakly expressed genes and likely increases with the transcription rate α (but is always inferior to 1). The $i \to j$ substitution can result from the mispaired bases $j : i^c$ and $i : j^c$. In non-transcribed regions (intergenic regions in our model), the mispaired base has a 50% chance to result in the $i \to j$ substitution:

$$\text{intergenic} \quad \begin{matrix} i \\ j^c \end{matrix} \xrightarrow{\frac{1}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1}{2}} \begin{matrix} j \\ i^c \end{matrix} \tag{29}$$

In transcribed regions (sense and antisense genic regions in our model), TCR with a probability $P_{TCR}[\alpha]$ repairs towards the coding strand:

$$(+) \quad \text{gene} \quad \begin{matrix} i \\ j^c \end{matrix} \xrightarrow{\frac{1-P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1+P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ i^c \end{matrix} \tag{30}$$

$$(-) \quad \text{gene} \quad \begin{matrix} i \\ j^c \end{matrix} \xrightarrow{\frac{1+P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1-P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ i^c \end{matrix} \tag{31}$$

where as usual the reference strand is the upper strand. The relations Eqs. (29)-(31) can be written in the compact form:

$$\begin{matrix} i \\ j^c \end{matrix} \xrightarrow{\frac{1\mp P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1\pm P_{TCR}[\alpha]}{2}} \begin{matrix} j \\ i^c \end{matrix} \tag{32}$$

for $(\pm)$ genes and where $\alpha = 0$ ($P_{TCR}[\alpha] = 0$) corresponds to the intergenic case (Eq. (29)). Green et al. (2003) assumed that the replication induced misinsertions

occur with equal frequencies on the two strands. This is the case if we neglect strand asymmetry due to replication, *e.g.* when the replication fork polarity $p = 0$. The mispaired bases are then created with rates (Eqs. (10) and (11)):

$$(p=0) \quad \begin{matrix} i \\ j^c \end{matrix} \xleftarrow{\frac{(\Sigma^c+\Delta^c)_{ji}}{2}} \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{\frac{(\Sigma+\Delta)_{ji}}{2}} \begin{matrix} j \\ i^c \end{matrix} . \tag{33}$$

Combining Eqs. (32) and (33), TCR acting on misinserted bases contributes to the $i \to j$ substitution by:

$$\begin{aligned} \frac{(1 \mp P_{TCR}[\alpha])}{2} \frac{(\Sigma^c + \Delta^c)_{ji}}{2} + \frac{(1 \pm P_{TCR}[\alpha])}{2} \frac{(\Sigma + \Delta)_{ji}}{2} \\ = \frac{(\Sigma^s + \Delta^s)_{ji}}{2} \pm P_{TCR}[\alpha] \frac{(\Sigma^a + \Delta^a)_{ji}}{2}. \end{aligned} \tag{34}$$

Consistently with our minimal model, we recover Eqs. (8) and (9) for $p = 0$ with:

$$\tau_R^s = \frac{(\Sigma^s + \Delta^s)}{2}, \quad \tau_T^a[\alpha] = P_{TCR}[\alpha] \frac{\Sigma^a + \Delta^a}{2}, \quad \text{and} \quad \tau_T^s[\alpha] = 0. \tag{35}$$

This model has therefore two consequences: it generates an asymmetry that depends on the error spectra of the two DNA polymerases, and the symmetrical rates are equal in genic and intergenic regions. Green et al (2003) proposed this model to explain the $(A \to G)^a > 0$ asymmetry observed in the coding strand of mammalian genes. This model is consistent with the observation that the symmetrical rate $(A \to G)^s$ has approximately the same value in genes and in their flanking intergenic regions (Green et al. 2003).

When reaching the limits of our minimal model (Eqs. (5) *to* (7)*).* When taking into account the strand asymmetry due to replication ($p \neq 0$), the mispaired bases are now created with rates (Eqs. (10) and (11)):

$$\begin{matrix} i \\ j^c \end{matrix} \xleftarrow{(p_{(-)}\Sigma^c + p_{(+)}\Delta^c)_{ji}} \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{(p_{(+)}\Sigma + p_{(-)}\Delta)_{ji}} \begin{matrix} j \\ i^c \end{matrix} . \tag{36}$$

Combining relations Eqs. (32) and (36) we get:

$$\tau[p, \alpha, (\pm)] = \frac{1 \mp P_{TCR}[\alpha]}{2} (p_{(-)}\Sigma^c + p_{(+)}\Delta^c) + \frac{1 \pm P_{TCR}[\alpha]}{2} (p_{(+)}\Sigma + p_{(-)}\Delta). \tag{37}$$

This complicated relation nonetheless satisfies strand exchange symmetry (Eq. (2)), as it should. If we develop the different terms, we find for the symmetrical part:

$$\tau^s[p, \alpha, (\pm)] = \frac{\Sigma^s + \Delta^s}{2} \pm p P_{TCR}[\alpha] \frac{\Sigma^s - \Delta^s}{2}. \tag{38}$$

A new term, that depends on the replication fork polarity p and gene orientation $(\pm)$, is found adding to the symmetrical part predicted by our minimal model (Eq. (8)).

Note that even with this additional term, the symmetrical part still obeys strand exchange symmetry (Eq. (3)). Interestingly, the asymmetrical part still satisfies Eq. (9) with:

$$\tau_R^a = \frac{\Sigma^a - \Delta^a}{2}, \quad \text{and} \quad \tau_T^a[\alpha] = P_{TCR}[\alpha]\frac{\Sigma^a + \Delta^a}{2}. \tag{39}$$

Remark. For the transcription-associated asymmetry $\tau_T^a[\alpha]$, we recover Eq. (35) derived in the $p = 0$ case. For the replication-associated asymmetry τ_R^a, we recover Eq. (14) previously derived in the $\alpha = 0$ ($P_{TCR}[\alpha] = 0$) case.

☛ *Consistently with our minimal model, all the molecular mechanisms considered here lead to Eq.* (9) *for the substitutional asymmetry. The symmetrical part also satisfies the predicted Eq.* (8)*, except when TCR is acting on misinserted bases, where a new term must be added into Eq.* (38)*. Importantly, the signs of the coefficients* τ_R^a*,* τ_T^a *and* τ_T^s *depend on the underlying molecular mechanisms and their relative strengths.*

I.2.3 Determination of substitution rates in the human genome

Our collaborators determined substitution rates in the human genome (Chen et al. 2010), from which we estimated the values of the different coefficients of our minimal model. To do so, we computed the average values in regions of given replication fork polarity and given transcriptional status. Importantly the substitutional asymmetry is found to be small.

Methodology

Substitutions were tabulated in the human lineage since its divergence with chimpanzee using macaca and orangutan as outgroups (Chen et al. 2010). Sequences were divided into CpG and non-CpG sites in the ancestral human-chimpanzee genome (CpG means a C followed by a G in the DNA sequence *i.e.* $5' - CG - 3'$). Cytosine when methylated can spontaneously deaminates into thymine. In vertebrates genomes, most CpG dinucleotides have their cytosine methylated with the exception of a few genomic regions called CpG islands, see (Suzuki and Bird 2008) for review. As a result the CpG dinucleotide is hypermutable, and the $CpG \to TpG$ and its reverse complementary $CpG \to CpA$ are by far the principal neighbor-dependent substitutions rates (Hess et al. 1994; Arndt and Hwa 2005). The twelve neighbor-independent substitution rates were determined on non-CpG sites. The two neighbor-dependent $CpG \to TpG$ and $CpG \to CpA$ substitution rates were determined on CpG sites. CpG islands and exons were excluded from the analysis as they are unlikely to evolve neutrally. The first and last 500 bp of intronic sequences were also excluded to avoid bias due to splicing sites (Touchon et al. 2004). Genomic regions were classed as genic (+), genic (−), and intergenic using RefGene transcripts. As an estimator of the replication fork polarity, for a reason that will be justified in Chapter II, we took the derivative of the mean replication timing

abbreviated as dMRT/dx. The dMRT/dx profile was computed as in (Baker et al. 2011), using experimental replication timing data obtained from (Chen et al. 2010; Hansen et al. 2010). Note that the MRT profile is expressed as a fraction of S-phase and has therefore no dimension. The dMRT/dx profile will be expressed in Mbp^{-1}. If we multiply the MRT by the duration of S-phase we get a reasonable proxy for the MRT expressed in time, although the conversion between S-phase fraction and time is not strictly linear (Blumenthal et al. 1974). The replication fork polarity can be estimated as:

$$p(x) \simeq v\, T_S\, \text{dMRT/dx}, \tag{40}$$

where v is the replication fork velocity and T_S the duration of S-phase. Note that the MRT profile was determined for seven cell lines: an ESC cell line (BG02), three lymphoblastoid cell lines (GM06990, H0287, TL010), a fibroblast cell line (BJ, replicates R1 and R2), an erythroid cell line (K562), and a HeLa cell line. The MRT profile, the replication fork velocity v, the duration of the S-phase T_S, and consequently the replication fork polarity are all cell type specific. For our study of strand asymmetry, we would like to have access to germline replication fork polarity, as only mutations occurring in the germline are transmitted to the descendants. Unfortunately, no germline replication timing data are available today. As a substitute to germline dMRT/dx, we use the dMRT/dx profile obtained in the BG02 embryonic stem cell line. The conservation of the dMRT/dx profile between different cell lines will be addressed in Chapters III, IV and V. For our current purpose we just point out that the dMRT/dx profile in one cell line correlates with the dMRT/dx profile in another cell line (Baker et al. 2011). For all the reasons mentioned above, we conjecture that the dMRT/dx profile in BG02 is, on average, proportional to the replication fork polarity in the germline. We estimated substitution rates by concatenating the sequences of all genomic regions of given transcriptional status (genic (+), intergenic, genic (−)) and given dMRT/dx. The substitution rates thus correspond to averaged values, usually on several Mbp of aligned sequences.

Results

We considered genomic regions with dMRT/dx > 1 Mbp^{-1} (in BG02 cell line) which are likely to have a positive replication fork polarity $\bar{p} > 0$ in the germline. The precise value of $\bar{p} > 0$ is unknown. The neighbor-independent (single nucleotide) substitution rates are tabulated in the form of a substitution rate matrix (see Section I.3.1). The substitution rate matrix $M_{\text{inter},\bar{p}}$ and $M_{\text{sense},\bar{p}}$ obtained in intergenic regions and genic sense (+) regions are equal to:

$M_{\text{inter},\bar{p}} =$

↙	T	A	G	C
T		0.638	1.234	3.804
A	0.606		3.639	1.189
G	0.778	3.254		1.244
C	2.873	0.809	1.139	

(41)

and:

$$M_{\text{sense},\bar{p}} = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & 0.601 & 1.118 & 3.535 \\ \hline \text{A} & 0.509 & & 3.392 & 0.926 \\ \hline \text{G} & 0.776 & 3.700 & & 1.383 \\ \hline \text{C} & 2.418 & 0.817 & 1.077 & \\ \hline \end{array} \tag{42}$$

where substitution rates are expressed in kbp^{-1}. We can readily see that substitution rates in Eqs. (41) and (42) do not respect PR1 (Section I.1.2). For instance, for the $M_{\text{inter},\bar{p}}$ substitution rate matrix, $A \to G = 3.254 \neq T \to C = 2.873$. According to Eq. (8), the symmetrical part of $M_{\text{inter},\bar{p}}$ provides an estimate of $M_0^s + M_R^s$:

$$M_0^s + M_R^s = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & 0.622 & 1.211 & 3.721 \\ \hline \text{A} & 0.622 & & 3.721 & 1.211 \\ \hline \text{G} & 0.794 & 3.063 & & 1.191 \\ \hline \text{C} & 3.063 & 0.794 & 1.191 & \\ \hline \end{array} . \tag{43}$$

Note that the symmetrical part, by definition, satisfies PR1. The asymmetrical part of $M_{\text{inter},\bar{p}}$, according to relation Eq. (9), provides an estimate of $\bar{p}M_R^a$:

$$\bar{p}M_R^a = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & 0.016 & 0.022 & 0.082 \\ \hline \text{A} & -0.016 & & -0.082 & -0.022 \\ \hline \text{G} & -0.015 & 0.190 & & 0.052 \\ \hline \text{C} & -0.190 & 0.015 & -0.052 & \\ \hline \end{array} . \tag{44}$$

The non null substitutional asymmetry $\bar{p}M_R^a \neq 0$ is responsible for the breaking of PR1 in Eq. (41). We further remark from Eqs.(41) and (42) that substitution rates are not equal in intergenic and genic regions. For instance, $A \to G = 3.254$ for the $M_{\text{inter},\bar{p}}$ matrix whereas $A \to G = 3.700$ for the $M_{\text{genic }(+),\bar{p}}$ matrix. According to relation Eq. (6), $M_{\text{sense},\bar{p}} - M_{\text{inter},\bar{p}}$ provides an estimate of M_T:

$$M_T = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & -0.036 & -0.116 & -0.268 \\ \hline \text{A} & -0.097 & & -0.246 & -0.263 \\ \hline \text{G} & -0.002 & 0.446 & & 0.139 \\ \hline \text{C} & -0.456 & 0.007 & -0.062 & \\ \hline \end{array} \tag{45}$$

which symmetrical and asymmetrical parts are respectively:

$$M_T^s = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & -0.067 & -0.189 & -0.257 \\ \hline \text{A} & -0.067 & & -0.257 & -0.189 \\ \hline \text{G} & 0.003 & -0.005 & & 0.039 \\ \hline \text{C} & -0.005 & 0.003 & 0.039 & \\ \hline \end{array} \tag{46}$$

and

$$M_T^a = \begin{array}{|c|c|c|c|c|}\hline \swarrow & \text{T} & \text{A} & \text{G} & \text{C} \\ \hline \text{T} & & 0.030 & 0.074 & -0.011 \\ \hline \text{A} & -0.030 & & 0.011 & -0.074 \\ \hline \text{G} & -0.005 & 0.451 & & 0.101 \\ \hline \text{C} & -0.451 & 0.005 & -0.101 & \\ \hline \end{array} . \tag{47}$$

For the neighbor-dependent rate $r = CpG \to TpG$, we obtained:

$$r_{\text{inter},\bar{p}} = 53.738, \qquad r^c_{\text{inter},\bar{p}} = 47.122, \tag{48}$$

$$\text{and} \quad r_{\text{genic (+)},\bar{p}} = 51.842, \quad r^c_{\text{genic (+)},\bar{p}} = 42.885, \tag{49}$$

Note that, as expected, the neighbor-dependent $CpG \to TpG$ rate is much higher (about 10 fold) than the $C \to T$ rate. We obtained the coefficients:

$$r_0^s + r_R^s = 50.43, \quad \bar{p} r_R^a = 3.308, \quad r_T^s = -3.067, \quad \text{and} \quad r_T^a = 1.171, \tag{50}$$

where as already mentioned substitution rates are expressed in kbp^{-1}.

Remark. The transcription-associated component τ_T was determined over all genic (+) regions, whatever their transcription rates α. The given transcription-associated component τ_T thus corresponds to the $\tau_T[\alpha]$ coefficient averaged over α. The given replication-associated asymmetry $\bar{p}\tau_R^s$ depends on the (unknown) average replication fork polarity $\bar{p}$. We independently measured substitution rates in genomic regions with $\text{dMRT/dx} < -1$ Mbp^{-1} which are likely to have the opposite replication fork polarity $-\bar{p} < 0$. We got approximately the same values for the coefficients as in Eqs. (43) and (44). We also independently measured the net effect of transcription using genic (−) regions. We got approximately the same values for the coefficients in Eqs. (46) and (47). The compliance of substitution rates with Eqs. (8) and (9) will be explicitly addressed in Chapter III.

The substitutional asymmetry is small

Note that the asymmetrical parts $\bar{p}\tau_R^a$ and τ_T^a are small as compared to the symmetrical part $\tau_0^s + \tau_R^s$. The symmetrical coefficients τ_T^s are also small as compared to the symmetrical part $\tau_0^s + \tau_R^s$. More precisely, for the values reported above we have:

$$\bar{p}|\tau_R^a| = \epsilon|\tau_0^s + \tau_R^s|, \quad \text{with} \quad \epsilon \leq 0.07 \tag{51}$$

$$|\tau_T^a| = \epsilon|\tau_0^s + \tau_R^s|, \quad \text{with} \quad \epsilon \leq 0.15 \tag{52}$$

$$|\tau_T^s| = \epsilon|\tau_0^s + \tau_R^s|, \quad \text{with} \quad \epsilon \leq 0.16 \, . \tag{53}$$

☛ *The coefficients $\tau_R^a, \tau_T^a, \tau_T^s$ are small as compared to the symmetrical $\tau_0^s + \tau_R^s$ substitution rates.*

I.2.4 From substitutional to compositional asymmetry

The DNA sequences, under the repetitive exposure to substitutional patterns, change their nucleotide compositions over time (Graur and Li 1999). How does the substitutional asymmetry reflect on the DNA composition evolution ? When the substitution rates are equal on the two DNA strands (PR1), the compositions are equal on the two DNA strands (PR2) (Lobry 1995; Lobry and Lobry 1999). On the opposite we expect that a substitutional asymmetry gives rise to a compositional asymmetry. For instance, if we have the $(A \to G)^a > 0$ asymmetry, *i.e.* an excess of $A \to G$ versus $T \to C$ substitutions, we expect after long evolutionary time $[G] > [C]$ and $[T] > [A]$. The next two mathematical Sections I.3 and I.4 formalize the evolution of the compositional asymmetry and prove this assertion. In Section I.4 we use, as a central hypothesis, the smallness of the substitutional asymmetry to relate directly the compositional asymmetry to the substitutional one. An important consequence of this mathematical treatment is that the decomposition of the substitutional asymmetry into transcription and replication components in Eq. (9) directly reflects on the compositional asymmetry. This key result, that will be exploited all along this thesis manuscript, is reported in Section I.5.

☛ *Readers not interested in the mathematical formalism are encouraged to skip Sections I.3 and I.4 and to go directly to Section I.5.*

I.3 DNA composition evolution

We first recall the general formalism of DNA composition evolution (Graur and Li 1999), for neighbor-independent and time homogeneous substitutions. We then exploit the symmetry of strand exchange to rewrite the equations under a more suitable form for the study of strand asymmetry (Lobry and Lobry 1999).

I.3.1 General formalism

Time evolution of the composition

In the case of neighbor-independent and time homogeneous substitutions, the time evolution of the DNA composition is given by (Graur and Li 1999):

$$\frac{d}{dt}X(t) = MX(t), \tag{54}$$

where $X(t)$ is the frequency (or probability) vector; for a nucleotide $i \in \{T, A, G, C\}$, $X_i(t)$ is the frequency (or probability) of i in the DNA sequence at time t. M is called the **substitution rate matrix**; for $i \neq j \in \{T, A, G, C\}$, the element M_{ij} is the substitution rate $j \to i$ (expressed in per bp per unit of time). Diagonal elements of M are such that sum over rows are null: $M_{jj} = -\sum_{i \neq j} M_{ij}$. When $X(t)$ is thought as a probability vector, Eq. (54) is called the master equation; it

is the time continuous formulation of a Markov chain. The general properties of a Markov chain are well known (Van Kampen 2007), we briefly expose them below.

Evolution towards equilibrium

First, we can easily integrate Eq. (54) to get the composition $X(t)$ at any time t, knowing the initial composition $X(t_0)$ at a time t_0:

$$X(t) = W(t, t_0)X(t_0), \quad \text{where} \quad W(t, t_0) = e^{(t-t_0)M}. \tag{55}$$

The matrix $W(t, t_0)$ gives the substitution probabilities between t_0 and t; for $i, j \in \{T, A, G, C\}$, $W_{ij}(t, t_0) = \text{Prob}(i \text{ at time } t \,|\, j \text{ at time } t_0)$. We recover in this form the time discrete formulation of a Markov chain. Note that from this formulation we have necessarily $\sum_i W_{ij}(t, t_0) = 1$, which in the limit $t \to t_0$ gives the condition $\sum_i M_{ij} = 0$ for M. This property also ensures that $\sum_i X_i(t) = 1$ at all time t. The spectral properties of M are important to give the asymptotic behaviour of $X(t)$. There is a unique vector X^*, called the **equilibrium vector**, such as:

$$MX^* = 0 \quad \text{and} \quad \sum_i X_i^* = 1. \tag{56}$$

So X^* is an eigenvector of M with eigenvalue 0. The three other eigenvalues have all a strictly negative real part:

$$MX^{(a)} = \left[-\frac{1}{\tau^{(a)}} + i\omega^{(a)} \right] X^{(a)} \quad \text{with} \quad \tau^{(a)} > 0, \quad a \in \{1, 2, 3\}. \tag{57}$$

These spectral properties have been demonstrated in many different ways. One can for instance use the fact that $W(t, t_0)$ belongs to the class of "stochastic matrices" ($0 \leq W_{ij}(t, t_0) \leq 1$ and $\sum_i W_{ij}(t, t_0) = 1$) and apply Perron-Frobenius theorem (Van Kampen 2007). There are some exceptional cases where Eqs. (56) and (57) are not verified, but such cases are never encountered in DNA composition evolution and therefore not relevant for our purpose. It follows from Eqs. (56) and (57) that the composition $X(t)$ converges asymptotically towards the equilibrium value X^*, whatever the initial composition $X(t_0)$:

$$\boxed{X(t) = e^{M(t-t_0)}X(t_0) \to X^* \quad \text{when} \quad t \to \infty} \tag{58}$$

☛ *Convergence towards equilibrium is the central property of the time evolution Eq.* (54)*.*

I.3.2 Exploiting strand exchange symmetry

We previously defined strand exchange symmetry for substitution rates (Eq. (2)). More abstractly, we can define strand exchange symmetry as the change of strand

referential (exchanging the reference strand and the complementary strand). We can study more easily strand asymmetry if we systematically define variables that are invariant under strand exchange symmetry and variables that change sign. Using this reformulation, we easily recover PR2 (Section I.1.2).

Defining strand-symmetric and strand-asymmetric variables

The time evolution on the complementary strand is given by:

$$\frac{d}{dt}X^c(t) = M^c X^c(t), \tag{59}$$

where $X^c(t)$ is the frequency vector on the complementary strand, and M^c is the substitution rate matrix computed on the complementary strand. For a nucleotide $i \in \{T, A, G, C\}$, let us denote by $i^c \in \{A, T, C, G\}$ the corresponding complementary nucleotide. By reverse complementarity we have $X_i^c(t) = X_{i^c}(t)$ and $M_{ij}^c = M_{i^c j^c}$. We can decompose M into a symmetrical and an asymmetrical part under strand exchange symmetry:

$$M = M^s + M^a \quad \text{with} \quad M^s = \frac{M + M^c}{2}, \quad M^a = \frac{M - M^c}{2}. \tag{60}$$

It is more convenient to consider the evolution of DNA composition through the following variables (Lobry and Lobry 1999):

$$Y = \begin{pmatrix} \theta \\ S \end{pmatrix} = \begin{pmatrix} \theta_{TA} \\ \theta_{GC} \\ S_{TA} \\ S_{GC} \end{pmatrix} = UX = \begin{pmatrix} X_T + X_A \\ X_G + X_C \\ X_T - X_A \\ X_G - X_C \end{pmatrix}, \tag{61}$$

where S_{TA} and S_{GC} are the compositional skews and θ_{TA} and θ_{GC} are the TA and GC contents. The TA and GC contents are invariant under strand exchange symmetry, whereas the compositional skews change sign. The change of coordinate matrix U and its inverse are given by:

$$U = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad U^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{62}$$

It is easy to get the evolution of Y through a linear transformation of Eq. (54):

$$\frac{dY(t)}{dt} = NY(t) \quad \text{with} \quad N = UMU^{-1}. \tag{63}$$

Similarly it is straightforward to get the equilibrium composition Y^* through a linear transformation of Eq. (56):

$$NY^* = 0 \quad \text{and} \quad \theta_{TA}^* + \theta_{GC}^* = 1. \tag{64}$$

The symmetry properties of M^s and M^a imply the following block forms for:

$$N^s = UM^sU^{-1} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \tag{65}$$

$$N^a = UM^aU^{-1} = \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}. \tag{66}$$

The A and D matrices are invariant under strand exchange symmetry whereas the B and C matrices change sign. More explicitly, in the $\{T, A, G, C\}$ coordinates, the symmetrical and asymmetrical parts of M have the following forms:

$$M^s = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \mu & \nu & \kappa & \epsilon \\ \nu & \mu & \epsilon & \kappa \end{pmatrix}, \quad M^a = \begin{pmatrix} a & -b & c & -d \\ b & -a & d & -c \\ m & -n & k & -e \\ n & -m & e & -k \end{pmatrix} \tag{67}$$

The matrices A and D introduced in Eq. (65) are equal to:

$$A = \begin{pmatrix} \alpha+\beta & \gamma+\delta \\ \mu+\nu & \kappa+\epsilon \end{pmatrix}, \quad D = \begin{pmatrix} \alpha-\beta & \gamma-\delta \\ \mu-\nu & \kappa-\epsilon \end{pmatrix}, \tag{68}$$

and the matrices B and C introduced in Eq. (66) are equal to:

$$B = \begin{pmatrix} a+b & c+d \\ m+n & k+e \end{pmatrix}, \quad C = \begin{pmatrix} a-b & c-d \\ m-n & k-e \end{pmatrix}. \tag{69}$$

Following (Duret and Arndt 2008), the coefficients of the matrix A can also be expressed as substitution rates between weak ($W = A, T$) and strong ($S = G, C$) nucleotides:

$$\mu + \nu = (T \to G)^s + (T \to C)^s = (W \to S), \tag{70}$$

$$\gamma + \delta = (G \to T)^s + (G \to A)^s = (S \to W). \tag{71}$$

The spectral properties of the matrices A and D will be needed for the time evolution of the composition. The eigenvalues and eigenvectors of A are given by:

$$A\theta_A = 0 \qquad \text{with} \quad \theta_A = \frac{1}{\mu+\nu+\gamma+\delta}\begin{pmatrix} \gamma+\delta \\ \mu+\nu \end{pmatrix}, \tag{72}$$

$$A\begin{pmatrix} 1 \\ -1 \end{pmatrix} = -\frac{1}{\tau_A}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{with} \quad \tau_A = \frac{1}{\mu+\nu+\gamma+\delta}. \tag{73}$$

Expressed in terms of weak to strong and strong to weak substitution rates, the GC component of θ_A and τ_A are equal to (Duret and Arndt 2008):

$$\theta_{A,GC} = \frac{(W \to S)}{(W \to S) + (S \to W)}, \quad \tau_A = \frac{1}{(W \to S) + (S \to W)}. \tag{74}$$

The two eigenvalues of D have a strictly negative real part:

$$DS^{(a)} = \left[-\frac{1}{\tau_D^{(a)}} + i\omega^{(a)} \right] S^{(a)} \quad \text{with} \quad \tau_D^{(a)} > 0, \quad a \in \{1, 2\}. \tag{75}$$

Hence, D is invertible and $e^{tD} \to 0$ when $t \to \infty$. As it will become clear in the next paragraph, θ_A characterizes the equilibrium composition, while τ_A and $\tau_D^{(1,2)}$ are characteristic time-scales of the DNA composition evolution, when the substitution rate matrix satisfies PR1.

Proof. As $\sum_i M_{ij}^s = 0$ and for $i \neq j, M_{ij}^s > 0$, we know according to Eqs. (56) and (57) that M^s has 0 as eigenvalue, and three eigenvalues with a strictly negative real part. N^s and M^s are similar, they have therefore the same eigenvalues. As 0 and $-\frac{1}{\tau_A}$ are the two eigenvalues of A, the two remaining eigenvalues of N^s, those of D, have a strictly negative real part.

Strand symmetry: evolution under PR1

We consider here the case where there is no substitutional asymmetry $M^a = 0$, in other words the substitution rate matrix is symmetrical $M = M^s$ and satisfies PR1 (Sueoka 1995). It implies in turn that the matrices B and C are null (Eq. (66)). The equilibrium TA and GC contents and the equilibrium skews satisfy:

$$A\theta^* = 0 \quad \text{and} \quad DS^* = 0, \tag{76}$$

with the constraint $\theta_{TA}^* + \theta_{GC}^* = 1$ (Eq. (64)). According to the spectral properties of A and D derived just above (Eqs. (72) and (75)), the solutions of Eq. (76) are:

$$\theta^* = \theta_A \quad \text{and} \quad S^* = 0. \tag{77}$$

The equations of evolution for the TA and GC contents and the compositional skews are respectively given by:

$$\frac{d}{dt}\theta(t) = A\theta(t), \tag{78}$$

$$\frac{d}{dt}S(t) = DS(t), \tag{79}$$

whose solutions are:

$$\theta(t) = e^{A(t-t_0)}\theta(t_0) \to \theta_A \quad \text{when} \quad t - t_0 \gg \tau_A, \tag{80}$$

$$S(t) = e^{D(t-t_0)}S(t_0) \to 0 \quad \text{when} \quad t - t_0 \gg \tau_D^{(1)}, \tau_D^{(2)}. \tag{81}$$

We recover the result of Lobry (1995): if the substitution rate matrix is symmetrical (PR1), then the compositional skews are null at equilibrium (PR2):

$$\boxed{\text{if} \quad M = M^s \text{ (PR1)} \quad \text{then} \quad S_{TA}^* = S_{GC}^* = 0 \text{ (PR2)}}. \tag{82}$$

The TA and GC contents converge exponentially toward their equilibrium values θ_A with the characteristic time scale τ_A. More explicitly the evolution of the GC content[1] is given by:

$$\theta_{GC}(t) = e^{-\frac{(t-t_0)}{\tau_A}} \theta_{GC}(t_0) + \left(1 - e^{-\frac{(t-t_0)}{\tau_A}}\right) \theta^*_{GC}. \tag{83}$$

Hence the half time $t_{1/2}$ of the GC content evolution, defined as the time necessary to divide by two the difference between the GC content and the equilibrium GC content, is given by (Duret and Arndt 2008):

$$\frac{\theta^*_{GC} - \theta_{GC}(t)}{\theta^*_{GC} - \theta_{GC}(t_0)} = e^{-\frac{(t-t_0)}{\tau_A}} = \frac{1}{2}, \quad \text{for} \quad t - t_0 = t_{1/2} = \ln 2 \; \tau_A. \tag{84}$$

The compositional skews decay towards zero with two time-scales $\tau_D^{(1)}$ and $\tau_D^{(2)}$. More precisely, the projections of the compositional skew $S(t)$ onto the eigenvectors $S^{(1)}$ and $S^{(2)}$ of D (Eq. (75)) decay exponentially with respective characteristic time-scales $\tau_D^{(1)}$ and $\tau_D^{(2)}$, and the corresponding half times are given by $\ln 2 \; \tau_D^{(1)}$ and $\ln 2 \; \tau_D^{(2)}$.

Numerical test. In Fig. 8, we illustrate the time evolution under PR1, using the symmetrical substitution rate matrix $M = M_0^s + M_R^s$ given in Eq. (43). To express substitution rates in per bp per Myrs units, I used 5 Myrs as an estimation of the human-chimpanzee divergence. As predicted by Eq. (80), the TA and GC contents converge towards their equilibrium values whatever their initial values (Fig. 8A,B). The equilibrium GC content is equal to $\theta^*_{GC} = 44\%$ and the characteristic time scales are equal to $\tau_A = 568$ Myrs (corresponding half time $t_{1/2} = 393$ Myrs), $\tau_D^{(1)} = 566$ Myrs and $\tau_D^{(2)} = 1398$ Myrs (corresponding half times 385 Myrs and 969 Myrs). The dynamics of the GC content and the skews are therefore extremely slow. As predicted by Eq. (81), the TA and GC skews decay towards 0 whatever their initial values (Fig. 8C,D).

Comment on the GC content evolution. The symmetrical substitution rates (and therefore the GC content evolution) depend on many variables not taken into account here: for instance recombination rates in the context of the biased gene conversion (BGC) model (Duret and Arndt 2008), or replication timing (Stamatoyannopoulos et al. 2009; Chen et al. 2010). The value found for GC^* (44%) corresponds to the highest values found in reference (Duret and Arndt 2008), for reasons currently unclear. The substitution rate matrix $M_0^s + M_R^s$ was determined in regions of high replication polarity, and could maybe correspond to regions of high recombination rate. In the BGC model, the half time $t_{1/2}$ strongly depends on the recombination

[1]The TA content evolution is somehow redundant with the GC content evolution, as at all time we have $\theta_{TA}(t) + \theta_{GC}(t) = 1$.

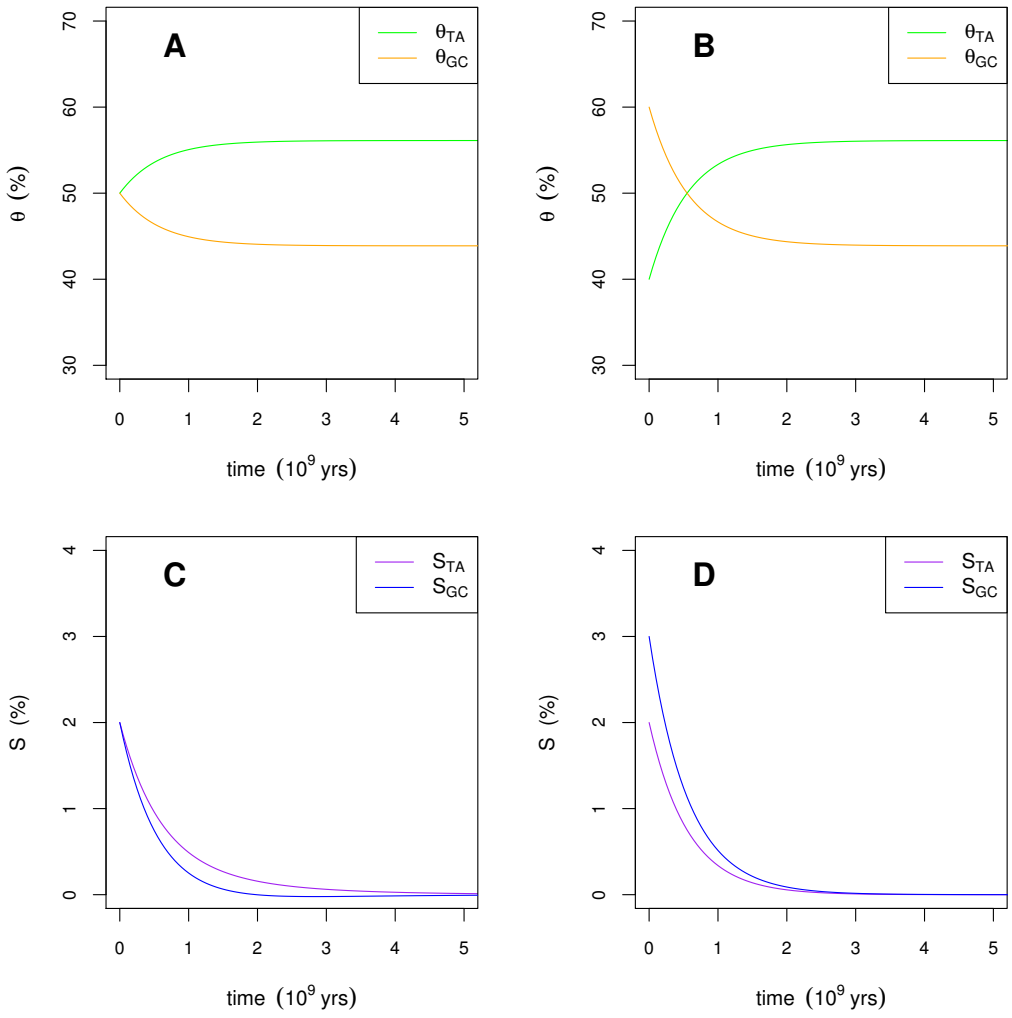

Figure 8: **DNA composition evolution under PR1 satisfies PR2 asymptotically.** The substitution rate matrix is symmetric $M = M_0^s + M_R^s$ (Eq. (43)). Time evolution of the TA and GC contents with the initial conditions (A) $\theta_{TA}(0) = \theta_{GC}(0) = 50\%$, and (B) $\theta_{TA}(0) = 40\%$ and $\theta_{GC}(0) = 60\%$. Time evolution of the TA and GC skews with the initial conditions (C) $S_{TA}(0) = S_{GC}(0) = 2\%$, and (D) $S_{TA}(0) = 2\%$ and $S_{GC}(0) = 3\%$.

rate and the effective population size. In absence of recombination, the BGC model predicts $t_{1/2} \sim 470$ Myrs, and the GC content evolution is extremely slow (Duret and Arndt 2008). But in genomic region of high recombination rate, for species with large effective population size, the GC content evolution is predicted to be much faster $t_{1/2} \sim 62$ Myrs (Duret and Arndt 2008). Our value for $t_{1/2}$ (393 Myrs) therefore corresponds to an intermediate value between the two extremes proposed by the BGC model.

Strand asymmetry

When the PR1 symmetry is broken $M^a \neq 0$, the matrices B and C are no longer null (Eq. (66)). The equilibrium TA and GC contents and the equilibrium skews are now solutions of the equations:

$$A\theta^* + BS^* = 0 \quad \text{and} \quad C\theta^* + DS^* = 0, \tag{85}$$

with the constraint $\theta^*_{TA} + \theta^*_{GC} = 1$. The evolutions of the TA and GC contents and the skews are now governed by the following ordinary differential equations:

$$\frac{d}{dt}\theta(t) = A\theta(t) + BS(t), \tag{86}$$

$$\frac{d}{dt}S(t) = C\theta(t) + DS(t). \tag{87}$$

How are the time evolutions of the skews and the GC content affected by the substitutional asymmetry ? How are their equilibrium values modified ? Is PR2 still verified ? These fundamental questions are addressed in Section I.4 using pertubative analysis. Under the assumption that the substitutional asymmetry is small, we will solve the time evolution Eqs. (86) and (87) and the equilibrium composition Eq. (85), and show that PR2 is explicitly broken.

☛ *We systematically defined symmetric and asymmetric variables under strand exchange symmetry. Among the strand-symmetric variables we have: the TA and GC contents θ, the symmetrical substitution rates M^s, the 2×2 matrices A and D defined from M^s, the equilibrium TA and GC contents θ_A, the time scale τ_A of TA and GC contents evolution, the time scales $\tau_D^{(1)}$ and $\tau_D^{(2)}$ of the compositional skews evolution. Among the strand-asymmetric variables we have: the TA and GC skews S, the substitutional asymmetries M^a, and the 2×2 matrices B and C defined from M^a.*

I.4 Pertubative analysis of the compositional asymmetry

Under symmetrical substitution rates (PR1), the compositional skews decay towards 0 (PR2) as demonstrated by Lobry (1995). In this section we are interested in the

opposite situation: establishment and maintenance of the compositional skews in the presence of a substitutional asymmetry. In the human genome, we found that the substitutional asymmetry was small (Eqs. (51) and (52)). As we know how the compositional skews evolve under symmetrical substitution rates, we will use perturbation theory to determine the skews evolution, adding a small perturbation that will be the substitutional asymmetry. This section is devoted to this pertubative resolution. Importantly, for substitution rates satisfying our minimal model (Eqs. (8) and (9)), we demonstrate that the compositional asymmetry can be decomposed into transcription- and replication-associated components. We also extend this result for neighbor-dependent substitutions (Arndt et al. 2003) and formally for time-dependent substitution rates (Lobry and Lobry 1999).

I.4.1 General principles

Pertubative analysis of $X(t)$

Let us illustrate the principles of perturbation theory on the time evolution of the composition $X(t)$ governed by Eq. (54). For other quantities, we will just give the results, as the same method will be used repeatedly. Here we consider the symmetrical part M^s as order $O(1)$ and the asymmetrical part M^a as a small perturbation of order $O(\epsilon)$. We define the expansion of the composition $X(t)$ in order of ϵ (Eqs. (51)-(53)):

$$X(t) = X^{(0)}(t) + \epsilon X^{(1)}(t) + \epsilon^2 X^{(2)}(t) + \cdots \tag{88}$$

We have then to solve the time evolution Eq. (54) order by order in ϵ, considering M^a of order ϵ:

$$\frac{d}{dt}X^{(0)}(t) = M^s X^{(0)}(t), \tag{89}$$

$$\epsilon^n \frac{d}{dt}X^{(n)}(t) = \epsilon^n M^s X^{(n)}(t) + \epsilon^{n-1} M^a X^{(n-1)}(t), \quad \text{for} \quad n \geq 1, \tag{90}$$

with the initial condition $X(t_0) = X^{(0)}(t_0)$ that we choose not to depend upon ϵ. The solutions of these differential equations are:

$$X^{(0)}(t) = e^{M^s(t-t_0)}X(t_0), \tag{91}$$

$$\epsilon X^{(1)}(t) = \int_{t \geq t_1 \geq t_0} dt_1 e^{M^s(t-t_1)} M^a e^{M^s(t_1-t_0)} X(t_0), \tag{92}$$

$$\epsilon^2 X^{(2)}(t) = \int_{t \geq t_2 \geq t_1 \geq t_0} dt_2 dt_1 e^{M^s(t-t_2)} M^a e^{M^s(t_2-t_1)} M^a e^{M^s(t_1-t_0)} X(t_0), \tag{93}$$

and so on. We finally get the pertubative solution of the composition $X(t)$:

$$X(t) = e^{M^s(t-t_0)}X(t_0) + \int_{t_0}^{t} du e^{M^s(t-u)} M^a e^{M^s(u-t_0)} X(t_0) + O(\epsilon^2). \tag{94}$$

The zero-order term $X^{(0)}(t)$ corresponds to the PR2 solution when there is no substitutional asymmetry $M^a = 0$. The first-order term $X^{(1)}(t)$ gives small corrections to the composition evolution when there is a small asymmetry $M^a \neq 0$.

Pertubative analysis in the (θ, S) coordinates

We can also perform the pertubative analysis directly in the (θ, S) coordinates, the computations are just slightly more involved. If we start with initial null skews $S(t_0) = 0$, the pertubative resolution of Eq. (63) gives the following time evolutions:

$$\theta(t) = e^{A(t-t_0)}\theta(t_0) + O(\epsilon^2), \tag{95}$$

$$S(t) = \int_{t_0}^{t} du e^{D(t-u)} C e^{A(u-t_0)}\theta(t_0) + O(\epsilon^2). \tag{96}$$

We can also use perturbation theory to find the composition at equilibrium. The method is similar, but instead of having differential equations we have algebraic equations to solve. The pertubative resolution of Eq. (64) gives the equilibrium values:

$$\theta^* = \theta_A + O(\epsilon^2), \tag{97}$$

$$S^* = -D^{-1}C\theta_A + O(\epsilon^2). \tag{98}$$

We can demonstrate that the pertubative expansions of $\theta(t)$ (Eq. (95)) and $S(t)$ (Eq. (96)) consistently converge towards the pertubative expansions of θ^* (Eq. (97)) and S^* (Eq. (98)), and this at all orders in the expansion parameter ϵ. At first order in the substitutional asymmetry, the GC content time evolution (Eq. (95)) and its equilibrium value (Eq. (97)) are not affected. The dynamic of the compositional skews in Eq.(96) is controlled by the time scales $\tau_D^{(1)}$ and $\tau_D^{(2)}$ (Eq. (75)) but also by the time scale τ_A (Eq. (73)). As expected the compositional skews depend linearly on the substitutional asymmetry through the matrix C in Eqs. (96) and (98). The skews at equilibrium (Eq. (98)) show that PR2 is explicitly broken when there is a substitutional asymmetry.

I.4.2 Impact of replication fork polarity, gene orientation and transcription rate

Working hypotheses

We now address the impact of replication fork polarity p, gene orientation $(\pm)$ and transcription rate α on the compositional asymmetry. To this purpose, we apply the pertubative analysis for substitution rates satisfying our model Eqs. (8) and (9). We recall that the coefficients $\tau_R^a, \tau_T^a, \tau_T^s$ of the model were all found to be small as compared to the $\tau_0^s + \tau_R^s$ substitution rate in the human genome (Eqs. (51)-(53)). We will therefore treat the $\tau_0^s + \tau_R^s$ substitution rates as order $O(1)$, and the $\tau_R^a, \tau_T^a, \tau_T^s$ coefficients as small perturbations of order $O(\epsilon)$. In our minimal model symmetrical substitutions rates follow Eq. (8), rewritten here for the reader's convenience:

$$\tau^s[p, \alpha, (\pm)] = \tau_0^s + \tau_R^s + \tau_T^s[\alpha], \tag{8}$$

which implies in turn the same decomposition for the symmetrical part M^s, and for the matrices A and D defined from M^s in Eq. (65). Similarly the Eq. (9) for the

substitutional asymmetry:

$$\tau^a[p,\alpha,(\pm)] = p\tau^a_R \pm \tau^a_T[\alpha] \tag{9}$$

implies in turn the same decomposition for the asymmetrical part M^a, and for the matrices B and C defined from M^a in Eq. (66).

Weak impact on the TA and GC contents

The pertubative resolution of Eq. (63) gives the following time evolution of the TA and GC contents (Eq. (95)):

$$\theta[p,\alpha,(\pm)](t) = \tilde{\theta}_0(t) + \theta_T[\alpha](t) + O(\epsilon^2), \tag{99}$$

where:

$$\tilde{\theta}_0(t) = e^{[A_0+A_R](t-t_0)}\,\theta(t_0), \tag{100}$$

$$\theta_T[\alpha](t) = \int_{t_0}^{t} du\, e^{[A_0+A_R](t-u)}\, A_T[\alpha]\,\tilde{\theta}_0(u). \tag{101}$$

The pertubative resolution of Eq. (64) yields the following equilibrium TA and GC contents (Eq. (97)):

$$\theta^*[p,\alpha,(\pm)] = \tilde{\theta}^*_0 + \theta^*_T[\alpha] + O(\epsilon^2), \tag{102}$$

where:

$$\tilde{\theta}^*_0 = \theta_{[A_0+A_R]}, \tag{103}$$

$$\theta^*_T[\alpha] = \tau_{[A_0+A_R]}\, A_T[\alpha]\,\tilde{\theta}^*_0. \tag{104}$$

As expected the GC content does not depend on replication fork polarity and gene orientation. Hence our minimal model (Eqs. (8) and (9)) do not provide a satisfactory treatment of the GC content evolution. The GC content is almost equal to its PR2 value and depends on all the variables that determine the symmetrical substitution rates, and they are many. More relevant explanatory variables, such as recombination rate (Duret and Arndt 2008), should be considered to account for the GC content evolution. Our model only predicts a slight dependence of the GC content upon transcription rate through the $\theta_T[\alpha]$ coefficient. This change is however presumably small as compared to the variation of the GC content with recombination rate.

The skews can be decomposed into transcription- and replication-associated components

The pertubative resolution of Eq. (63), with initial null skews $S(t_0) = 0$, gives the following time evolution of the TA and GC skews (Eq. (96)):

$$\boxed{S[p,\alpha,(\pm)](t) = pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2)}, \tag{105}$$

where:

$$S_R(t) = \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, C_R\, \tilde{\theta}_0(u), \tag{106}$$

$$S_T[\alpha](t) = \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, C_T[\alpha]\, \tilde{\theta}_0(u). \tag{107}$$

The pertubative resolution of Eq. (64) gives the following equilibrium TA and GC skews (Eq. (98)):

$$\boxed{S^*[p,\alpha,(\pm)] = pS_R^* \pm S_T^*[\alpha] + O(\epsilon^2)}, \tag{108}$$

where:

$$S_R^* = -[D_0 + D_R]^{-1}\, C_R\, \tilde{\theta}_0^*, \tag{109}$$

$$S_T^*[\alpha] = -[D_0 + D_R]^{-1}\, C_T[\alpha]\, \tilde{\theta}_0^*. \tag{110}$$

Therefore we recover for the compositional asymmetry the same additive decomposition into a replication and a transcription contribution, as originally hypothesized for the substitutional asymmetry (Eq. (9)).

☛ *In our minimal model, the compositional asymmetry can be decomposed into transcription- and replication-associated components. The replication-associated compositional asymmetry is proportional to the replication fork polarity, the transcription-associated one increases in magnitude with transcription rate and changes sign with gene orientation.*

Numerical test. In Fig. 9, we compare the exact and pertubative solutions for the toy model substitution rate matrix:

$$M[p] = M_0^s + M_R^s + pM_R^a = M_0^s + M_R^s + (p/\bar{p})\,\bar{p}M_R^a, \tag{111}$$

where $M_0^s + M_R^s$ and $\bar{p}M_R^a$ are the substitution rate matrices computed in the human genome as explained in Section I.2.3 (Eqs. (43) and (44)). To express the substitution rates in per bp per Myrs units, I used 5 Myrs as an estimation of the human-chimp divergence. According to our minimal model Eq. (8) and (9), $M[p]$ is equal to the substitution rate matrix obtained in intergenic regions of replication fork polarity p. As shown in Fig. 9A,B, the pertubative solutions for the time evolution of θ_{GC}, θ_{TA}, and S_{GC}, S_{TA} are indistinguishable from the exact solutions. The pertubative solutions for the equilibrium values are also indistinguishable from the exact solutions (Fig. 9C,D). For the given experimental substitution rate matrices, the first-order correction is already an excellent approximation, and there is no need to take into account higher order corrections. As predicted by Eq. (108) the skews at equilibrium are proportional to p (Fig. 9D). Similarly, as predicted by Eq. (102), the TA and GC contents at equilibrium do not depend upon p (Fig. 9C). As governed by Eq. (99), the time evolution of the GC content (Fig. 9A) is not affected by the

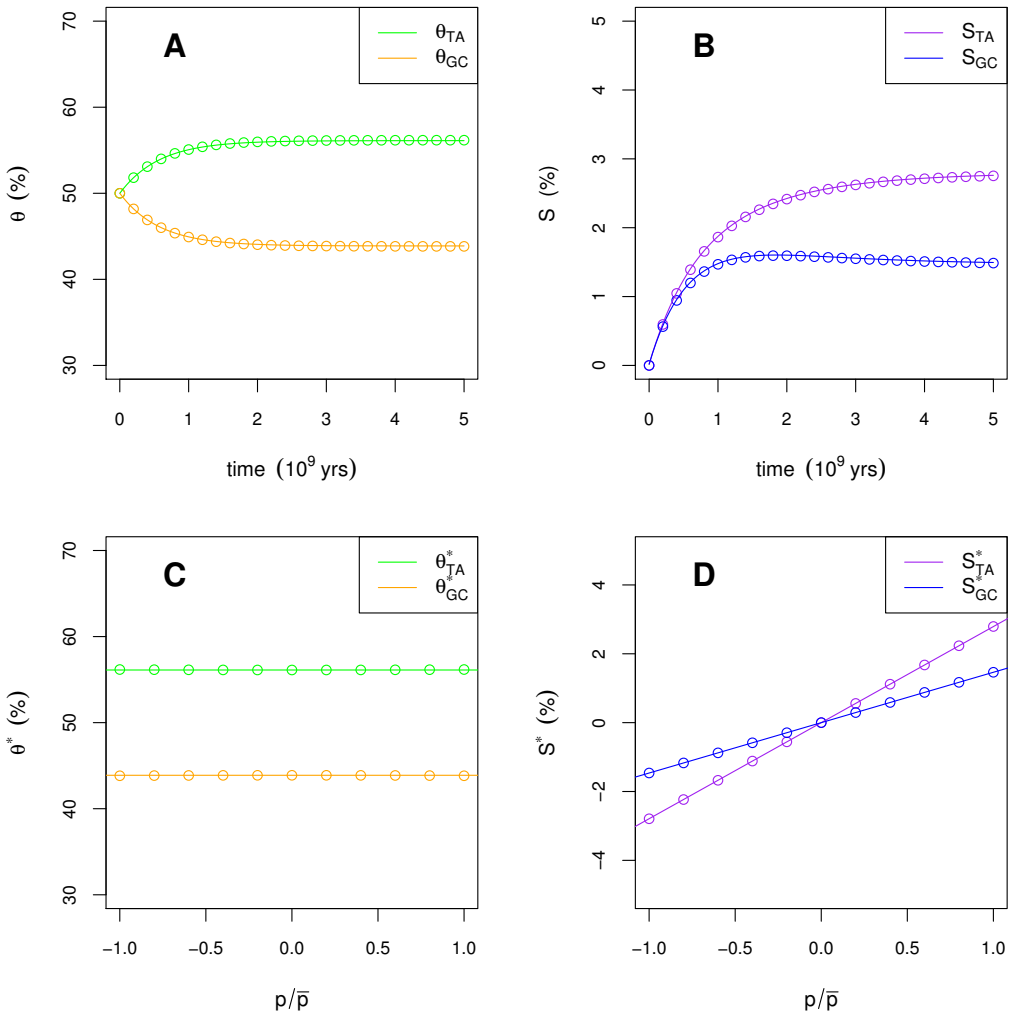

Figure 9: **DNA composition evolution in presence of strand asymmetry: comparison of exact and pertubative solutions.** Toy model substitution rate matrix $M[p] = M_0^s + M_R^s + pM_R^a$ depends on the replication fork polarity p (Eq. (111)). Exact solution is represented as circles, pertubative solution as solid line. Time evolution of the TA and GC contents (A) and of the TA and GC skews (B) for the initial conditions $\theta_{TA}(0) = \theta_{GC}(0) = 50\%$ and $S_{TA}(0) = S_{GC}(0) = 0\%$, and $p = \bar{p}$. Equilibrium TA and GC contents (C) and TA and GC skews (D) versus p.

substitutional asymmetry $\bar{p}M_R^a$, which explains that we recover the time evolution under the symmetrical matrix $M_0^s + M_R^s$ previously shown in Fig. 8A. According to Eq. (106), the skews converge towards their equilibrium values with time scales $\tau_{A_0+A_R}$, $\tau^{(1)}_{D_0+D_R}$ and $\tau^{(2)}_{D_0+D_R}$. For the $M_0^s + M_R^s$ matrix given in Eq. (43), these time scales are equal to 568 Myrs, 566 Myrs and 1398 Myrs. Hence the convergence of the skews towards their equilibrium values is a very long process.

Comment on the long term memory of the initial skews. If the skews are initially null, they increase according to Eqs. (105)-(107) to ultimately reach their equilibrium values. Depending linearly on the substitutional asymmetry (Eqs. (108)-(110)), the equilibrium skews are of order $O(\epsilon)$. Therefore under a small substitutional asymmetry, the skews cannot reach values larger than $O(\epsilon)$. Hence in our pertubative analysis, if we take initial non null skews $S(t_0) \neq 0$, we will nonetheless assume that they are of order $O(\epsilon)$. Under this assumption the time evolution of the skews is governed by:

$$S[p, \alpha, (\pm)](t) = S_{\text{ini}}(t) + pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2), \tag{112}$$

where:

$$S_{\text{ini}}(t) = e^{[D_0+D_R](t-t_0)} S(t_0). \tag{113}$$

The skews at equilibrium are of course unchanged as they do not depend on the initial composition. But if the skews have not reached equilibrium, their time evolution keeps memory of the initial skews through the additional term $S_{\text{ini}}(t)$. We recognize this term as the PR2 solution (Eq. (81)) under the symmetrical matrix $M_0^s + M_R^s$. As we have already discussed for the PR2 solution, this term slowly decays towards zero with time scales $\tau^{(1)}_{[D_0+D_R]}$ and $\tau^{(2)}_{[D_0+D_R]}$.

I.4.3 Accounting for neighbor-dependent substitution rates

We first exposed DNA composition evolution (Eq. (54)) in the case of time homogeneous and neighbor-independent substitution rates. However, substitution rates do depend on the flanking nucleotides (Hwang and Green 2004; Arndt and Hwa 2005). In vertebrates, CpG sites are hypermutable, the substitution rate $C \to T$ depends dramatically (ten fold) on whether the cytosine belongs to a CG dinucleotide or not (Hess et al. 1994; Arndt and Hwa 2005). The $r = CpG \to TpG$ and its reverse complementary $r^c = CpG \to CpA$ are by far the principal neighbor-dependent substitutions rates in the human genome (Hess et al. 1994; Arndt and Hwa 2005). The neighbor dependency is difficult to handle mathematically and we will follow the model introduced in (Arndt et al. 2003). First we expose the model in its general form and some of its general properties (odd ratios, PR2). Then we specify the model when only the neighbor-dependent $r = CpG \to TpG$ substitution rate is taken into account. Finally we perform the pertubative analysis for this simple case.

☛ *As far as the pertubative analysis is concerned, the neighbor dependency does not change anything to the conclusions made in the neighbor-independent model.*

Neighbor-dependent model

Following (Arndt et al. 2003) the sequence evolves by two processes:

* single-nucleotide (neighbor-independent) substitution rates;
* dinucleotide (neighbor-dependent) substitution rates.

We define a single-nucleotide substitution rate matrix M and a dinucleotide substitution rate matrix Q. We keep the same definition for M as before: for nucleotides $i, a \in \{T, A, G, C\}$, the element M_{ia} is the (neighbor-independent) substitution rate $a \to i$. For $i, j \in \{T, A, G, C\}$ and $a, b \in \{T, A, G, C\}$, the element $Q_{ij,ab}$ is the (neighbor-dependent) substitution rate $ab \to ij$. Sums over rows are null for both M and Q:

$$\sum_i M_{ia} = 0 \quad \text{and} \quad \sum_{ij} Q_{ij,ab} = 0. \tag{114}$$

The evolution of the composition is now given by:

$$\frac{d}{dt}X_i(t) = \sum_a M_{ia}X_a(t) + \sum_{abc} Q_{ia,bc}X_{bc}(t) + \sum_{abc} Q_{ai,bc}X_{bc}(t), \tag{115}$$

where $X_i(t)$ is still the frequency (or probability) of nucleotides i at time t, and $X_{ij}(t)$ the frequency (or probability) of dinucleotides ij at time t. As before the first term accounts for the single nucleotide substitutions. The second and third terms account for all the dinucleotide substitutions that give rise to the nucleotide i. A dinucleotide bc can substitute into a dinucleotide ia (second term) with a nucleotide i on the first base, or into a dinucleotide ai (third term) with a nucleotide i on the second base.

The neighbor-dependent model is not a closed system

According to Eq. (115), the time evolution of the composition in nucleotides $X_i(t)$ depends on the composition in dinucleotides $X_{ij}(t)$. Hence, in order to solve this equation, we need to determine the composition in dinucleotides. The evolution of the composition in dinucleotides is given by:

$$\begin{aligned}\frac{d}{dt}X_{ij}(t) = &\sum_{ab}[M \otimes \mathbb{I} + \mathbb{I} \otimes M]_{ij,ab}X_{ab}(t) \\ &+ \sum_{ab} Q_{ij,ab}X_{ab}(t) + \sum_{abc} Q_{ja,bc}X_{ibc}(t) + \sum_{abc} Q_{ai,bc}X_{bcj}(t),\end{aligned} \tag{116}$$

where $X_{ijk}(t)$ is the frequency (or probability) of trinucleotides ijk at time t, $\otimes$ is the Kronecker tensor product, and $\mathbb{I}$ is the 4×4 identity matrix. The first term accounts for single nucleotide substitutions, the three last terms for dinucleotide substitutions.

Dinucleotide substitutions can give rise to the dinucleotide ij in different ways. Of course a dinucleotide ab can substitute into the dinucleotide ij (second term). But we can also get the dinucleotide ij if a trinucleotide ibc containing a i on the first base, undergoes a substitution $bc \to ja$ and becomes a trinucleotide ija (third term). We also get the dinucleotide ij if a trinucleotide bcj containing a j on the third base, undergoes a substitution $bc \to ai$ and becomes a trinucleotide aij (fourth term).

The time evolution for the composition in dinucleotides (Eq. (116)) therefore depends on the composition in trinucleotides, whose time evolution will in turn depend on the composition in quadrinucleotides, and so on. We are thus faced with an infinite hierarchy of equations (Arndt et al. 2003). This is the main mathematical difficulty of the neighbor-dependent model, as the infinite hierarchy of equations cannot be solved exactly in general. To my knowledge, only Bérard et al. (2008) succeeded in proving exact results regarding the neighbor-dependent model. The authors in (Arndt et al. 2003) proposed instead to truncate the infinite hierarchy using the **two-cluster approximation**:

$$X_{ijk} \simeq \frac{X_{ij} X_{jk}}{X_j}. \tag{117}$$

This approximation is equivalent to state that the sequence is a first-order Markov chain (in genomic position). Then Eq. (116) is closed, with trinucleotide frequencies given by Eq. (117). In further numerical examples, we will use the two-cluster approximation to compute the time evolution of the dinucleotide frequencies.

Remark. Note that the infinite hierarchy of equations is in fact highly redundant. The composition in nucleotides can be obtained from the composition in dinucleotides, which can be obtained from the composition in trinucleotides, and so on:

$$X_i = \sum_a X_{ia} = \sum_a X_{ai}, \quad X_{ij} = \sum_a X_{ija} = \sum_a X_{aij}, \quad \dots \tag{118}$$

Using Eq. (118) between compositions in nucleotides and dinucleotides and the time evolution Eq. (116) for the composition in dinucleotides, one recovers the time evolution Eq. (115) for the composition in nucleotides. Similarly, the time evolution for composition in n-nucleotides implies all the time evolutions for lower numbers of nucleotides through relations like Eq. (118).

Odds ratios and PR2

When there are no neighbor-dependent substitution rates, *i.e.* when $Q = 0$, the solution of Eq. (116) is $X_{ij}(t) = X_i(t) X_j(t)$. The observed frequencies of dinucleotides X_{ij} are then equal to their expected value $X_i X_j$. Odds ratios (or observed over expected values) $\rho_{ij} = \frac{X_{ij}}{X_i X_i}$ clearly different from 1 have been an indication, directly derived from the sequence, that neighbor-independency does not hold in many genomes (Burge et al. 1992). Regarding strand symmetry, one can extend

PR1 and PR2 symmetries to dinucleotide frequencies. We recall that the reverse complementary of a dinucleotide ij is the dinucleotide $j^c i^c$. Hence four dinucleotides are their own reverse complementary:

$$(TA)^c = TA, \quad (AT)^c = AT, \quad (GC)^c = GC \quad \text{and} \quad (CG)^c = CG. \tag{119}$$

The dinucleotide frequencies computed on the complementary strand are given by reverse complementarity:

$$X_{ij}^c = X_{(ij)^c} = X_{j^c i^c}. \tag{120}$$

We note that the CG frequency is strand-symmetric: $X_{CG}^c = X_{CG}$. The neighbor-dependent substitution rate matrix computed on the complementary strand is also given by reverse complementarity:

$$Q_{ij,ab}^c = Q_{(ij)^c,(ab)^c} = Q_{j^c i^c, b^c a^c}. \tag{121}$$

We can decompose Q into symmetrical and asymmetrical parts under strand exchange symmetry:

$$Q = Q^s + Q^a, \quad \text{with} \quad Q^s = \frac{Q + Q^c}{2}, \quad Q^a = \frac{Q - Q^c}{2}. \tag{122}$$

In the neighbor-dependent model, PR2 extends to the dinucleotides frequencies. Under symmetrical substitution rates (PR1), the frequencies of reverse complementary dinucleotides are equal at equilibrium (PR2):

$$\boxed{\text{if} \quad M = M^s \text{ and } Q = Q^s \text{ (PR1)} \quad \text{then} \quad X_{ij}^* = X_{j^c i^c}^* \text{ (PR2)}} \tag{123}$$

Focusing on the $CpG \to TpG$ substitution only

As previously mentioned, the $r = CpG \to TpG$ and its reverse complementary $r^c = CpG \to CpA$ are by far the principal neighbor-dependent substitutions rates in the human genome (Hess et al. 1994; Arndt and Hwa 2005). Following (Arndt et al. 2003), we take all the elements of the matrix Q null except:

$$Q_{TG,CG} = r = r^s + r^a, \quad Q_{CA,CG} = r^c = r^s - r^a, \quad Q_{CG,CG} = -2r^s. \tag{124}$$

The evolution of the composition then simplifies to:

$$\frac{d}{dt}X(t) = MX(t) + X_{CG}(t) \begin{pmatrix} r^s + r^a \\ r^s - r^a \\ -r^s + r^a \\ -r^s - r^a \end{pmatrix} \tag{125}$$

in the $\{T, A, G, C\}$ coordinates and to:

$$\frac{d}{dt}Y(t) = NY(t) + X_{CG}(t) \begin{pmatrix} 2r^s \\ -2r^s \\ 2r^a \\ 2r^a \end{pmatrix} \tag{126}$$

in the $\{\theta_{TA}, \theta_{GC}, S_{TA}, S_{GC}\}$ coordinates. Therefore we recover the time evolutions Eqs. (54) and (63) with a source term depending on the CG frequency. The composition at equilibrium is given by:

$$NY^* + X^*_{CG} \begin{pmatrix} 2r^s \\ -2r^s \\ 2r^a \\ 2r^a \end{pmatrix} = 0 \quad \text{and} \quad \theta^*_{TA} + \theta^*_{GC} = 1. \tag{127}$$

Therefore a source term depending on the equilibrium CG frequency is also added to the equilibrium composition Eq. (63).

Numerical test of odds ratio and PR2

To illustrate the properties of the neighbor dependency, we investigated observed dinucleotide frequencies X_{ij} versus expected dinucleotide frequencies X_iX_j at equilibrium for four different models (Fig. 10). The dinucleotide frequencies at equilibrium were determined by integrating numerically the differential Eq. (116) with the two-cluster approximation (Eq. (117)). The four models correspond to special cases of the model Eq. (126), with or without neighbor dependency, and evolving under PR1 or PR1 breaking. For the two neighbor-independent models ($r = 0$ in Fig. 10A,B), observed dinucleotide frequencies are equal to their expected values. On the opposite, for the neighbor-dependent models ($r \neq 0$ in Fig. 10C,D) the odd ratios are clearly different from 1. As expected the odd ratio of the CG dinucleotide is decreased, whereas the odds ratio of the TG and CA dinucleotides are increased. For models under PR1 ($M = M^s, r = r^s$ in Fig. 10A,C), observed frequencies of reverse complementary dinucleotides are equal, as their expected values. The composition does satisfy PR2 ($[G] = [C]$ and $[T] = [A]$), as verified in Fig. 10A,C, the values are clustered into three groups along the x-axis, corresponding to the only three different expected dinucleotide frequencies $[X][Y]$ values ($[G][G]$, $[T][T]$ or $[G][T]$). Furthermore, for the neighbor-dependent model under PR1 (Fig. 10C), the observed dinucleotide frequencies are not equal to their expected values, but the observed frequencies of reverse complementary dinucleotides are nonetheless equal. For instance in Fig. 10C, TG and CA observed frequencies are equal $[TG] = [(TG)^c] = [CA]$ whereas the CG observed frequency is only equal to itself $[CG] = [(CG)^c]$. Of course when the PR1 symmetry is broken (Fig. 10B,D), the PR2 symmetry is broken for both expected and observed dinucleotide frequencies.

Pertubative resolution of the neighbor-dependent model

The neighbor-dependent case (Eq. (126)) differs from the neighbor-independent model (Eq. (63)) by the additional source term related to the CG frequency. The neighbor dependency does not change anything to our conclusions, but a source term related to the CG frequency is systematically added.

Pertubative analysis in the (θ, S) coordinates. Here the symmetrical parts M^s and r^s

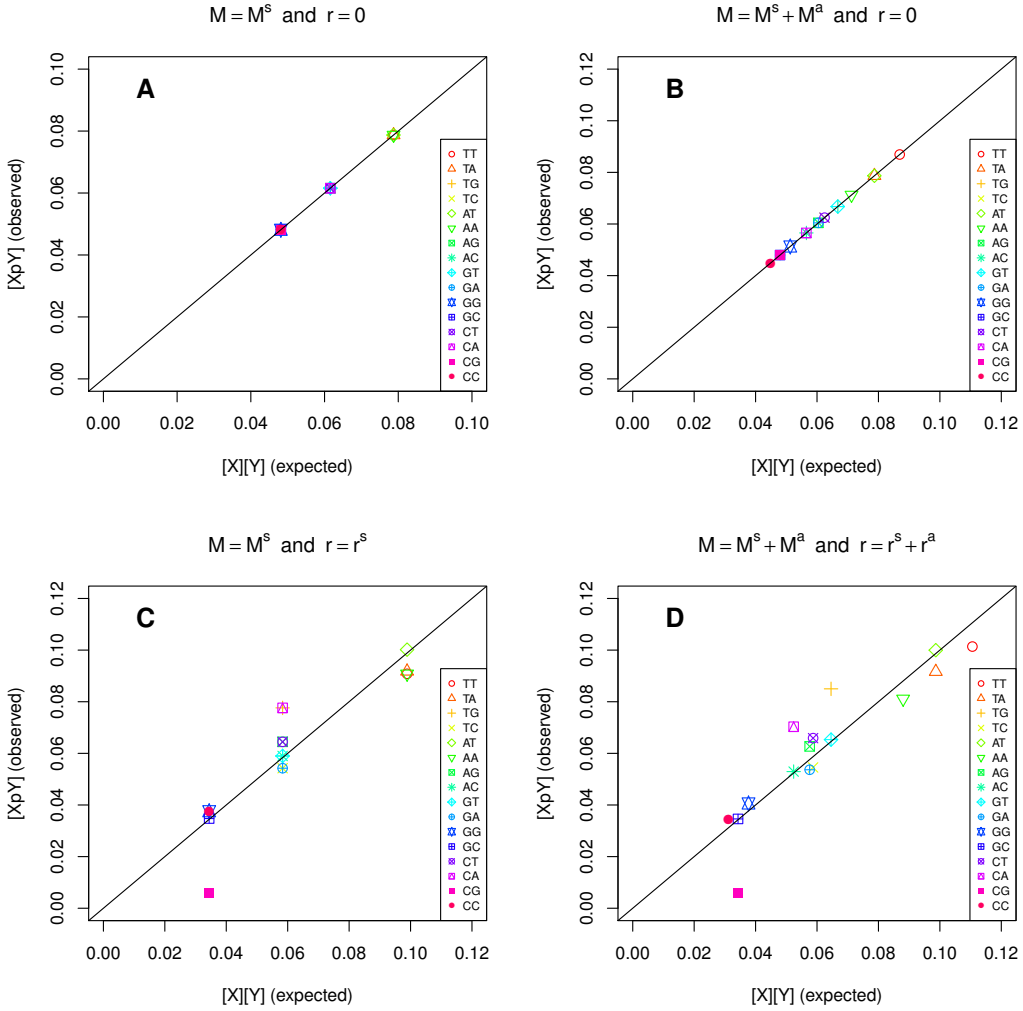

Figure 10: **Observed and expected dinucleotide frequencies at equilibrium.** (A) Neighbor-independent model ($r = 0$) with symmetrical substitution rates ($M = M^s$). (B) Neighbor-independent model ($r = 0$) with substitutional asymmetry ($M^a \neq 0$). (C) Neighbor-dependent model ($r \neq 0$) with symmetrical substitution rates ($M = M^s$, $r = r^s$). (D) Neighbor-dependent model ($r \neq 0$) with substitutional asymmetry ($M^a \neq 0$, $r^a \neq 0$).

are considered $O(1)$, while the asymmetrical parts M^a and r^a are considered $O(\epsilon)$. If we start with initial null skews $S(t_0) = 0$, the pertubative resolution of Eq. (126) gives the following time evolutions:

$$\theta(t) = e^{A(t-t_0)}\,\theta(t_0) + \int_{t_0}^{t} du\, e^{A(t-u)}\, X_{CG}(u) \begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + O(\epsilon^2), \tag{128}$$

$$\begin{aligned} S(t) = & \int_{t_0}^{t} du\, e^{D(t-u)}\, C\, e^{A(u-t_0)}\, \theta(t_0) \\ & + \int_{t_0}^{t} du\, e^{D(t-u)}\, C \int_{t_0}^{u} dv\, e^{A(u-v)}\, X_{CG}(v) \begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} \\ & + \int_{t_0}^{t} du\, e^{D(t-u)}\, X_{CG}(u) \begin{pmatrix} 2r^a \\ 2r^a \end{pmatrix} + O(\epsilon^2). \end{aligned} \tag{129}$$

The pertubative resolution of Eq. (127) yields the equilibrium values:

$$\theta^* = \theta_A + \tau_A X^*_{CG} \begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + O(\epsilon^2), \tag{130}$$

$$S^* = -D^{-1} \left\{ C\theta_A + C\tau_A X^*_{CG} \begin{pmatrix} 2r^s \\ -2r^s \end{pmatrix} + X^*_{CG} \begin{pmatrix} 2r^a \\ 2r^a \end{pmatrix} \right\} + O(\epsilon^2). \tag{131}$$

We note that both the GC content and the skews are affected by the neighbor-dependent rate r.

Weak impact on the TA and GC contents. Here we assume that the substitution rates, and in particular the r substitution rate and the substitution rate matrix M, follow our minimal model Eqs. (8) and (9). We recall that the coefficients $\tau^a_R, \tau^s_T, \tau^a_T$ are $O(\epsilon)$. The pertubative resolution of Eq. (126) gives the following time evolution of the TA and GC contents:

$$\theta[p, \alpha, (\pm)](t) = \tilde{\theta}_0(t) + \theta_T[\alpha](t) + O(\epsilon^2), \tag{132}$$

where:

$$\begin{aligned} \tilde{\theta}_0(t) = & e^{[A_0+A_R](t-t_0)}\, \theta(t_0) \\ & + \int_{t_0}^{t} du\, e^{[A_0+A_R](t-u)}\, X_{CG}(u) \begin{pmatrix} 2(r^s_0 + r^s_R) \\ -2(r^s_0 + r^s_R) \end{pmatrix}, \end{aligned} \tag{133}$$

$$\begin{aligned} \theta_T[\alpha](t) = & \int_{t_0}^{t} du\, e^{[A_0+A_R](t-u)}\, A_T[\alpha]\, \tilde{\theta}_0(u) \\ & + \int_{t_0}^{t} du\, e^{[A_0+A_R](t-u)}\, X_{CG}(u) \begin{pmatrix} 2r^s_T[\alpha] \\ -2r^s_T[\alpha] \end{pmatrix}. \end{aligned} \tag{134}$$

The pertubative resolution of Eq. (127) gives the following equilibrium TA and GC contents:

$$\theta^*[p, \alpha, (\pm)] = \tilde{\theta}^*_0 + \theta^*_T[\alpha] + O(\epsilon^2), \tag{135}$$

where:

$$\tilde{\theta}_0^* = \theta_{[A_0+A_R]} + \tau_{[A_0+A_R]}\, X_{CG}^* \begin{pmatrix} 2(r_0^s + r_R^s) \\ -2(r_0^s + r_R^s) \end{pmatrix}, \tag{136}$$

$$\theta_T^*[\alpha] = \tau_{[A_0+A_R]} \left\{ A_T[\alpha]\, \tilde{\theta}_0^* + X_{CG}^* \begin{pmatrix} 2r_T^s[\alpha] \\ -2r_T^s[\alpha] \end{pmatrix} \right\}. \tag{137}$$

As compared to the neighbor-independent model Eqs. (99)-(104), we note that the neighbor-dependent rate r impacts on both the $\tilde{\theta}_0$ and $\theta_T[\alpha]$ coefficients in Eqs. (132)-(137).

The skews still decompose into transcription- and replication-associated components. We still assume that the substitution rates obey our minimal model Eqs. (8) and (9). The pertubative resolution of Eq. (63), with initial null skews $S(t_0) = 0$, yields the following time evolution of the TA and GC skews:

$$\boxed{S[p, \alpha, (\pm)](t) = pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2)}, \tag{138}$$

where:

$$\begin{aligned} S_R(t) = & \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, C_R\, \tilde{\theta}_0(u) \\ & + \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, X_{CG}(v) \begin{pmatrix} 2r_R^a \\ 2r_R^a \end{pmatrix}, \end{aligned} \tag{139}$$

$$\begin{aligned} S_T[\alpha](t) = & \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, C_T[\alpha]\, \tilde{\theta}_0(u) \\ & + \int_{t_0}^{t} du\, e^{[D_0+D_R](t-u)}\, X_{CG}(v) \begin{pmatrix} 2r_T^a[\alpha] \\ 2r_T^a[\alpha] \end{pmatrix}. \end{aligned} \tag{140}$$

The pertubative resolution of Eq. (64) gives the equilibrium TA and GC skew values:

$$\boxed{S^*[p, \alpha, (\pm)] = pS_R^* \pm S_T^*[\alpha] + O(\epsilon^2)}, \tag{141}$$

where:

$$S_R^* = -[D_0 + D_R]^{-1} \left\{ C_R\, \tilde{\theta}_0^* + X_{CG}^* \begin{pmatrix} 2r_R^a \\ 2r_R^a \end{pmatrix} \right\}, \tag{142}$$

$$S_T^*[\alpha] = -[D_0 + D_R]^{-1} \left\{ C_T[\alpha]\, \tilde{\theta}_0^* + X_{CG}^* \begin{pmatrix} 2r_T^a[\alpha] \\ 2r_T^a[\alpha] \end{pmatrix} \right\}. \tag{143}$$

Therefore we recover for the compositional asymmetry the same decomposition as for the substitutional asymmetry Eq. (9). From the comparison with the corresponding Eqs. (105)-(110) of the neighbor-independent model, we note that the neighbor-dependent rate r also impacts on the S_R and $S_T[\alpha]$ coefficients.

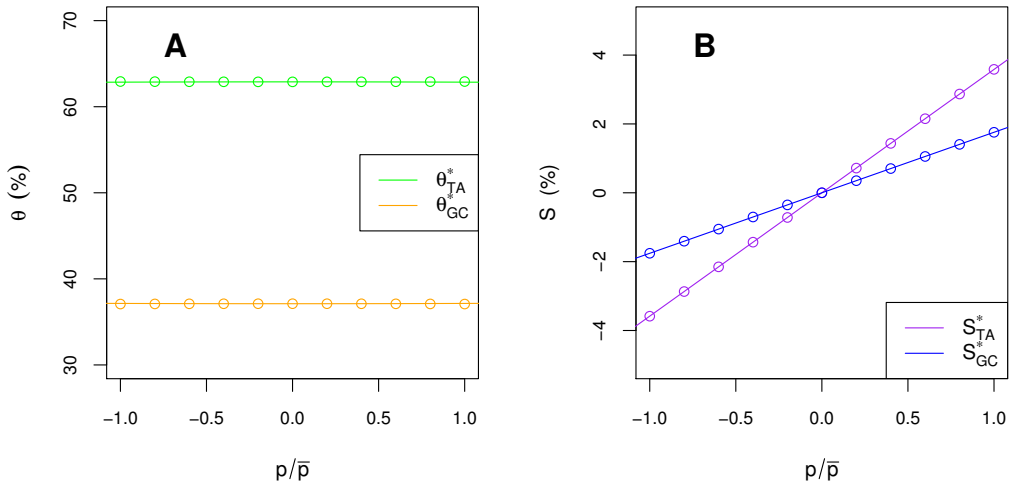

Figure 11: **Comparison of the exact and pertubative solutions in the neighbor-dependent model.** The substitution rate matrix $M[p] = M_0^s + pM_R^a$ and the neighbor-dependent substitution rate $r[p] = r_0^s + pr_R^a$ depend on the replication fork polarity p. Exact solution is represented as circles, pertubative solution as solid line. (A) Equilibrium TA and GC contents versus p. (B) Equilibrium TA and GC skews versus p.

Numerical test. In Fig. 11, we compare the exact and pertubative solutions for the toy model substitution rates matrix:

$$M[p] = M_0^s + M_R^s + pM_R^a = M_0^s + M_R^s + (p/\bar{p})\,\bar{p}M_R^a, \tag{144}$$

when taking into account the toy model neighbor-dependent substitution rate:

$$r[p] = r_0^s + r_R^s + pr_R^a = r_0^s + r_R^s + (p/\bar{p})\,\bar{p}r_R^a, \tag{145}$$

where the experimental matrices $M_0^s + M_R^s$ and $\bar{p}M_R^a$ are given in Eqs. (43) and (44), and the experimental neighbor-dependent rate $r_0^s + r_R^s$ and $\bar{p}r_R^a$ in Eq. (50). According to our minimal model Eq. (8), these substitution rates are equal to those obtained in intergenic regions of replication fork polarity p. As shown in Fig. 11, the pertubative solutions for the equilibrium values are indistinguishable from the exact solutions. As predicted by Eq. (141) the skews at equilibrium are proportional to p (Fig. 11B). In contrast, as predicted by Eq. (135), the TA and GC contents at equilibrium do not depend upon p (Fig. 11A). When comparing these values to the ones previously obtained with the neighbor-independent model ($r = 0$, Fig. 9), we see that the neighbor-dependent rate r does impact on both the skews and the GC content.

I.4.4 Time dependency of substitution rates

Substitutional patterns have been also shown to depend on time (Hwang and Green 2004; Mugal et al. 2009). In this case all substitution rates, and all the parameters

derived from them, depend explicitly on time. Importantly, on the contrary to the time homogeneous case, the nucleotide composition does not necessarily converge towards its equilibrium value. PR2 formally extends to the time-dependent case (Lobry and Lobry 1999). For the pertubative analysis of the compositional asymmetry (violation of PR2), I briefly indicate how to formally take into account the time dependency.

☛ *This subsection will not lead to any practical, or even numerical, applications.*

Equilibrium composition interpreted as the current direction of evolution

If we take into account the time dependency of the substitution rates, the neighbor-independent model evolution Eq.(54) becomes:

$$\frac{d}{dt}X(t) = M(t)X(t). \tag{146}$$

In general the solution of Eq. (146) does not converge. At each time t, the composition starts converging towards the equilibrium value $X^*(t)$, defined from the matrix $M(t)$. As the equilibrium composition $X^*(t)$ changes over time, the composition $X(t)$ may never actually reach an equilibrium state. In this perspective the equilibrium composition $X^*(t)$ gives the current direction of evolution, not the long term asymptotic value (that may even not exist) of the composition. With this interpretation in mind, the pertubative solutions for the equilibrium GC content and the skews are still valid, but they of course depend explicitly on time.

PR2 is still valid for time-dependent substitution rates

PR2 was first proved (Lobry 1995) for the equilibrium composition for time-independent substitution rates (Eq. (82)). But if the composition never reaches the equilibrium state, are we certain that PR2 is still satisfied ? Under symmetrical substitution rates (PR1), the evolution of the skews decouples from the evolution of the TA and GC contents (Lobry and Lobry 1999):

$$\frac{d}{dt}\theta(t) = A(t)\theta(t), \tag{147}$$

$$\frac{d}{dt}S(t) = D(t)S(t). \tag{148}$$

Note that Eqs. (147) and (148) are the equivalent of Eqs. (78) and (79) for time-dependent substitution rates. Lobry and Lobry (1999) showed that the GC content on one side and the skews on the other have distinct long term behaviour. The GC content generally never reaches equilibrium, $GC^*(t)$ only gives the ever changing direction of evolution. On the opposite the skews always decay towards zero, as $S^*(t) = 0$ gives at all times the same direction of evolution. Therefore whatever the time-dependent substitutional pattern, under symmetrical substitution rates (PR1), the nucleotide composition will always satisfy PR2 asymptotically.

Solving time-dependent differential equations

If we take into account the time dependency of the substitution rates, the neighbor-dependent model Eq. (126) may be rewritten as

$$\frac{d}{dt}Y(t) = N(t)Y(t) + X_{CG}(t)\begin{pmatrix} 2r^s(t) \\ -2r^s(t) \\ 2r^a(t) \\ 2r^a(t) \end{pmatrix} \tag{149}$$

in the $\{\theta_{TA}, \theta_{GC}, S_{TA}, S_{GC}\}$ coordinates. This models falls into the class of time-dependent linear differential equations:

$$\frac{d}{dt}Y(t) = N(t)Y(t) + Z(t), \tag{150}$$

where $Z(t)$ is a source term. When solving pertubatively Eq. (149), we systematically encounter differential equations of this class. To solve Eq. (150) we need to introduce the time-ordered exponential (Van Kampen 2007):

$$Te^{\int_{t_0}^{t} duN(u)} = 1 + \sum_{n\geq 1} \int_{t\geq t_n \geq \cdots \geq t_1 \geq t_0} dt_n \ldots dt_1 N(t_n) \cdots N(t_1). \tag{151}$$

Importantly the time-ordered exponential satisfies the following properties:

$$\frac{d}{dt}\, Te^{\int_{t_0}^{t} duN(u)} = N(t)\, Te^{\int_{t_0}^{t} duN(u)}, \quad \text{and} \quad Te^{\int_{t_0}^{t} duN(u)}\Big|_{t=t_0} = \mathbb{I}, \tag{152}$$

where $\mathbb{I}$ is the identity matrix. The solution of Eq. (150) with the initial condition $Y(t_0)$ at t_0 is given by:

$$Y(t) = Te^{\int_{t_0}^{t} duN(u)} Y(t_0) + \int_{t_0}^{t} du\, Te^{\int_{t_0}^{u} dvN(v)} Z(u). \tag{153}$$

Proof. The proof is very simple. The solution Eq. (153) satisfies both Eq. (150) and the initial condition thanks to the properties of the time-ordered exponential (Eq. (152)). Uniqueness of the solution is ensured by the Cauchy-Lipschitz theorem.

We recall that the solution of differential Eq. (150) in the time-independent case $N(t) = N$ is given by:

$$Y(t) = e^{(t-t_0)N} Y(t_0) + \int_{t_0}^{t} du\, e^{(u-t_0)N} Z(u). \tag{154}$$

Consistently the time-ordered exponential reduces to the ordinary exponential in the time-independent case:

$$Te^{\int_{t_0}^{t} duN(u)} = e^{(t-t_0)N} \quad \text{when} \quad N(t) = N. \tag{155}$$

Therefore the resolution of the time-dependent differential equation actually amounts to replace the ordinary exponential in Eq. (154) by the time-ordered exponential in Eq. (153).

Pertubative analysis of the compositional asymmetry with time-dependent substitution rates

All the result derived in the previous pertubative analyses are extended to the time-dependent case if we systematically replace ordinary exponentials by time-ordered exponentials. For example, Eq. (128) becomes in the time-dependent case:

$$\theta(t) = Te^{\int_{t_0}^{t} duA(u)}\theta(t_0) + \int_{t_0}^{t} du\, Te^{\int_{u}^{t} dvA(v)} X_{CG}(u) \begin{pmatrix} 2r^s(u) \\ -2r^s(u) \end{pmatrix} + O(\epsilon^2). \quad (156)$$

In the time-dependent case the skews can still be decomposed into a transcription- and a replication-associated contribution. The proportionality of the replication-associated skew with the replication fork polarity further depends on whether the replication fork polarity changes over time or not. In our minimal model Eq. (9), the replication-associated substitutional asymmetry can change over time either because the replication fork polarity changes or because the coefficient τ_R^a changes. We could imagine for instance that the error spectra of the DNA polymerase, which affect τ_R^a (Eq. (14)), change over evolutionary time scales. We could also invoke that the replication program, which determines p, undergoes major changes over evolutionary time scales. If the replication fork polarity changes over time, we are no longer certain of the proportionality between the skew and the replication fork polarity. On the opposite, if the replication fork polarity is relatively constant, but the coefficient τ_R^a is time-dependent, then the proportionality will still hold.

I.5 Compositional asymmetry

Definition of the compositional skews

The compositional asymmetry is measured by the compositional skews (Eq. (61)):

$$S_{TA} = [T] - [A] \quad \text{and} \quad S_{GC} = [G] - [C], \quad (157)$$

where $[i]$ denotes the frequency of nucleotide $i \in \{T, A, G, C\}$ in the DNA sequence. The skew S_{TA} (resp. S_{GC}) is also equal to the difference in frequencies of T (resp. G) between the two DNA strands, hence its ability to measure compositional asymmetry. In the biological literature the skews are often normalized by the GC and TA contents (Brodie of Brodie et al. 2005; Touchon et al. 2005):

$$S_{TA} = \frac{[T] - [A]}{[T] + [A]} \quad \text{and} \quad S_{GC} = \frac{[G] - [C]}{[G] + [C]}. \quad (158)$$

In the human genome, as the TA and GC skews correlate (Touchon et al. 2003), the total skew defined as the sum of the TA and GC skews is also often considered. In the following S will denote generically the compositional skews, no matter their definitions.

The skew decomposes into transcription- and replication-associated components

We showed in Sections I.3 and I.4 that a substitutional asymmetry (breaking of PR1) in turn generates a compositional asymmetry (breaking of PR2). For substitution rates following our minimal model Eqs. (8) and (9), and starting from initial null skews, we demonstrated that the compositional skews are at all times decomposable into the sum of a transcription-associated component and a replication-associated component:

$$S[p, \alpha, (\pm)](t) = pS_R(t) \pm S_T[\alpha](t). \tag{159}$$

If the substitutional pattern given by Eqs. (8) and (9) lasts for ever, the skews converge towards their **equilibrium** values, and asymptotically reach them after long evolutionary time (Graur and Li 1999). The equilibrium skews can be directly computed from the substitution rates. More generally, as the substitutional pattern may change over time, the equilibrium skews computed from today's substitution rates give the current direction of evolution. We demonstrated that the equilibrium skews follow the same decomposition into transcription- and replication-associated components:

$$S^*[p, \alpha, (\pm)] = pS^*_R \pm S^*_T[\alpha]. \tag{160}$$

As reported in Section I.4, the proof of Eqs. (159) and (160) relies importantly on the smallness of the coefficients τ^a_R, τ^a_T and τ^s_T as observed in the human genome (Eqs. (51) to (53)). We also took into account the neighbor-dependent $r = CpG \to TpG$ substitution rate using the model introduced in (Arndt et al. 2003).

Finally what about the **observed** compositional skews, computed from the current DNA sequence? If the substitutional pattern has robustly satisfied Eqs. (8) and (9) for a sufficiently long period of time, we can suppose that the theoretical time evolution Eq. (159) has significantly contributed to shape the observed compositional skews. Therefore we expect the current observed compositional skews to follow more or less the same decomposition into transcription- and replication-associated components:

$$\boxed{S[p, \alpha, (\pm)] = pS_R \pm S_T[\alpha]}. \tag{161}$$

☛ *In our minimal evolution model, the compositional asymmetry can be linearly decomposed into a transcription-associated component and a replication-associated component. The replication-associated compositional asymmetry is proportional to the replication fork polarity, whereas the transcription-associated asymmetry increases in magnitude with transcription rate and changes sign with gene orientation.*

Numerical test. In Fig. 12, we illustrate the additive decomposition of the equilibrium skews into replication- and transcription-associated components (Eq. (160))

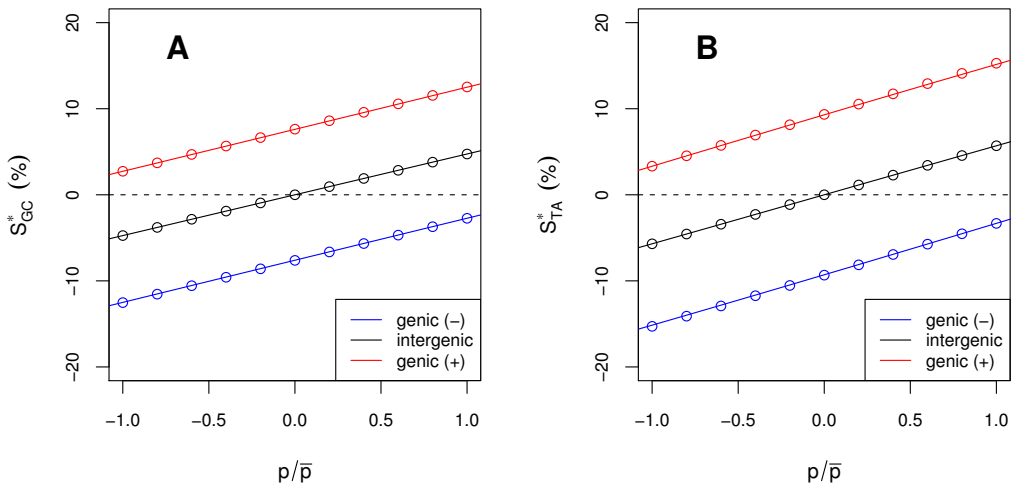

Figure 12: **Decomposition of the equilibrium skews into transcription- and replication-associated components.** In intergenic regions substitution rates depend on the replication fork polarity $\tau[p] = \tau_0^s + \tau_R^s + p\tau_R^a$. In genic regions substitution rates depend on the replication fork polarity and gene orientation $\tau_{\text{genic }(\pm)}[p] = \tau_{\text{intergenic}}[p] + \tau_T^s \pm \tau_T^a$. Exact solution is represented as circles, pertubative solution as solid line. Equilibrium GC skew (A) and TA skew (B) versus replication fork polarity p. S_{TA} and S_{GC} skews are defined in Eq. (158).

for the toy model substitution rates:

$$\tau_{\text{intergenic}}[p] = \tau_0^s + \tau_R^s + p\tau_R^a = \tau_0^s + \tau_R^s + (p/\bar{p})(\bar{p}\tau_R^a), \tag{162}$$

$$\tau_{\text{genic }(\pm)}[p] = \tau_{\text{intergenic}}[p] + \tau_T^s \pm \tau_T^a, \tag{163}$$

where the coefficients $\tau_0^s + \tau_R^s$, $\bar{p}\tau_R^a$, τ_T^s and τ_T^a were estimated in the human genome as described in Section I.2.3 (Eqs. (43)-(47) and Eq. (50)). According to our minimal model Eqs. (5)-(7), the substitution rates correspond to those obtained in intergenic and genic ($\pm$) regions of replication fork polarity p. In Fig. 12, as predicted by Eq. (160), the replication-associated skew in intergenic regions is found proportional to p. In genic regions we recover the same linear dependence upon p, adding up for sense genes or subtracting down for antisense genes a constant corresponding to the (mean) transcription-associated skew, in agreement with Eq. (160).

Note for the readers of Sections I.3 and I.4. As shown in Fig. 12, the pertubative solutions for the equilibrium values are indistinguishable from the exact solutions. The coefficients S_R^* and S_T^* in the biologist convention (Eq. (158)) are easily deduced from the coefficients given in Section I.4.

Summary of Chapter I

To determine the DNA sequence, we have to choose arbitrary one of the two strands, for instance the published strand. The sequence is read in the $5' \to 3'$ direction on this **reference strand**. The sequence on the **complementary strand** is then given by reverse complementarity. The orientation ($\pm$) of genes (Fig. 5), as well as the orientation of replication forks (Fig. 7), were defined relatively to the reference strand. Over cell cycles, a locus x is replicated by a proportion $p_{(\pm)}(x)$ of ($\pm$) forks. The difference of these proportions defines the **replication fork polarity**:

$$p(x) = p_{(+)}(x) - p_{(-)}(x). \tag{164}$$

Gene orientation and transcription rate are the natural parameters to describe the strand asymmetry due to transcription. Replication fork polarity is the natural parameter to describe the strand asymmetry due to replication.

For a substitution rate τ (*e.g.* $A \to G$), we denote by τ^c the reverse complementary substitution rate (*e.g.* $T \to C$). It is much more convenient to study strand asymmetry using the symmetrical part $\tau^s = [\tau + \tau^c]/2$ and asymmetrical part $\tau^a = [\tau - \tau^c]/2$ of substitution rates. The symmetrical part corresponds to the average of a substitution rate on the two DNA strands, whereas the asymmetrical part measures the **substitutional asymmetry** between the two DNA strands. The compositional asymmetry is measured by the **compositional skews**:

$$S_{TA} = \frac{[T] - [A]}{[T] + [A]}, \quad S_{GC} = \frac{[G] - [C]}{[G] + [C]}. \tag{165}$$

In our minimal evolution model, the impact of replication fork polarity p, transcription rate α and gene orientation ($\pm$) on substitution rates and compositional skews is described by the following equations:

$$\tau^s[p, \alpha, (\pm)] = \tau_0^s + \tau_R^s + \tau_T^s[\alpha], \tag{166}$$

$$\tau^a[p, \alpha, (\pm)] = p\tau_R^a \pm \tau_T^a[\alpha], \tag{167}$$

and

$$S[p, \alpha, (\pm)] = pS_R \pm S_T[\alpha], \tag{168}$$

where $\alpha = 0$ ($\tau_T[0] = 0$, $S_T[0] = 0$) corresponds to the intergenic case. The compositional asymmetry, as well as the substitutional asymmetry, can be decomposed into two distinct contributions, one associated to transcription and the other one to replication. The replication-associated asymmetry is proportional to the replication fork polarity p, while the transcription-associated asymmetry increases in magnitude with transcription rate α and changes sign with gene orientation ($\pm$).

Chapter II

Theory of the spatio-temporal DNA replication program

In this Chapter, we present a rigorous analysis of the spatio-temporal program of DNA replication. Importantly, we indicate how to measure replication fork polarity from replication timing data, which will be crucial for the interpretation of compositional strand asymmetry. We first recall the complexity of the spatio-temporal program of DNA replication in higher eukaryotes, which motivates the introduction of a rigorous theoretical framework.

II.1 Introduction

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. According to the so-called "replicon" paradigm derived from prokaryotes (Jacob et al. 1963), this process starts with the binding of some "initiator" protein complex to a specific "replicator" DNA sequence called origin of replication. The recruitment of additional factors initiates the bi-directional progression of two divergent replication forks along the chromosome. One strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand) (see Figs. 6 and 7 of Chapter I). In eukaryotic cells, this event is initiated at a number of replication origins and propagates until two converging forks collide at a terminus of replication (Bell and Dutta 2002; DePamphilis 2006). The initiation of different replication origins is coupled to the progression through S phase but there is a definite flexibility in the usage of the replication origins at different developmental stages (Hyrien and Méchali 1993; Gerbi and Bielinsky 2002; Schübeler et al. 2002; Anglana et al. 2003; Fisher and Méchali 2003). Also, it can be strongly influenced by the distance and timing of activation of neighboring replication origins, by the transcriptional activity and by the local chromatin structure (Gerbi and Bielinsky 2002; Schübeler et al. 2002; Anglana et al. 2003; Fisher and Méchali 2003). Actually, sequence requirements for a replication origin vary significantly

between different eukaryotic organisms. In the unicellular eukaryote *S. cerevisiae*, the replication origins spread over 100-150 bp and present some highly conserved motifs (Bell and Dutta 2002). However, among eukaryotes, *S. cerevisiae* seems to be the exception that remains faithful to the replicon model. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp (Bell and Dutta 2002). In multicellular organisms, the nature of initiation sites of DNA replication is even more complex (DePamphilis 2006). Metazoan replication origins are rather poorly defined and initiation may occur at multiple sites distributed over a thousand of base pairs (Gilbert 2001). The initiation of replication at random and closely spaced sites was repeatedly observed in *Drosophila* and *Xenopus* early embryo cells, presumably to allow for extremely rapid S phase, suggesting that any DNA sequence can function as a replicator (Hyrien and Méchali 1993; Coverley and Laskey 1994; Sasaki et al. 1999). A developmental change occurs around midblastula transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites (Hyrien and Méchali 1993; Sasaki et al. 1999). Thus, although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types (Bogan et al. 2000; Gilbert 2004; DePamphilis 2006). As recently proposed by many authors (Demeret et al. 2001; Méchali 2001; McNairn and Gilbert 2003), the need to fulfill specific requirements that result from cell diversification may have led high eukaryotes to develop various epigenetic controls over the replication origin selection rather than to conserve specific replication sequence. This might explain that for many years, very few replication origins have been identified in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human. In several recent studies, replication bubbles and small nascent DNA strands synthetized at origins were purified by various methods (Mesner et al. 2006; Lucas et al. 2007; The ENCODE Project Consortium 2007; Cadoret et al. 2008; Karnani et al. 2009; Mesner et al. 2011) and hybridized to microarrays, to map a few hundred putative origins over a small fraction ($\lesssim 1\%$) of the human genome. However, the concordance between the different studies is very low (from $< 5\%$ to $< 25\%$), even when they employ the same technique (Cadoret et al. 2008; Karnani et al. 2009), for reasons that are currently unclear (Hamlin et al. 2010). The reliable detection of individual origins seems today still an extremely difficult experimental task. This contrasts with the flourishing availability of genome-wide replication timing data, for several eukaryotic organisms ranging from yeast (Raghuraman et al. 2001), to drosophila (Schübeler et al. 2002), to mouse (Hiratani et al. 2008), and to human (Woodfine et al. 2005). Very recently genome-wide replication timing data has been determined in several human cell types (Woodfine et al. 2005; Desprat et al. 2009; Chen et al. 2010; Hansen et al. 2010; Ryba et al. 2010; Yaffe et al. 2010), which permits to study changes in the replication program across differentiation.

In this experimental context, what kind of information can we extract from replication timing data about the mechanisms underlying the spatio-temporal replication program? Note that the replication timing at a given locus depends on the local initiation properties, but it depends equally on the initiation properties of neighboring sites as replication forks propagate (de Moura et al. 2010; Yang et al. 2010). As a very challenging inverse problem, is it possible to infer the underlying initiation properties from the replication timing data? How much the replication fork propagation does affect the link between replication timing, transcriptional activity and chromatin context? Motivated by our study of compositional strand asymmetry in Chapter I, is there a way to experimentally measure the replication fork polarity? In order to avoid any overstatement on the replication program in higher eukaryotes, it seems necessary from the early beginning to provide a rigorous theoretical framework (de Moura et al. 2010; Hyrien and Goldar 2010; Yang et al. 2010), where required assumptions are explicitly stated.

II.2 Constant replication fork velocity

Here we assume that replication origins are bidirectional and the replication fork velocity v is constant. From these simple assumptions we can already extract fundamental clues such as the replication fork polarity and the difference between initiation site density and termination site density (Baker et al. 2011; Rappailles et al. 2011).

☛ *Importantly for Chapters III, IV and V, we demonstrate that the replication fork polarity is related to the derivative of the mean replication timing.*

Replication program in one cell cycle

If the replication fork velocity v is constant, the spatio-temporal program of replication for one cell cycle is completely specified by the position x_i and the firing time t_i of the n activated bidirectional replication origins O_i (Fig. 1A). From each bidirectional origin, two divergent forks propagate at velocity v, until they meet a fork of opposite orientation (Fig. 1A). Let T_i be the termination locus where the fork coming from O_i meets the fork coming from O_{i+1}. Straightforward calculations lead to the space-time coordinates (y_i, u_i) for T_i:

$$y_i = \frac{1}{2}(x_{i+1} + x_i) + \frac{v}{2}(t_{i+1} - t_i), \quad u_i = \frac{1}{2v}(x_{i+1} - x_i) + \frac{1}{2}(t_{i+1} + t_i). \qquad (1)$$

In Fig. 1, the x-axis is conventionally oriented in the $5' \to 3'$ direction of the reference strand. Hence sense $(+)$ and antisense $(-)$ forks defined in Chapter 1, correspond respectively to rightward and leftward moving forks in Fig. 1B. Around origin O_i (for $x \in [y_{i-1}, y_i]$), the replication timing $t_R(x)$ and the fork orientation $o(x) = \pm 1$ are given by:

$$t_R(x) = t_i + |x - x_i|/v \quad \text{and} \quad o(x) = \text{sign}(x - x_i). \qquad (2)$$

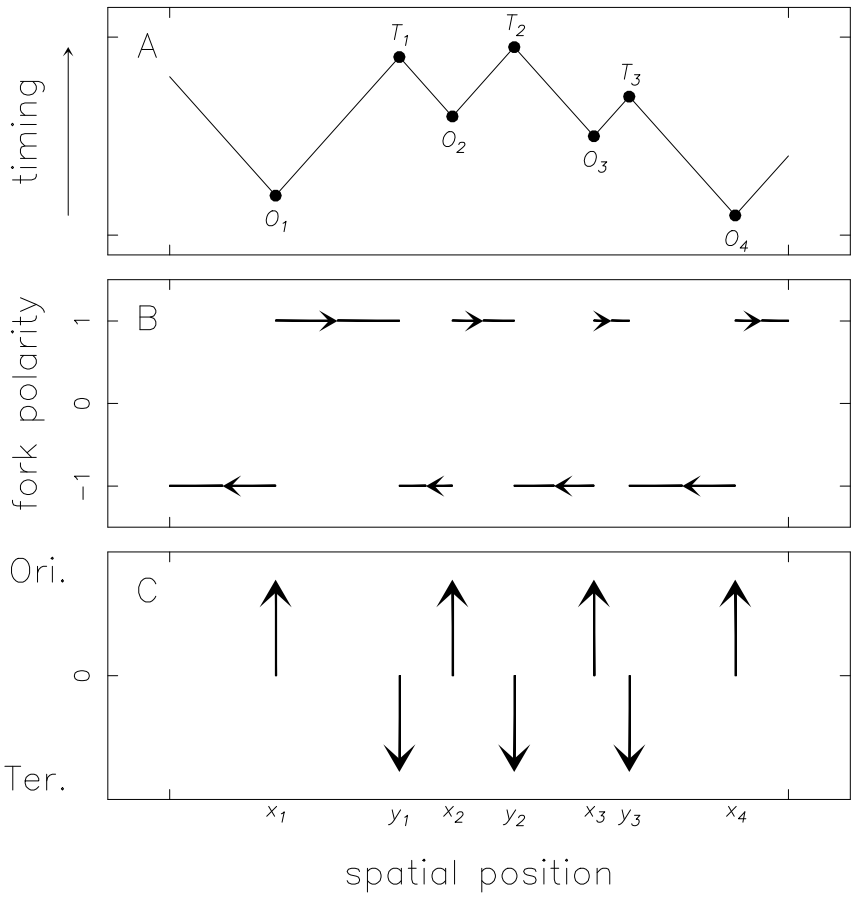

Figure 1: **Replication program in one cell cycle.** (A) Replication timing $t_R(x)$, (B) replication fork orientation $o(x)$ and (C) spatial location of replication origins (upward arrows) and termination sites (downward arrows). $O_i = (x_i, t_i)$ corresponds to the origin i positioned at location x_i and firing at time t_i. Fork coming form O_i meets the fork coming from O_{i+1} at termination site T_i with space-time coordinates (y_i, u_i) given in Eq. (1). Note that we can deduce the fork orientation in (B) (resp. origin and termination site locations in (C)) by simply taking successive derivatives of the timing profile in (A) (Eqs. (3) and (4)). If we take the population average, we can deduce in turn the replication fork polarity, and the difference between initiation site density and termination site density by simply taking successive derivatives of the mean replication timing (Eqs. (7) and (8)). The fundamental hypothesis is that the replication fork velocity v is constant.

Finally, using the Dirac distribution δ to represent origin locations $\delta(x - x_i)$ and termination sites $\delta(y - y_i)$ (Fig. 1C), we obtain the following fundamental relationships:

$$v \frac{d}{dx} t_R(x) = o(x), \tag{3}$$

$$v \frac{d^2}{dx^2} t_R(x) = \sum_i \delta(x - x_i) - \sum_i \delta(x - y_i). \tag{4}$$

In other words, we can extract, up to a multiplicative constant, the fork orientation $o(x)$ (Fig. 1B) and the location of origin and termination sites (Fig. 1C) by simply taking successive derivative (Eqs. (3) and (4)) of the timing profile $t_R(x)$ (Fig. 1A).

Extension to population average

Single cell determination of the replication timing profile is beyond current experimental abilities. Replication timing data are usually obtained using thousands to millions of cells, and thus only determine the population average replication program. If we assume a nearly deterministic replication program, where at each cell cycle nearly the same set of replication origins is used, and where all replication origins fire at specific times, then the population average replication program reflects faithfully what occurs in each cell cycle. However there is now increasing evidence that the replication program is stochastic (Friedman et al. 1997; Patel et al. 2006; Rhind 2006; Czajkowsky et al. 2008), it is now believed that no two cell cycles use the same set of replication origins and the same firing times. But if we take into account the stochasticity of the replication program, some care is needed in interpreting mean replication timing profiles (de Moura et al. 2010; Rhind et al. 2010; Retkute et al. 2011). For instance, it has been proposed that the mean replication timing gradient measures the replication fork velocity (Raghuraman et al. 2001). However, this is true only for a nearly deterministic replication program. In whole generality, the mean replication timing gradient also reflects the proportion over cell cycles of leftward and rightward moving forks replicating a locus (de Moura et al. 2010). It has also been proposed that replication origins correspond to minima of the mean replication timing profile (Raghuraman et al. 2001), but this intuitive claim turns out to be incorrect (Retkute et al. 2011). A careful and rigorous analysis is therefore necessary to interpret mean replication timing profiles. The successive derivatives of the mean replication timing profile give direct access to the replication fork polarity and to the distribution of initiation and termination events, as proposed and analysed in (de Moura et al. 2010; Retkute et al. 2011). We propose here an elementary and very general derivation of these relationships.

As Eqs. (3) and (4) are true for each cell cycle, they are in particular true if we average over cell cycles. We note first that the replication fork polarity $p(x)$ is equal to the population average of the fork orientation $o(x)$:

$$p(x) = p_{(+)}(x) - p_{(-)}(x) = \langle o(x) \rangle, \tag{5}$$

where $p_{(\pm)}(x)$ denote the proportions, over cell cycles, of $(\pm)$ forks replicating the locus x. The initiation site density $d_{\mathrm{Ori}}(x)$ and termination site density $d_{\mathrm{Ter}}(x)$ are equal to the population averages:

$$d_{\mathrm{Ori}}(x) = \left\langle \sum_i \delta(x - x_i) \right\rangle, \quad d_{\mathrm{Ter}}(x) = \left\langle \sum_i \delta(x - y_i) \right\rangle. \tag{6}$$

As taking the spatial derivative commutes with population average, Eq. (3) can be rewritten as:

$$\boxed{v \, \frac{d}{dx} \langle t_R(x) \rangle = p(x)}, \tag{7}$$

as well as Eq. (4) into:

$$\boxed{v \, \frac{d^2}{dx^2} \langle t_R(x) \rangle = d_{\mathrm{Ori}}(x) - d_{\mathrm{Ter}}(x)}, \tag{8}$$

where $\langle t_R(x) \rangle$ is the mean replication timing. Hence the successive derivatives of the mean replication timing give access, up to a multiplicative constant, to the replication fork polarity and the difference between initiation site density and termination site density.

Remark: application to experimental data. Replication timing data have always a finite spatial resolution (from 1kbp in yeast to tens of kbp in human). Moreover, as the experimental mean replication timing profiles are noisy, we cannot simply take the naive derivative of the mean replication timing, which would yield numerically unstable results. We note that as taking the spatial derivative also commutes with the spatial average, Eqs. (7) and (8) are still valid when averaging over the spatial coordinate. There are several ways to perform spatial averaging, but they all consist in convolving the original signal by a weight function (positive function of integral 1). The spatial averaging can be performed at various scales. Over larger and larger scales, the noise is more and more reduced but at the expense of a loss in the spatial resolution. The average of the signal f at scale a is defined as:

$$f_a(x) = \int dy \phi_a(x - y) f(y), \tag{9}$$

where $\phi_a(x) = \frac{1}{a}\phi(x/a)$ is the weight function ϕ dilated at scale a. For instance if we take for the weight function the rectangular function:

$$\sqcap(x) = 1 \quad \text{if} \quad |x| < \frac{1}{2}, \quad \text{and} \quad 0 \quad \text{elsewhere}, \tag{10}$$

we recover the usual definition of the spatial average at scale a:

$$f_a(x) = \frac{1}{a} \int_{x-a/2}^{x+a/2} dy f(y). \tag{11}$$

In practical situations, the rectangular function often turns out to be a bad choice, as it is not a sufficiently smooth function. Indeed the rectangular function present jumps (hence singularities) at $x = \pm\frac{1}{2}$. One usually takes smoother weight functions, such as the gaussian, or any sufficiently smooth weight function.

Initiation site density and origin efficiency

If initiations can only occur at predefined sites x_k, the initiation site density profile formally reduces to:

$$d_{\mathrm{Ori}}(x) = \sum_k E_k \delta(x - x_k), \tag{12}$$

where E_k, the **observed efficiency** of locus x_k, is defined as the fraction of cells where an initiation is observed at x_k. In some eukaryotes, such as yeast, initiations predominantly occur at specific sites (well-positioned replication origins), that usually spread over several hundreds bp. In this situation there are sharp peaks in the initiation site density profile, and the height of these peaks are given by the observed efficiencies. This contrasts with the termination site density profile, which is generally expected to be a rather smooth profile, as the variable firing times lead to greatly dispersed termination sites over cell cycles (Retkute et al. 2011). Therefore well-positioned replication origins reveal themselves as singularities in the $d_{\mathrm{Ori}}(x)$ profile, and consequently according to Eq. (8), as singularities in the second derivative of the mean replication timing profile. Local minima of the mean replication timing are singularities in the second derivative but the converse is not true. A minima is present only if the derivative is negative upstream and positive downstream, in other words if the replication fork polarity switch from negative to positive values (Eq. (7)). It clearly depends on the efficiency of the origin, and equally on the main directionality of forks that passively replicate the origin. Therefore well-positioned origins do not necessarily correspond to minima of the mean replicating timing (Retkute et al. 2011). Finally we note that in some cases, for instance in the mouse β-globin locus, initiation sites are dispersed over a large genomic region (extended initiation zones) that can reach several hundreds kbp (Aladjem 2007). In extended initiation zones the efficiency of each locus is very weak, and we can no longer speak of well-positioned replication origins. The resulting initiation site density profile is also smoother in such genomic regions.

II.3 Independent firing of replication origins

We still assume that origins are bidirectional and that the replication fork velocity v is constant. In this section, we further assume the independent firing of replication origins. In a pioneering work, Bechhoefer and his group (Jun et al. 2005; Jun and Bechhoefer 2005) observed that the DNA replication program is formally equivalent to a 1D nucleation-and-growth process. In the late 1930s, Kolmogorov (1937), Johnson and Mehl (1939), and Avrami (1939; 1940; 1941) independently derived a model (the so-called KJMA model) for nucleation-and-growth processes that describes the

phase transition kinetics. Bechhoefer's group generalized and adapted the KJMA model to the study of DNA replication kinetics. They demonstrated that once the intrinsic firing properties of replication origins are given, most features of the DNA replication program can be analytically predicted, including the observed density of initiations and the replication timing distribution (Yang et al. 2010). Recently many modelling efforts (de Moura et al. 2010; Luo et al. 2010; Yang et al. 2010) have been devoted to infer, from the replication timing distribution, the intrinsic firing properties of replication origins. Here, we first briefly present the KJMA model. Then we report our main result, namely the analytical inversion of the KJMA model. Finally we discuss some shortcomings of the theoretical inversion approach when applied to currently available experimental data.

☛ *We demonstrate that the local initiation properties can be analytically predicted from replication timing data. However application to currently available experimental data still requires further investigation.*

II.3.1 The KJMA model applied to DNA replication kinetics

Analogy between DNA replication and 1D nucleation-and-growth process

Nucleation-and-growth processes model the irreversible transition from an old (untransformed) phase into a new (transformed) phase. In such processes the phase transition decomposes into three steps: nucleation (appearance of the new phase), growth of transformed domains, and coalescence (merging of two transformed domains). Examples of such processes are widespread in physics (crystallization, freezing phase change, ...) and in material sciences (random deposition on a surface, film growth, ...), see (Evans 1993; Fanfoni and Tomellini 1998; Christian 2002) for reviews. In our biological context, it is clear from Fig. 2 that DNA replication is formally a 1D nucleation-and-growth process (Jun and Bechhoefer 2005):

- the initiation, in other words the activation or firing of a replication origin, corresponds to a nucleation event;
- the replication eye, or replication bubble, corresponds to an island of transformed domain (replicated DNA);
- the elongation (growth of the replication bubble due to the propagating replication forks) corresponds to the expansion of a transformed domain;
- the merging of two expanding domains (coalescence) corresponds, in the DNA replication context, to a termination event.

The phase transition kinetics depends on the laws governing the nucleation and the growth of the transformed domains. In its simplest form, the KJMA model (Kolmogorov 1937; Johnson and Mehl 1939; Avrami 1939, 1940, 1941) assumes a homogeneous and constant nucleation rate I and a linear growth law (constant growth velocity v). The fraction of transformed phase $f(t)$ at time t can then be

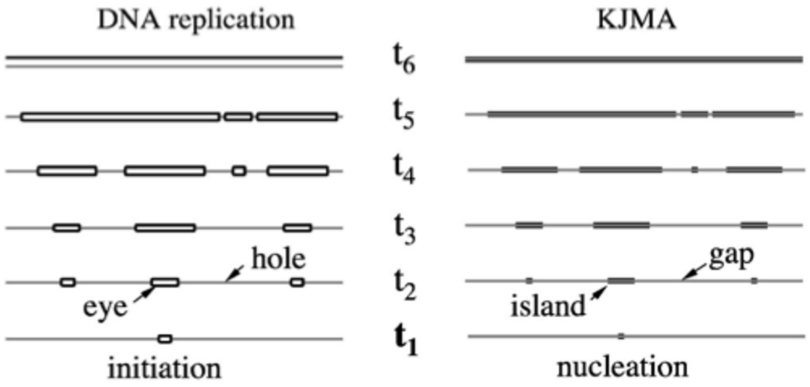

Figure 2: **Analogy between DNA replication and the 1D KJMA model.** Reproduced with permission from (Jun and Bechhoefer 2005). Copyright (2005) by the American Physical Society.

analytically derived from the nucleation rate and the growth velocity (Christian 2002). Sekimoto (1984a; 1984b; 1986) pushed further the KJMA formalism for 1D growth and showed that the distribution of eye and hole sizes can equally be derived. Later on, Bechhoefer and collaborators (Jun et al. 2005; Jun and Bechhoefer 2005) generalized these results to a spatially homogeneous but time-dependent nucleation rate $I(t)$, with the specific goal to model DNA replication kinetics, and in particular the genome wide distribution of replication eye and hole sizes determined by DNA combing experiments (Herrick et al. 2000, 2002). As noted by Jun et al. (2005), the KJMA model can be further extended to account for a space-time dependent nucleation rate $I(x,t)$. This opens new perspectives in the modeling of DNA replication kinetics as for instance when the replication program is spatially structured, as observed in almost all eukaryotes (Yang et al. 2010). For applications of the KJMA model to biological problems, *e.g.* the random completion paradox in *Xenopus* early embryo (Hyrien et al. 2003), we refer the reader to the various publications of the Bechhoefer's group (Herrick et al. 2002; Zhang and Bechhoefer 2006; Bechhoefer and Marshall 2007; Yang and Bechhoefer 2008; Gauthier and Bechhoefer 2009; Yang et al. 2010).

Theoretical remark: limitation of the KJMA formalism. Unfortunately, the KJMA formalism cannot be applied to any growth law or when the nucleation events are correlated. For some growth laws, *e.g.* diffusion-limited growth[1], the KJMA derivation is not correct, as stated by the "no-protrusion condition" in (Sekimoto 1986) or equivalently the "phantom overgrowth" in (Tomellini and Fanfoni 1997). In the present work, as we assume that the replication fork velocity is constant, we will not

[1] If τ denotes the time lapsed since the nucleation, the radius of the transformed domain increases very slowly as $R(\tau) \sim \sqrt{\tau}$.

encounter this issue and the KJMA derivation is perfectly rigorous. Importantly, the KJMA formalism also assumes that nucleations occur independently of each other. Indeed, if we formally define a local nucleation rate (depending or not on space and time), we are explicitly assuming that the nucleation will occur independently of the nucleations events in the surroundings. As Kolmogorov (1937) already suggested in his title "On the static theory of crystallization in metal", the independent nucleation hypothesis corresponds to a static view of nucleation-and-growth processes. In more dynamical models, one could allow nucleation to favor, or on the contrary to inhibit, subsequent nucleation events in the nearby untransformed phase. When nucleations are spatially or temporally correlated, for instance when nucleation is forbidden close to the frontier of a transformed domain (hard-core repulsion), this explicitly breaks the independence assumption of the KJMA model. Tomellini and Fanfoni (2003) extensively studied those shortcomings of the KJMA formalism, and generalized the KJMA model to include spatial correlations.

Definitions of interest: the observed density of initiations $n(x,t)$ and the unreplicated fraction $s(x,t)$

The **unreplicated fraction** $s(x,t)$ is defined as the fraction of cells where the locus x is not yet replicated at time t. The unreplicated fraction characterizes the DNA replication kinetics. From $s(x,t)$, we can define the replicated fraction $f(x,t) = 1 - s(x,t)$ as the fraction of cells where the locus x is already replicated at time t. From $s(x,t)$ we can also extract the probability distribution of the replication timing. Note that a locus x is unreplicated at time t iff[2] its replication timing $t_R(x)$ is greater than t. Therefore the probability distribution $P(x,t)$ of the replication timing at locus x is given by:

$$P(x,t) = -\partial_t\, s(x,t), \tag{13}$$

as $s(x,t) = \mathrm{Prob}(t_R(x) \geq t)$. To characterize the distribution of initiation events, we introduce the **observed density of initiations** $n(x,t)$ at locus x and time t. In other words $n(x,t)$ is the probability[3] to observe an initiation at locus x and time t. Several characteristics of the DNA replication program can be directly deduced from $s(x,t)$ and $n(x,t)$, for instance:

* the probability distribution of the replication timing, and consequently the median and the mean replication timing, according to Eq. (13);
* the length of replicated DNA in a genomic region $\mathcal{R}$ at time t: $l_{\mathcal{R}}(t) = \int_{\mathcal{R}} dx\, f(x,t)$;
* the length of unreplicated DNA in $\mathcal{R}$ at time t: $\bar{l}_{\mathcal{R}}(t) = \int_{\mathcal{R}} dx\, s(x,t)$;
* the number of initiations in $\mathcal{R}$ per unit of time: $n_{\mathcal{R}}(t) = \int_{\mathcal{R}} dx\, n(x,t)$;
* the rate of DNA synthesis in $\mathcal{R}$: $dl_{\mathcal{R}}(t)/dt = \int_{\mathcal{R}} dx\, P(x,t)$;

[2] abbreviation for "if and only if"

[3] Strictly speaking, $n(x,t)dxdt$ is the probability of observing an initiation at $[x, x+dx]\times[t, t+dt]$. I will often make a similar slight abuse of terminology.

- the initiation rate in $\mathcal{R}$, defined as the number of initiations per unit time per unit length of unreplicated DNA: $I_{\mathcal{R}}(t) = n_{\mathcal{R}}(t)/\bar{l}_{\mathcal{R}}(t)$;
- the initiation site density, introduced in Eq. (6), $d_{\mathrm{Ori}}(x) = \int dt\, n(x,t)$, and consequently the observed efficiency of well-positioned origins.

Several experimental techniques were developed to investigate the replication kinetics and the distribution of initiation events:

- Time-course microarray experiments (Raghuraman et al. 2001) measure the replicated fractions at different time points through S-phase;
- Repli-Seq experiments (Chen et al. 2010; Hansen et al. 2010) measure the probability distribution of the replication timing for several S-phase fractions;
- Nascent strand studies (Cadoret et al. 2008), trapping of replication bubbles (Mesner et al. 2011), and other techniques (Gilbert 2010) map well-positioned replication origins and estimate their efficiencies. The mapping of replication origins is still a difficult experimental task, in contrasts with replication timing experiments that are currently performed with relative ease (Gilbert 2010);
- DNA combing (Bensimon et al. 1994) has found several applications in the study of DNA replication (Herrick et al. 2000, 2002; Anglana et al. 2003; Conti et al. 2007; Courbet et al. 2008; Czajkowsky et al. 2008; Marheineke et al. 2009; Rappailles et al. 2011). As DNA combing is a single-molecule technique, it permits to visualize single-cell snapshots that reveal much more information on the spatio-temporal replication program than the population average quantities we were considering above.

☛ *From a theoretical view point, our main goal is to predict $s(x,t)$ and $n(x,t)$ from the parameters of the model. Many relevant characteristics of the spatio-temporal program of DNA replication can then be deduced.*

Analytical prediction of $n(x,t)$ and $s(x,t)$

Under the assumption of independent firing of replication origins, the firing probability of an unreplicated locus does not depend on the nearby initiations events. This probability can however depend both on the locus and on the time of S-phase. If the locus is already replicated it can no longer fire since in eukaryotes re-replication is not allowed (DePamphilis 2006). We can then define the **local initiation rate** $I(x,t)$ as the probability the locus x, if not replicated at time t, to fire at t. In models that assume independent firing, a local initiation rate can always be defined, and it completely specifies the intrinsic firing properties (as detailed in the next paragraph). Almost all stochastic models of the DNA replication program proposed so far (Lygeros et al. 2008; Blow and Ge 2009; de Moura et al. 2010; Luo et al. 2010; Yang et al. 2010; Retkute et al. 2011) assume independent firing of replication origins, and are thus special cases of the KJMA model. The space-time dependent initiation rate $I(x,t)$ can thus be considered as the basic ingredient of the model. In

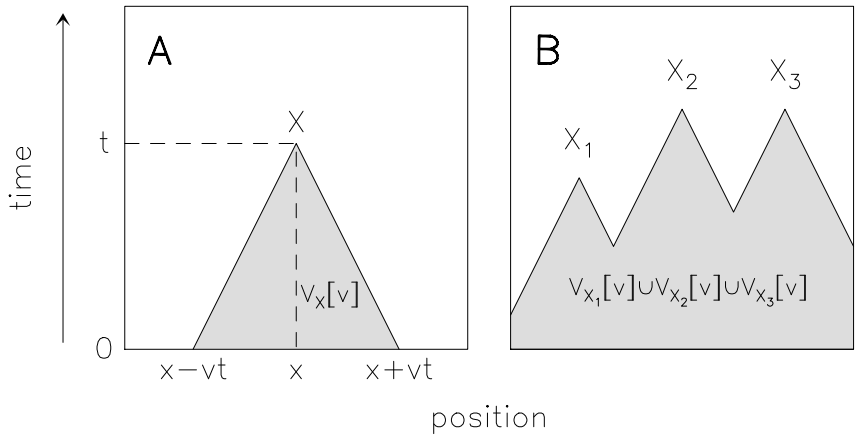

Figure 3: **The Kolmogorov argument.** (A) A locus x is unreplicated at time t iff no initiation occurred in the past light cone $V_X[v]$ of $X = (x, t)$ (grey region). (B) All the loci x_1, x_2, x_3 are unreplicated at respective times t_1, t_2, t_3 iff no initiation occurred in $V_{X_1}[v] \cup V_{X_2}[v] \cup V_{X_3}[v]$ (grey region).

the previously mentioned stochastic models of the replication program, most groups use numerical simulations to estimate $s(x,t)$ and $n(x,t)$ from the intrinsic firing properties (free parameters of the models), a task which is rather time consuming. Bechhoefer's group proved that $s(x,t)$ and $n(x,t)$ can be analytically derived from $I(x,t)$, generalizing the KJMA formalism to a space-time dependent nucleation (initiation) rate $I(x,t)$ (Yang et al. 2010).

Proof. First, as illustrated in Fig. 3A, we remark that the locus x is unreplicated at time t iff no initiation occurred in the past light cone $V_X[v]$ of $X = (x,t)$. Thus $s(x,t)$ is equal to the probability $P_0(V_X[v])$ that no initiation occurred in $V_X[v]$ (Kolmogorov (1937) argument). As the origins fire independently, this probability is equal to:

$$P_0(V_X[v]) = \lim_{\Delta Y \to 0} \prod_{Y \in V_X[v]} [1 - I(Y)\Delta Y] = e^{-\int_{V_X[v]} dY I(Y)}. \tag{14}$$

Therefore the unreplicated fraction is equal to (Yang et al. 2010):

$$\boxed{s(x,t) = e^{-\int_{V_X[v]} dY I(Y)}}. \tag{15}$$

Next we note that we can observe an initiation at X iff no initiation occurred in V_X and one initiation occurred at X (Fig. 3A). As the origins fire independently, the probability to observe an initiation at X is equal to $n(X) = I(X)P_0(V_X[v])$. Thus the observed density of initiations is given by (Yang et al. 2010):

$$\boxed{n(x,t) = I(x,t)e^{-\int_{V_X[v]} dY I(Y)}}. \tag{16}$$

Let us emphasize that from Eqs. (15) and (16), both the unreplicated fraction and the observed density of initiations at X depend on the local initiation properties in the whole past light cone of X. Therefore the replication timing distribution $P(x,t)$ at locus x, the initiation site density $d_{\mathrm{Ori}}(x)$ at locus x, the observed efficiency if x corresponds to a well-positioned origin, all depend not only on the local initiation properties at x, but also on the local initiation properties of neighboring loci, as replication forks propagate.

Theoretical digression: correlations in observed initiation events and replication kinetics

A lot of information can be inferred from the correlations observed between replication eye sizes or between replication hole sizes (Fig. 2). For instance, if nearby replication eyes have comparable sizes, it indicates that the origins have fired synchronously (Blow et al. 2001). The distribution of eye-to-eye distances also provides information on the typical origin spacings during a replication round. The analysis of molecular combing experiments in early-embryo *Xenopus* for instance reveal that the distribution of eye-to-eye distances and the correlations between replication eye sizes were not compliant with a completely homogeneous firing rate (Jun et al. 2004). It has been proposed that DNA polymerases linked to the chromatin fiber could segregate, leading to the so-called "replication factories" where the chromatin forms loops (Cook 1999). Interestingly the physical properties of the chromatin fiber, which controls the typical size of the chromatin loops and in turn the origin spacings, can explain both the observed regularity of origin spacings and the existence of an origin exclusion zone (Jun et al. 2004).

Here, we investigate, in a more formal approach, the correlations between observed initiation events and the correlations between the replication timing of different loci. To this purpose, we introduce the N-point unreplicated fraction $s_N(X_1,\cdots,X_N)$, where X_i denotes the space-time point (x_i,t_i). It is defined as the fraction of cells where each loci x_i is unreplicated at time t_i. The joint probability distribution of the replication timing is given by:

$$P_N(X_1,\ldots,X_N) = (-1)^N \partial_{t_1}\cdots\partial_{t_N} s_N(X_1,\cdots,X_N), \tag{17}$$

where $s_N(X_1,\cdots,X_N) = \mathrm{Prob}(t_R(x_1)\geq t_1 \text{ and } \ldots \text{ and } t_R(x_N)\geq t_N)$. As regards the distribution of initiation events, we introduce the N-point joint observed density $n_N(X_1,\ldots,X_N)$. We can define $n_N(X_1,\ldots,X_N)$ as the probability to observe, during the same cell cycle, an initiation at each X_i. As combing gives snapshots of the replication state for individual DNA fragments, it could in principle be used to determine correlations between observed initiation events. As shown just below, $s_N(X_1,\cdots,X_N)$ and $n_N(X_1,\cdots,X_N)$ can be easily derived from the local initiation rate $I(x,t)$.

Proof. We assume first that no X_i belongs to the past light cone of another X_j, as

depicted in Fig. 3B. We remark on this figure that each loci x_i is unreplicated at time t_i iff no initiation occurred in $V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]$. Therefore $s_N(X_1, \ldots, X_N)$ is equal to $P_0(V_{X_1}[v] \cup \cdots \cup V_{X_N}[v])$. As the origins fire independently this probability is equal to:

$$P_0(V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]) = e^{-\int_{(V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]} dY I(Y)}. \tag{18}$$

Therefore the N-point unreplicated fraction is equal to:

$$\boxed{s_N(X_1, \cdots, X_N) = e^{-\int_{V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]} dY I(Y)}}. \tag{19}$$

In his study of the 1D KJMA model, Sekimoto (1986) introduced the space-time correlation functions $C_N(X_1, \ldots, X_N)$ which are equal to the $s_N(X_1, \ldots, X_N)$ functions. Consistently, the formula (3.8) of (Sekimoto 1986) reduces to Eq. (19) when the velocity v is constant. Now, as illustrated in Fig. 3B, we remark that we can observe an initiation at each X_i iff no initiation occurred in $V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]$ and an initiation occurs at each X_i. As the origins fire independently, the probability to observe an initiation at each X_i is equal to $n_N(X_1, \ldots, X_N) = I(X_1) \cdots I(X_N) P_0(V_{X_1}[v] \cup \cdots \cup V_{X_N}[v])$. Therefore the observed joint densities are equal to:

$$\boxed{n_N(X_1, \cdots, X_N) = I(X_1) \cdots I(X_N) e^{-\int_{V_{X_1}[v] \cup \cdots \cup V_{X_N}[v]} dY I(Y)}}. \tag{20}$$

To illustrate why the replication fork propagation necessarily creates correlations between the replication timings at different loci and also between the observed initiation densities, let us specify these expressions for $N = 2$. The 2-point unreplicated fraction is equal to:

$$s_2(X_1, X_2) = s(X_1) s(X_2) e^{\int_{V_{X_1}[v] \cap V_{X_2}[v]} dY I(Y)}. \tag{21}$$

Hence the distribution of replication timings at loci x_1 and x_2 are correlated due to the possible initiation events in their common past-light cone $V_{X_1}[v] \cap V_{X_2}[v]$. Indeed if $I(Y)$ is not identically zero in $V_{X_1}[v] \cap V_{X_2}[v]$ then $s_2(X_1, X_2) \neq s(X_1) s(X_2)$. Similarly we have for the observed joint densities:

$$n_2(X_1, X_2) = n(X_1) n(X_2) e^{\int_{V_{X_1}[v] \cap V_{X_2}[v]} dY I(Y)}. \tag{22}$$

The independent firing of origins, that generates the observed joint densities, is a stochastic process where the firings (probability $I(X)$ of firing at X) are by hypothesis not correlated. But the observed joint densities at X_1 and X_2 are possibly correlated due to the initiation events that may occur in their common past-light cone $V_{X_1}[v] \cap V_{X_2}[v]$. Finally, if one of the X_i belongs to the past light cone of another X_j, $n_N(X_1, \cdots, X_N)$ is necessary null. As re-replication is not allowed, we cannot observe an initiation in the future-light cone of another. This is a trivial correlation verified by the observed joint densities.

Intrinsic firing properties

If initiation can only occur at predefined sites called **potential origins**, the local initiation rate has the following form (Yang et al. 2010):

$$I(x,t) = \sum_i \delta(x - x_i) I_i(t), \tag{23}$$

where x_i is the position of the potential origin i, and $I_i(t)$ is its firing rate. The firing rate $I_i(t)$ gives the probability, if the locus x_i is unreplicated at time t, that the potential origin i fires at time t. If the locus x_i is replicated by a fork coming from a neighboring origin, the potential origin i will not be activated during this cell cycle, the potential origin is said to be **passively replicated**. The replication kinetics observed at a locus does not necessarily reflects the local initiation properties, due to the confounding effect of passive replication (de Moura et al. 2010; Hyrien and Goldar 2010; Yang et al. 2010). For instance, if a potential origin is observed early replicating, it does not necessary imply that the origin is intrinsically early firing: the origin could be close to a very efficient and early firing origin. In the same way, the observed efficiency of a potential origin depends as much on the context (is it close or not to other potential origins? what are the firing properties of the neighboring origins?) than on its individual firing properties. Following (Yang et al. 2010), let us now define, from the local initiation rate $I(x,t)$, several quantities that characterize the local initiation properties.

Intrinsic firing properties of a potential origin. In order to define the intrinsic firing properties of the potential origin i, let us isolate, by an experience of thought, the potential origin i from its neighbors. In other words, in our gedanken experiment, the potential origin i is **prevented from passive replication**. The potential origin i, if it has not fired before time t, fires with probability $I_i(t)$ at time t. We can then define the intrinsic unreplicated fraction $s_i(t)$ as the probability that the potential origin i has not yet fired at time t:

$$s_i(t) = \lim_{\delta_t \to 0} \prod_{u \leq t} [1 - I_i(u)\delta_t] = e^{-\int_0^t du I_i(u)}. \tag{24}$$

Similarly the intrinsic replicated fraction $f_i(t) = 1 - s_i(t)$ is defined as the probability that i has fired before t. Note that $s_i(t)$ (resp. $f_i(t)$) is equal to the probability that the firing time of i is greater (resp. less) than t. Hence the **intrinsic firing time distribution**, or origin activation time in (de Moura et al. 2010), is given by:

$$\phi_i(t) = -\partial_t s_i(t) = I_i(t) e^{-\int_0^t du I_i(u)}. \tag{25}$$

We note that $\phi_i(t)$ is also equal to the probability to observe a firing event at time t, hence $\phi_i(t)$ could also be called the intrinsic density of initiations. Consistently, the probability that i fires at t given that it has not fired before t is equal to the firing rate:

$$I_i(t) = \frac{\phi_i(t)}{s_i(t)}. \tag{26}$$

This last equality also permits to derive the firing rate for any intrinsic firing time distribution:

$$I_i(t) = \frac{\phi_i(t)}{\int_0^t du \phi_i(u)}. \tag{27}$$

Finally the **intrinsic efficiency** $\mathcal{E}_i$, defined as the probability that i has fired before the end of S phase ($t = t_{\text{end}}$), is equal to:

$$\mathcal{E}_i = \int_0^{t_{\text{end}}} dt \phi_i(t) = f_i(t_{\text{end}}). \tag{28}$$

Note that due to passive replication, the observed origin efficiency E_i (Eq. (12)) is generally not equal to and smaller than the intrinsic efficiency $\mathcal{E}_i$. Similarly the observed replication timing distribution $P(x_i, t)$ at origin i (Eq. (13)) is generally not equal to the intrinsic firing time distribution $\phi_i(t)$.

Intrinsic firing properties of a genomic region. Experimental replication timing data have always a finite spatial resolution (several kbp); usually we cannot resolve individually the potential origins. Moreover, if we want to probe the local initiation properties at different scales, we also need to consider larger genomic regions that may harbor several potential origins. Fortunately, we can easily extend the previous definitions to a genomic region $\mathcal{R}$, prevented from passive replication. Let us first consider that $\mathcal{R}$ gathers a cluster of potential origins $x_i \in \mathcal{R}$. If we neglect the spatial extension, the potential origin that fires first gives its firing time to the region $\mathcal{R}$. Then the **intrinsic unreplicated fraction** is equal to the probability that none potential origin has fired during $[0, t]$:

$$s_{\mathcal{R}}(t) = \prod_{x_i \in \mathcal{R}} s_i(t) = e^{-\sum_{x_i \in \mathcal{R}} \int_0^t du I_i(u)}. \tag{29}$$

Hence the intrinsic unreplicated fraction of $\mathcal{R}$ is given by:

$$s_{\mathcal{R}}(t) = e^{-\int_0^t du I_{\mathcal{R}}(u)}, \tag{30}$$

where we introduce the **intrinsic firing rate** of $\mathcal{R}$:

$$I_{\mathcal{R}}(t) = \sum_{x_i \in \mathcal{R}} I_i(t) = \int_{\mathcal{R}} dx I(x, t). \tag{31}$$

Note that the second equality allows us to extend the discussion to a continuous $I(x, t)$. The **intrinsic firing time distribution** of $\mathcal{R}$ is given by:

$$\phi_{\mathcal{R}}(t) = -\partial_t s_{\mathcal{R}}(t) = I_{\mathcal{R}}(t) e^{-\int_0^t du I_{\mathcal{R}}(u)}. \tag{32}$$

The **intrinsic efficiency** of $\mathcal{R}$ writes:

$$\mathcal{E}_{\mathcal{R}} = \int_0^{t_{\text{end}}} dt \phi_{\mathcal{R}}(t). \tag{33}$$

If we neglect its spatial extension, a region $\mathcal{R}$ thus behaves as an effective potential origin with a firing rate $I_{\mathcal{R}}(t)$. When clustering several potential origins of a region $\mathcal{R}$, the effective firing rate $I_{\mathcal{R}}(t)$ is simply the sum of the firing rates $I_i(t)$ over $\mathcal{R}$.

☛ *Under the assumption of independent firing, models of the DNA replication program are exactly solvable. We can always define a local initiation rate $I(x,t)$ which specifies the intrinsic firing properties of replication origins, such as the intrinsic firing time distribution and the intrinsic efficiency. The unreplicated fraction $s(x,t)$ and the observed density of initiations $n(x,t)$, as well as the higher order N-point functions, can be explicitly expressed from any local initiation rate $I(x,t)$.*

II.3.2 Inversion of the KJMA model

Motivation: determining local initiation properties from the replication timing distribution

In usual applications of the KJMA model, the nucleation rate and the growth law are dictated by the physics of the problem. The kinetics predicted by the KJMA model can then be directly confronted to experiments. In our biological context however, we have no idea of what the local initiation rate should be. The mechanisms that determine the initiation properties in eukaryotes are currently unclear and are presumably very complex. Moreover, the location of replication origins is poorly known in eukaryotes, and even less is known about the mechanisms that regulate their firing times (Gilbert 2001; Aladjem 2007; Méchali 2010). Therefore, as a very challenging project, we would like to extract the local initiation properties from replication timing data, that are now widely available. This could help us in determining the position of replication origins as well as their intrinsic firing time distribution (does this observed origin fire intrinsically late or early? Is this observed origin intrinsically efficient?). By confronting the local initiation properties to other types of genomic or epigenetic data, we may hope getting insight into the underlying mechanisms that govern the spatio-temporal replication program in higher eukaryotes. Recently, various groups (de Moura et al. 2010; Luo et al. 2010; Yang et al. 2010) have attempted to infer the local initiation properties from replication timing data in yeast (where the position of potential replication origins are well characterized (Nieduszynski et al. 2007)). They all used a fitting strategy that consists in determining the set of parameters (positions of replication origins and their firing time properties) that best reproduces the replication timing data. Here we propose to solve the inverse problem and to determine $I(x,t)$ analytically. Theoretically, the analytical inversion of the KJMA model has many advantages: (i) it avoids overfitting since we do not have to impose some predefined constraints on the local initiation rate (always necessary if we consider a finite number of parameters for the fitting procedure), (ii) it can as easily predict well-positioned origins as extended initiation zones, and (iii) it is almost immediate in computation time. Unfortunately, from a practical point of view, the analytical approach has several shortcomings when applied to currently available experimental data. We do not

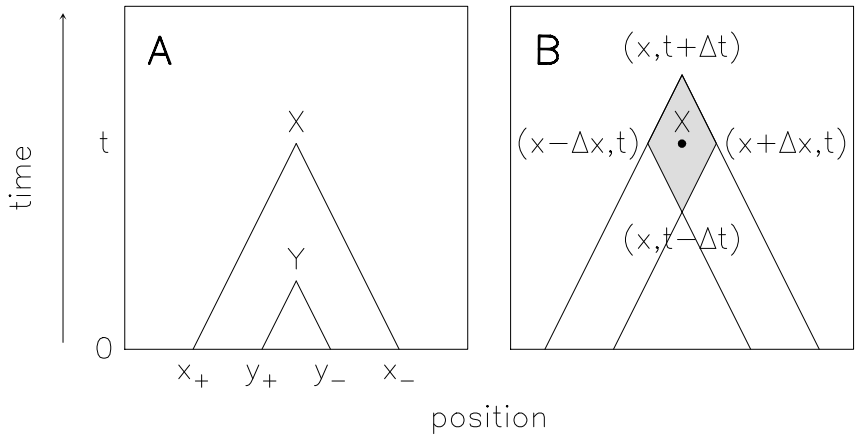

Figure 4: **Inversion of the KJMA model.** (A) Light cone coordinates $x_\pm = x \mp vt$ of space-time point $X = (x, t)$. Note that Y belongs to the past light cone of X iff $x_+ \leq y_+$ and $y_- \leq x_-$. (B) Diamond D_{X,Δ_X} of area $2\Delta_X\Delta_t$ (grey region) surrounding space-time point $X = (x, t)$. The corners of D_{X,Δ_X} are the space-time points $(x - \Delta_X, t)$, $(x, t + \Delta_t)$, $(x + \Delta_x, t)$ and $(x, t - \Delta_t)$.

know yet how to properly handle the noise, as well as other experimental issues (such as data normalization), in order to provide a robust and efficient determination of the local initiation properties directly from the analysis of replication timing data.

Analytical prediction of $I(x, t)$

An elegant proof of the inverse problem can be established when introducing the light cone coordinates (Wald 1984), as illustrated in Fig. 4A:

$$x_+ = x - vt, \quad x_- = x + vt. \tag{34}$$

In these coordinates, the past light cone has a simple expression:

$$V_X[v] = \{Y \quad \text{such that} \quad x_+ \leq y_+, \quad y_- \leq x_-\}. \tag{35}$$

In the light cone coordinates, it follows from Eq. (15):

$$\int_{x_+ \leq y_+, y_- \leq x_-} dy_+ dy_- I(y_+, y_-) = -\ln s(x_+, x_-). \tag{36}$$

Differentiating with respect to x_+ and x_-, we get:

$$I(x_+, x_-) = \partial_+ \partial_- \ln s(x_+, x_-). \tag{37}$$

Back to the original (x, t) coordinates, Eq. (37) becomes:

$$\boxed{I(x,t) = -\frac{v}{2}\Box \ln s(x,t)}, \tag{38}$$

where $\Box = \frac{1}{v^2}\partial_t^2 - \partial_x^2$ is the d'Alembertian operator. For numerical or experimental applications, we will consider the discrete and regularized versions of Eq. (38) described just below.

Application to finite resolution data: discrete version. We can give an alternative derivation of Eq. (38), more adapted to numerical and experimental applications, that have always a finite spatial and temporal resolution. Let Δ_t be a scale in time (in practice the time resolution) and $\Delta_x = v\Delta_t$ the corresponding scale in space. Let us consider D_{X,Δ_x}, the diamond of area $2\Delta_x\Delta_t$ surrounding X, as shown in Fig. 4B. The average of $I(x,t)$ in D_{X,Δ_x} is equal to:

$$I_{\Delta_x}(x,t) = \frac{1}{2\Delta_x\Delta_t}\int_{D_{X,\Delta_x}} dY I(Y), \tag{39}$$

$$= \frac{1}{2\Delta_x\Delta_t}\left[\int_{V_{(x,t+\Delta_t)}} dY I(Y) + \int_{V_{(x,t-\Delta_t)}} dY I(Y) - \int_{V_{(x+\Delta_x,t)}} dY I(Y) - \int_{V_{(x-\Delta_x,t)}} dY I(Y)\right], \tag{40}$$

$$= -\frac{1}{2\Delta_x\Delta_t}\left[\ln s(x,t+\Delta_t) + \ln s(x,t-\Delta_t) - \ln s(x+\Delta_x,t) - \ln s(x-\Delta_x,t)\right]. \tag{41}$$

The second equality follows from basic geometry (Fig. 4B), and the third equality from Eq. (15). We recognize in the third equality the discrete d'Alembertian. Indeed the discrete derivatives and d'Alembertian of a signal f are given by:

$$\partial_{x,\Delta_x} f(x,t) = \frac{1}{\Delta_x}[f(x+\frac{\Delta_x}{2},t) - f(x-\frac{\Delta_x}{2},t)], \tag{42}$$

$$\partial_{t,\Delta_t} f(x,t) = \frac{1}{\Delta_t}[f(x,t+\frac{\Delta_t}{2}) - f(x,t-\frac{\Delta_t}{2})], \tag{43}$$

$$\Box_{\Delta_x} f(x,t) = \frac{1}{\Delta_x^2}[f(x,t+\Delta_t) + f(x,t-\Delta_t) - f(x+\Delta_x,t) - f(x-\Delta_x,t)]. \tag{44}$$

Hence the average of $I(x,t)$ in D_{X,Δ_x} is equal to:

$$\boxed{I_{\Delta_x}(x,t) = -\frac{v}{2}\Box_{\Delta_x}\ln s(x,t) = \frac{1}{2\Delta_x\Delta_t}\ln\left[\frac{s(x+\Delta_x,t)s(x-\Delta_x,t)}{s(x,t+\Delta_t)s(x,t-\Delta_t)}\right]}, \tag{45}$$

which is nothing but the discrete version of Eq. (38).

Application to noisy data: regularized version. The d'Alembertian operator, as all differential operators, can amplify noise when applied to experimental data. It is often necessary to regularize $I(x,t)$ by a smooth weight function $W(x,t)$:

$$I_W(x,t) = (W * I)(x,t). \tag{46}$$

We can also perform the regularization at various scales, using the weight function W dilated at scale a_x in space and scale a_t in time: $W_{a_x,a_t}(x,t) = \frac{1}{a_x a_t} W(x/a_x, t/a_t)$. At larger scales, the noise will likely be reduced, but at the expense of a loss in the spatial and temporal resolution. If we convolve Eq. (38) by the weight function W and integrate by parts, we get:

$$\boxed{I_W(x,t) = -\frac{v}{2}\Box_W \ln s(x,t) = \left(\left[-\frac{v}{2}\Box W\right] * \ln s\right)(x,t)}, \tag{47}$$

where we have introduced the regularized d'Alembertian operator $\Box_W$, defined for any signal f as the convolution with the smooth function $\Box W$:

$$\Box_W f(x,t) = ([\Box W] * f)(x,t)\,. \tag{48}$$

Note that Eq. (47) in fact includes the discrete version Eq. (45) as a particular (but singular) case. If we consider for the weight function W the characteristic function of D_{0,Δ_x} (diamond of area $2\Delta_x\Delta_t$ surrounding $x = t = 0$), the regularization of f by W is equivalent to average f over D_{X,Δ_x}:

$$(W * f)(X) = \frac{1}{2\Delta_x\Delta_t}\int_{D_{X,\Delta_x}} dY\, f(Y). \tag{49}$$

Therefore the regularized $I_W(X)$ is equal to $I_{\Delta_x}(X)$ introduced in Eq. (39). Using the theory of distribution, we can prove that:

$$\Box W(x,t) = \frac{1}{\Delta_x^2}[\delta(x,t-\Delta_t) + \delta(x,t+\Delta_t) - \delta(x-\Delta_x,t) - \delta(x+\Delta_x,t)]. \tag{50}$$

We recover in turn the definition of the discrete d'Alembertian (Eq. (44)):

$$\Box_W f(X) = ([\Box W] * f)(x,t), \tag{51}$$

$$= \frac{1}{\Delta_x^2}[f(x,t+\Delta_t) + f(x,t-\Delta_t) - f(x+\Delta_x,t) - f(x-\Delta_x,t)], \tag{52}$$

$$= \Box_{\Delta_x} f(X). \tag{53}$$

Let us point out that the more general regularized version (Eq. (47)) offers many advantages as compared to the discrete version (Eq. (45)). First, for the discrete version we had to impose the scale Δ_x to be equal to $v\Delta_t$. Hence we can apply Eq. (45) only when on the one hand the spatial resolution divides $\Delta_x (= v\Delta_t)$ and on the other hand the temporal resolution divides Δ_t, which may not be possible for any replication fork velocity v. On the contrary, the operator $\Box_W$ is defined at any large enough scales a_x and a_t, whatever the replication fork velocity v, using dilated versions of the weight function W. Moreover it is clear from the Dirac distributions appearing in Eq. (50) that the discrete d'Alembertian corresponds to a singular version of $\Box_W$, and may amplify noise when applied to real data. On the opposite, for any sufficiently smooth weight function W, the regularized d'Alembertian $\Box_W$ corresponds to the convolution with the smooth function $\Box W$, which provides a robust and numerically stable way of estimating $I(x,t)$.

Numerical illustration

Here we illustrate the inversion of the KJMA model on "ideal" replication timing data distribution. We generated the theoretical unreplicated fraction of the multiple-initiator model proposed in (Yang et al. 2010). This model fits quite well the experimental replicated fractions obtained in yeast (McCune et al. 2008). The theoretical data generated is considered as ideal because: (i) the kinetics explicitly satisfies the two key assumptions of the KJMA model (constant replication fork velocity and independent firing of replication origins), (ii) the numerical data are not noisy, (iii) the unreplicated fraction are obtained at a fine spatial (2 kbp) and temporal (1 min) resolution, (iv) the unreplicated fraction cover the whole S-phase. The spatial resolution (2 kbp) corresponds to the resolution of the experimental data fitted by Yang et al (2010). In the following discussion, it will be implicitly considered that each locus x corresponds to a 2 kbp locus. The temporal resolution we choose (1 min) is finer that the temporal resolution (5 min) of the experimental data (McCune et al. 2008). Note that taking a temporal resolution of 5 min would not change considerably our numerical illustration, as all quantities could be consistently determined at the 5 min resolution.

Extracting the intrinsic firing time distribution from replication timing distribution. As exemplified in Fig. 5A for a 260 kbp fragment of yeast chromosome 4, containing 8 potential origins (O_1 to O_8), we generated using the forward KJMA formula (Eq. (15)) the theoretical unreplicated fraction $s(x,t)$ of the multiple-initiator model (Yang et al. 2010). From the theoretical unreplicated fraction $s(x,t)$, we computed the local initiation rate $I(x,t)$ according to the analytical inversion formula Eq. (45). The knowledge of $I(x,t)$ allows in turn to determine the intrinsic firing time distribution $\phi(x,t)$ (Fig. 5B), the observed density of initiations $n(x,t)$, the observed efficiency $E(x)$ and the intrinsic efficiency $\mathcal{E}(x)$ (Fig. 5C). We recover in Fig. 5B the location of the 8 potential origins of the multiple-initiator model. We can even distinguish the intrinsic firing time distribution of each potential origin, for instance O_6 is intrinsically early firing while O_7 fires intrinsically at the mid S-phase ($t \sim$30 min). As shown in Fig. 6 for the origins O_3, O_6 and O_7, the initiation rate $I(x,t)$ determined by the analytical inversion is in perfect agreement with the theoretical initiation rate of the multiple initiator model (Fig. 6A), as well as the intrinsic firing time distribution $\phi(x,t)$ (Fig. 6B). We notice from Fig. 5A that the origin O_7, detected by the numerical inversion, does not correspond to a local minima of the unreplicated fraction. About one origin over three in the multiple-initiator model is not recovered as a local minima in the unreplicated fractions (Yang et al. 2010). It is sometimes assumed that well-positioned origins correspond to local minima in the mean replication timing or the unreplicated fractions (Raghuraman et al. 2001). It is important to point out that the well-positioned origin O_7 would have been missed by such methods.

Impact of passive replication. Passive replication can strongly affects the replication

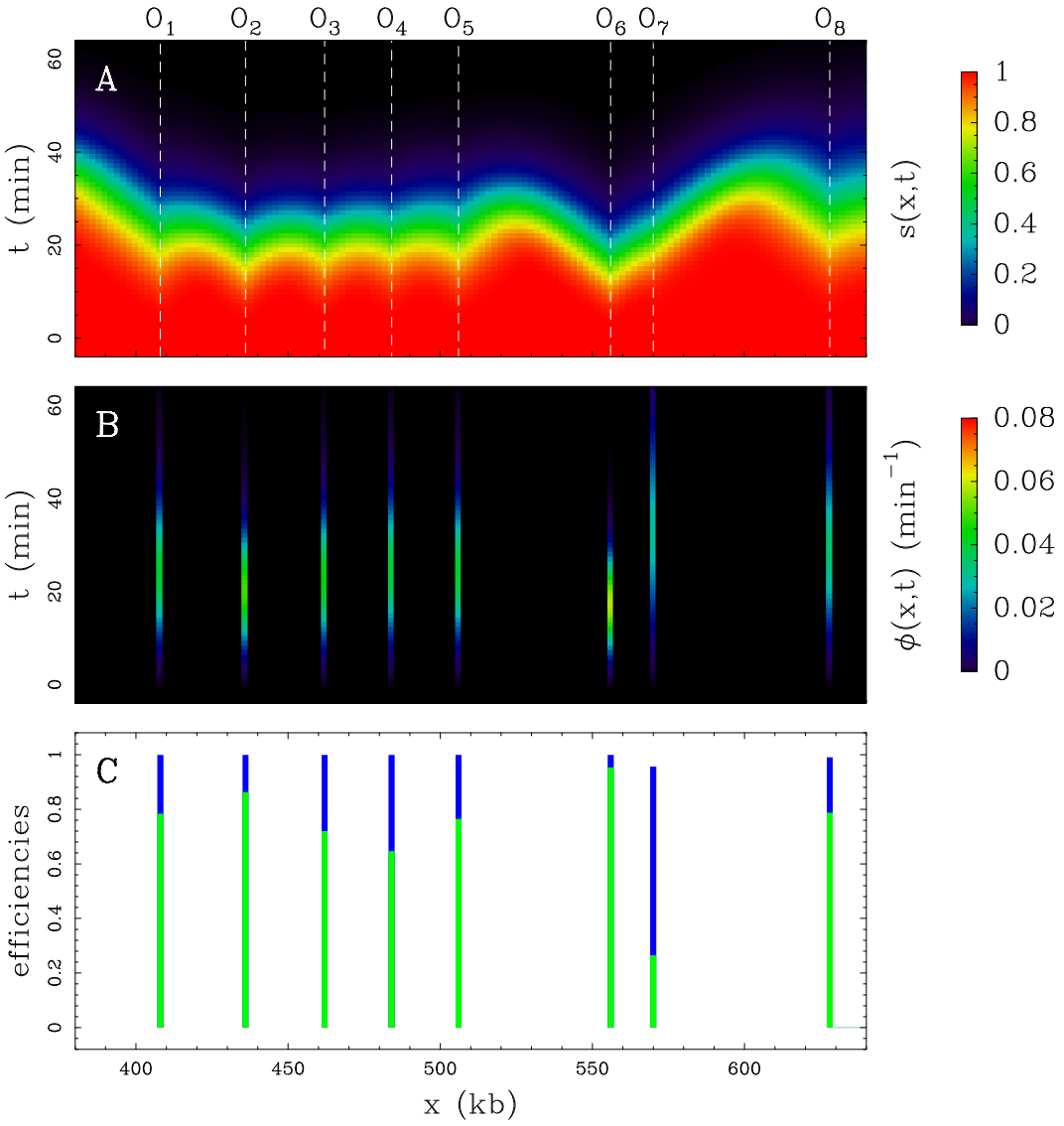

Figure 5: **Extracting the intrinsic firing time distribution from replication timing distribution.** On a 260 kbp fragment of yeast chromosome 4, containing 8 potential replication origins (O_1 to O_8), the unreplicated fraction $s(x,t)$ (A) given by the multiple-initiator model of (Yang et al. 2010) was generated by the forward KJMA formula (Eq. (15)). The local initiation rate $I(x,t)$ was computed from the unreplicated fraction $s(x,t)$ (A) using the inversion of the KJMA model Eq. (45). The intrinsic firing time distribution $\phi(x,t)$ (B), as well as the observed efficiency $E(x)$ (green bars in (C)) and intrinsic efficiency $\mathcal{E}(x)$ (blue bars in (C)), were then determined from the local initiation rate $I(x,t)$ according to their definitions provided in Section II.3.1.

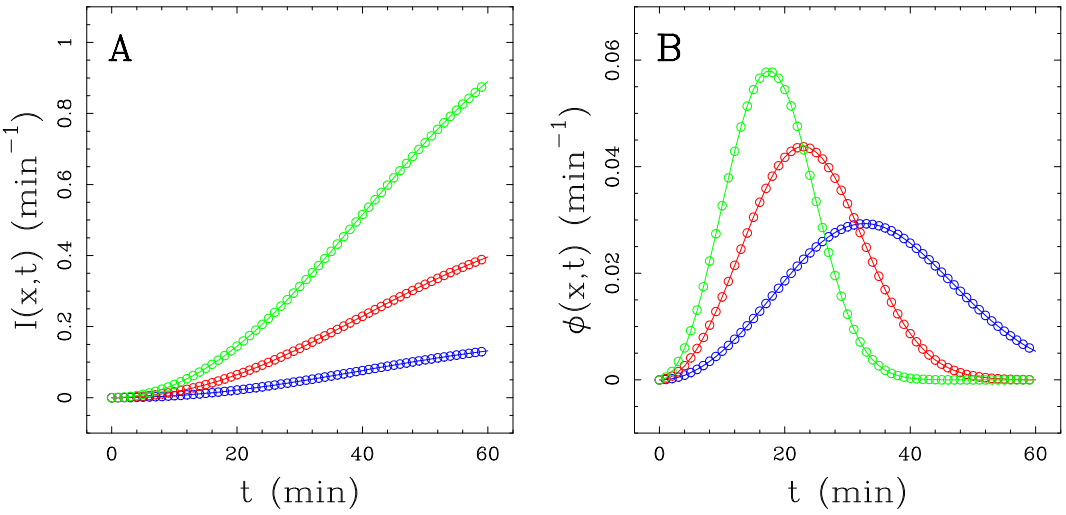

Figure 6: **Numerical test of our theoretical solution of the inverse problem.** Comparison of the solution obtained by the analytical inversion (circles) and the theoretical solution (solid line) for (A) the local initiation rate $I(x,t)$ (Eqs. (38) and (45)) and (B) the intrinsic firing time distribution $\phi(x,t)$ (Eqs. (25) and (32)), for the potential origins O_3 (red) , O_6 (green) and O_7 (blue) (see Fig. 5).

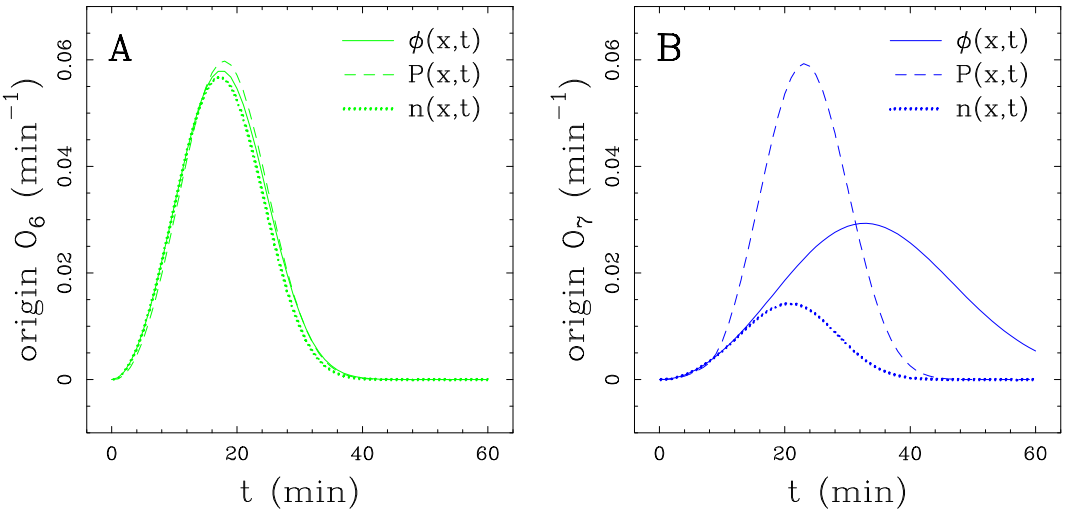

Figure 7: **Impact of passive replication.** Intrinsic firing time distribution $\phi(x,t)$ (solid line), replication timing distribution $P(x,t)$ (dashed line), and observed density of initiations $n(x,t)$ (dotted line) for (A) potential origin O_6 (green) and (B) potential origin O_7 (blue). The potential origin O_7 is often passively replicated by a fork originating from O_6, while O_6 is early firing and unlikely passively replicated (see Fig. 5).

kinetics observed at a locus and the observed efficiencies of replication origins, and can lead to erroneous interpretation of replication timing data (de Moura et al. 2010; Yang et al. 2010; Retkute et al. 2011). Let us illustrate, on the 260 kbp fragment in Fig. 5, some of the consequences of passive replication. We first note that all the potential origins have an intrinsic efficiency (blue bars in Fig. 5C) close to 1, as the smallest intrinsic efficiency (origin O_7) is equal to 96%. On the opposite the observed efficiencies (green bars in Fig. 5C) vary greatly (from 24% for origin O_7 to 94% for origin O_6), depending on the context. Passive replication has therefore a strong impact on the observed efficiencies of most origins. For a potential origin which is rarely passively replicated, we expect the replication timing to be equal to the firing time of the origin. We therefore expect for such an origin the replication timing distribution $P(x,t)$ to be close to the intrinsic firing distribution $\phi(x,t)$. As in such cases the firing time corresponds to an observed initiation event, we also expect the observed density of initiations $n(x,t)$ to be close to the intrinsic firing time distribution $\phi(x,t)$. As shown in Fig. 7A for the origin O_6, which is early firing and unlikely passively replicated (Fig. 5), we have indeed $P(x,t) \sim n(x,t) \sim \phi(x,t)$. However this is no longer true when the potential origin can be passively replicated. For instance the potential origin O_7, which can be passively replicated by a fork originating from O_6 (Fig. 5), has a replication timing distribution clearly different from its intrinsic firing time distribution (Fig. 7B). The replication timing distribution of O_7, as it is likely replicated by a fork coming from O_6, is very close to the firing time distribution of O_6, shifted by the time (7 min) necessary for a fork to propagate from O_6 to O_7 ($x_7 - x_6 = 14$ kbp and $v = 2$ kbp/min). At the onset of S-phase ($t < 16$ min), the origin O_7 is unlikely passively replicated, since only few forks coming from O_6 have the time to reach O_7, and the observed density of initiations at O_7 is very close to its intrinsic firing time distribution (Fig. 7B). At later times the observed density of initiations at O_7 is strongly reduced, as it becomes more and more likely that O_7 will be passively replicated by a fork coming from O_6. Even though the origin O_7 has intrinsically a high probability of firing for $t > 30$ min, we will almost never observe initiations at those times (Fig. 7B). Due to the context (early firing origin O_6 located nearby), the observed density of initiations and the replication timing at O_7 are strongly affected by the passive replication of O_7.

Application to experimental data?

The analytical prediction of the local initiation rate $I(x,t)$ from experimental replicated fractions requires some caution: the straightforward application of Eq. (45) to the experimental replicated fractions obtained in (McCune et al. 2008) leads to an unphysical $I(x,t)$, found negative in most space-time regions. The experimental replicated fractions present however several anomalies. For instance, at each locus x the replicated fraction $f(x,t)$ can only be increasing with time t, but the experimental replicated fractions violate this causality requirement: we observe on Fig. 8A that the $f(x, t = 45\text{min})$ profile (pink curve) is sometimes below the $f(x, t = 35\text{min})$ pro-

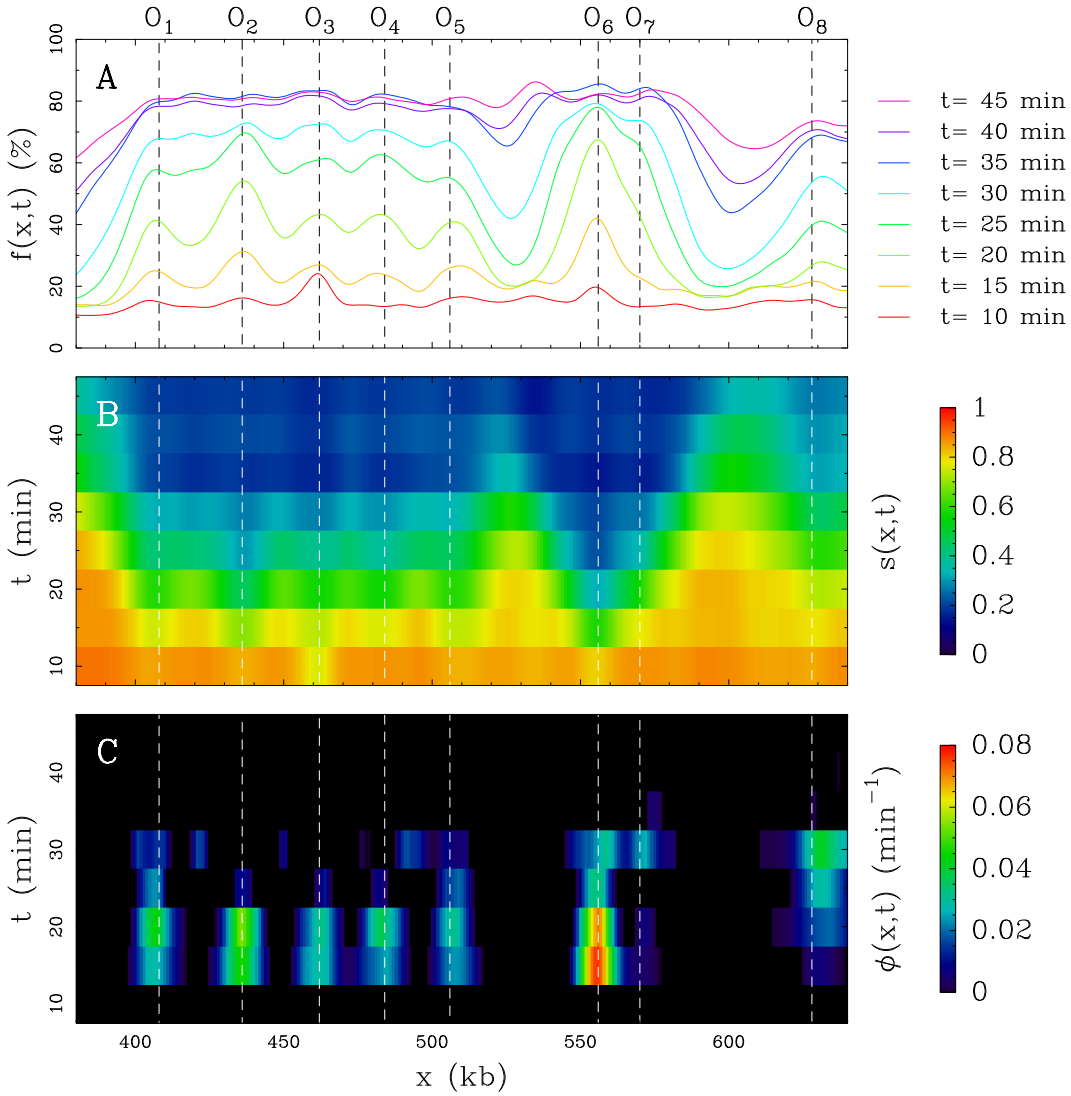

Figure 8: **Extracting the intrinsic firing time distribution from replication timing distribution: application to experimental data.** (A) Experimental replicated fraction $f(x,t)$ profiles, obtained at different S-phase times (from $t = 10$ min to $t = 45$ min every 5 min), along the same 260 kbp fragment of yeast chromosome 4 displayed in Fig. 5, which contains 8 potential replication origins (O_1 to O_8) in the multiple-initiator model of (Yang et al. 2010). The replicated fractions, obtained by time-course microarray experiments, were retrieved from (McCune et al. 2008). (B) Experimental unreplicated fraction $s(x,t)$. (C) Intrinsic firing time distribution $\phi(x,t)$, obtained from the experimental unreplicated fraction $s(x,t)$ as explained in the main text.

file (blue curve). It is therefore not surprising that the experimental replicated fractions are also in conflict with stronger requirements of the KJMA kinetics, namely the constant replication fork velocity and the independent firing of replication origins.

Applying the analytical inversion formula Eq. (45) on unreplicated fractions that do not respect the KJMA kinetics does not make sense, and inevitably yields an aberrant local initiation rate. Hence we must first modify the experimental unreplicated fractions to render them compliant with the KJMA kinetics:

(a) causality requires that $s(x,t)$ decreases with time t. We thus changed iteratively the unreplicated fractions according to: $s(x,t+\Delta_t) \leftarrow \min\left[s(x,t+\Delta_t), s(x,t)\right]$, where $\Delta_t = 5$ min is the time resolution.

(b) if replication forks propagate at velocity v, then for each space-time point X and for every Y in the past light-cone of X we have $s(X) \leq s(Y)$. To satisfy this requirement, it is sufficient to change iteratively the unreplicated fractions according to: $s(x,t+\Delta_t) \leftarrow \min\left[s(x,t+\Delta_t), \min_{y\in[x-\Delta_x,x+\Delta_x]} s(y,t)\right]$, where $\Delta_x = v\Delta_t$. We choose here $v = 2$ kbp/min.

(c) the independent firing of replication origins implies that $I_{\Delta_x}(X)$ is positive for any space-time point X. According to Eq. (45), this requirement is equivalent to $s(x,t+\Delta_t) \leq \frac{s(x+\Delta_x,t)s(x-\Delta_x,t)}{s(x,t-\Delta_t)}$. We therefore changed iteratively the unreplicated fractions according to: $s(x,t+\Delta_t) \leftarrow \min\left[s(x,t+\Delta_t), \frac{s(x+\Delta_x,t)s(x-\Delta_x,t)}{s(x,t-\Delta_t)}\right]$.

On the 260 kbp fragment used to illustrate the inversion of the multiple initiator model (Fig. 5), we modified the experimental unreplicated fractions (Fig. 8B) according to steps (a-c), we then applied Eq. (45) to obtain a local initiation rate $I(x,t)$, from which we finally deduced the intrinsic firing distribution $\phi(x,t)$ (Fig. 8C). The inversion works pretty well qualitatively: we recover the locations of the 8 potential origins O_1 to O_8 of the multiple initiator model (Fig. 5). Furthermore, the firing time distribution (Fig. 8C) is in good qualitative agreement with the firing time distribution of the multiple initiator model (Fig. 5C), for instance O_6 is a very efficient early-firing origin while O_7 is a less efficient origin firing at $t \sim 30$ min.

Arguably, the methodology presented here does not provide a robust and accurate determination of the local initiation rate from the replicated fractions. Causality, constant replication fork velocity, and independent firing of replication origins were imposed quite artificially on the unreplicated fractions. The rather crude and drastic modifications of the unreplicated fractions made in steps (a-c) are numerically unstable, which surely affects the determination of $I(x,t)$. The successful extraction of the local initiation rate from the experimental replicated fractions could considerably ease the analysis of replication kinetics, but it is a very challenging inverse problem. In future work, we would like to investigate several regularization schemes, and determine which can most accurately and most robustly predict the local initiation rate.

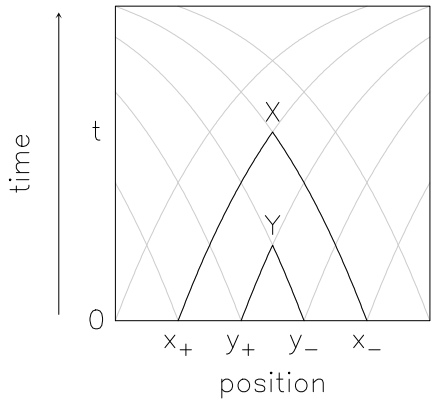

Figure 9: **Space-time dependent fork velocity.** Light gray curves represent the propagation lines of (+) forks (rightward moving forks) and (−) forks (leftward moving forks) in a space-time dependent fork velocity $v(x,t)$. The (±) fork propagation line passing by X intersects the x-axis at the $x_\pm$ coordinate, this defines the light cone coordinates $x_\pm$ of X. Note that Y belongs to the past light cone of X iff $x_+ \leq y_+$ and $y_- \leq x_-$.

Theoretical digression: space-time dependent fork velocity

Kolmogorov (1937) already considered a time-dependent growth velocity $v(t)$. We show in this paragraph that the KJMA model and its inversion can be easily generalized to a space-time dependent replication fork velocity $v(x,t)$. The integral curves of the velocity field $v(x,t)$ correspond in our biological context to the propagation lines of replication forks, depicted as gray lines in Fig. 9. We notice that the transition from a constant velocity v to a space-time dependent velocity $v(x,t)$ amounts formally to replace the Minkowski space-time $ds^2 = dx^2 - v^2dt^2$ by the pseudo-Riemannian space-time $ds^2 = dx^2 - v(x,t)^2dt^2$. In this geometrical viewpoint, the propagation lines of replication forks correspond to the geodesics of the $ds^2 = dx^2 - v(x,t)^2dt^2$ metric. The Kolmogorov argument, and consequently Eqs. (15) and (16) as well as Eqs.(19) and (20), are still true, keeping in mind that $V_X[v]$, the past light cone of X, depends functionally on the velocity field $v(x,t)$. The light cone coordinates, a useful mathematical trick encountered in general relativity (Wald 1984), can be defined in our present situation in the following simple way. As shown in Fig. 9, the space-time point X is at the intersection of a (+) fork and (−) fork propagation lines which respectively intersect the x-axis ($t = 0$) at the light cone coordinates x_+ and x_-. In the light cone coordinates, the past light cone is still given by Eq. (35) and the local initiation rate is still given by Eq. (37). Expressed in the original (x,t) coordinates the local initiation rate is equal to:

$$I(x,t) = \left\{ \frac{1}{2}\left(\frac{1}{v}\partial_t v\right)\frac{1}{v}\partial_t + \frac{1}{2}(\partial_x v)\partial_x - \frac{v}{2}\left(\frac{1}{v^2}\partial_t^2 - \partial_x^2\right)\right\} \ln s(x,t). \qquad (54)$$

The proof is given below. Note that we recover Eq. (38) when the velocity field is constant.

Proof. Let $s \to y_\pm(x_\pm, s)$ be the propagation line of the $(\pm)$ fork initially at $x_\pm$ (s denotes time and $y_\pm(x_\pm, s)$ the position at s). By definition $y_\pm(x_\pm, s)$ verifies:

$$y_\pm(x_\pm, 0) = x_\pm, \tag{55}$$
$$(\partial_s y_\pm)(x_\pm, s) = \pm v(y_\pm(x_\pm, s), s). \tag{56}$$

The light cone coordinates $x_\pm(x,t)$ of X are defined by $y_\pm(x_\pm(x,t),t) = x$. Conversely the origin coordinates $x(x_+, x_-)$ and $t(x_+, x_-)$ verify:

$$y_\pm(x_\pm, t(x_+, x_-)) = x(x_+, x_-). \tag{57}$$

Differentiating this relation with respect to x_+ and x_- yields:

$$(\partial_+ x) = -v(\partial_+ t) \quad \text{and} \quad (\partial_- x) = +v(\partial_- t). \tag{58}$$

The differential operators $\partial_\pm$ are then equal to:

$$\partial_+ = (\partial_+ t)\{\partial_t - v\partial_x\} \quad \text{and} \quad \partial_- = (\partial_- t)\{\partial_t + v\partial_x\}. \tag{59}$$

After some lines of algebra, we get the differential operators:

$$\partial_+\partial_- = (\partial_+\partial_- t)\{\partial_t + v\partial_x\} + (\partial_+ t)(\partial_- t)\{\partial_t - v\partial_x\}\{\partial_t + v\partial_x\}, \tag{60}$$
$$\partial_-\partial_+ = (\partial_-\partial_+ t)\{\partial_t - v\partial_x\} + (\partial_- t)(\partial_+ t)\{\partial_t + v\partial_x\}\{\partial_t - v\partial_x\}. \tag{61}$$

When applying the Schwarz's theorem $\partial_+\partial_- = \partial_-\partial_+$ to x, we get:

$$(\partial_+\partial_- t)v + (\partial_+ t)(\partial_- t)(\partial_t v) = 0. \tag{62}$$

Therefore the differential operator $\partial_+\partial_-$ is equal to:

$$\partial_+\partial_- = (\partial_+ t)(\partial_- t)\left[-\frac{1}{v}(\partial_t v)\{\partial_t + v\partial_x\} + \{\partial_t - v\partial_x\}\{\partial_t + v\partial_x\}\right]. \tag{63}$$

This relation simplifies into:

$$\partial_+\partial_- = -2v(\partial_+ t)(\partial_- t)\left\{\frac{1}{2}\left(\frac{1}{v}\partial_t v\right)\frac{1}{v}\partial_t + \frac{1}{2}(\partial_x v)\partial_x - \frac{v}{2}\left(\frac{1}{v^2}\partial_t^2 - \partial_x^2\right)\right\}. \tag{64}$$

The jacobian of the $(x,t) \to (x_+, x_-)$ change of variables is equal to:

$$J = \begin{vmatrix} (\partial_+ x) & (\partial_- x) \\ (\partial_+ t) & (\partial_- t) \end{vmatrix} = |\overset{>0}{2v}\,(\overset{<0}{\partial_+ t})\,(\overset{>0}{\partial_- t})| = -2v\,(\partial_+ t)\,(\partial_- t). \tag{65}$$

In other words we have $dxdt = Jdx_+dx_-$. As $I(x,t)dxdt = I(x_+,x_-)dx_+dx_-$ and $I(x_+,x_-) = \partial_+\partial_- \ln s(x_+,x_-)$, then:

$$I(x,t) = -\frac{1}{2v(\partial_+ t)(\partial_- t)}\partial_+\partial_- s(x,t), \tag{66}$$

When using the expression Eq. (64) of $\partial_+\partial_-$, we finally get:

$$I(x,t) = \left\{\frac{1}{2}\left(\frac{1}{v}\partial_t v\right)\frac{1}{v}\partial_t + \frac{1}{2}(\partial_x v)\partial_x - \frac{v}{2}\left(\frac{1}{v^2}\partial_t^2 - \partial_x^2\right)\right\}\ln s(x,t). \tag{67}$$

Summary of Chapter II

Under the assumptions that origins are bidirectional and that the replication fork velocity is constant, we have shown that the replication fork polarity is related to the derivative of the mean replication timing (see Fig. 1):

$$p(x) = v\frac{d}{dx} < t_R(x) >, \tag{68}$$

where $p(x)$ is the replication fork polarity, $< t_R(x) >$ the mean replication timing, and v the replication fork velocity. According to the relationship established in Chapter I between the replication-associated strand asymmetry and the replication fork polarity, this equality establishes a link between the replication-associated skew and replication timing data. We have also demonstrated that the difference between initiation site density and termination site density is related to the second derivative of the replication timing (see Fig. 1):

$$d_{\mathrm{Ori}}(x) - d_{\mathrm{Ter}}(x) = v\frac{d^2}{dx^2} < t_R(x) >, \tag{69}$$

where $d_{\mathrm{Ori}}(x)$ (resp. $d_{\mathrm{Ter}}(x)$) is the initiation (resp. termination) site density.

When further assuming that replication origins fire independently, the replication timing distribution, as well as the distribution of initiation events, can be analytically predicted from the intrinsic firing properties of replication origins (Yang et al. 2010). Reciprocally we have solved the inverse problem, by demonstrating that the intrinsic firing properties of replication origins can be analytically extracted from the replication timing distribution. However the application to currently available experimental data requires some caution, in order avoid artifacts induced by data normalization and signal to noise issues.

Chapter III

Transcription- and replication-associated strand asymmetries in the human genome

We show in this Chapter that substitution rates and compositional skews observed in the human genome are consistent with the model proposed in Chapter I. This model states that substitutional as well as compositional asymmetry can be decomposed into transcription- and replication-associated components. The transcription-associated asymmetry, as generally admitted, changes sign with gene orientation and increases in magnitude with gene expression. According to this model, the replication-associated asymmetry is proportional to the replication fork polarity. In Chapter II, under the assumption of constant replication fork velocity, we demonstrated that the replication fork polarity was directly related to the derivative of the mean replication timing. This theoretical result can be used to test the model proposed in Chapter I, using the derivative of the mean replication timing as an estimator of the replication fork polarity.

III.1 Analysis of substitution rates in the human genome

Genome wide substitution rates

The substitutions were tabulated in the human lineage since the divergence with chimpanzee (Chen et al. 2010) (Section I.2.3, §**Methodology**). As shown in Fig. 1, the genome-wide substitution rates are of the order of 10^{-3} substitutions per bp. The four transitions $C \to T$, $G \to A$, $A \to G$ and $T \to C$ are three fold higher than the eight transversions. The genome wide substitution rates respect parity rule type 1 (PR1) (Section I.1.2): reverse complementary substitutions, displayed with the same color coding in Fig. 1A, have nearly equal rates. However if we compute substitution rates separately in genic (+), intergenic and genic (−) regions, PR1 is explicitly broken. As argued in Chapter I (Section I.2.1 or the summary), it is more

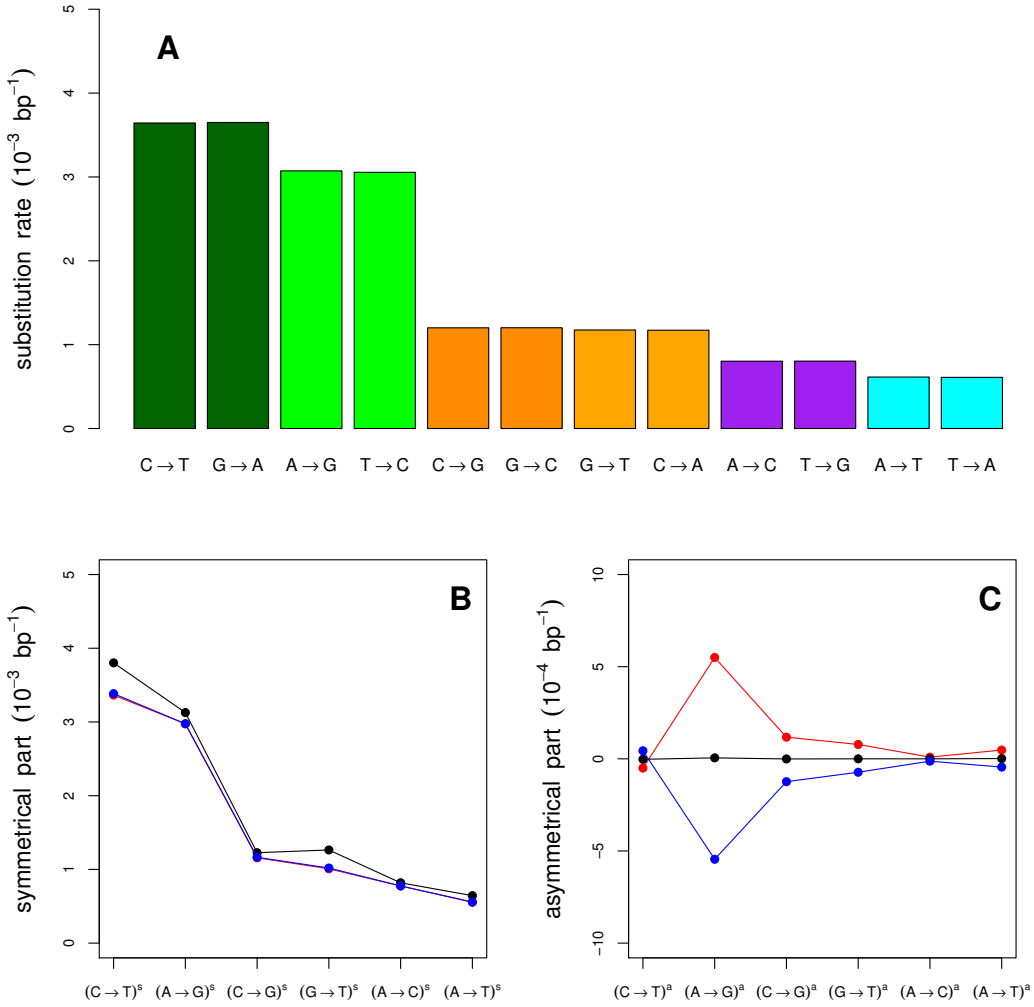

Figure 1: **Genome wide substitution rates.** (A) Genome wide substitution rates. Reverse complementary substitution rates, *e.g.* $C \to T$ and $G \to A$, have the same color coding. (B) Genome wide symmetrical substitution rates in genic sense (red), intergenic (black), and genic antisense (blue) regions. (C) Genome wide asymmetrical substitution rates in genic sense (red), intergenic (black), and genic antisense (blue) regions. Substitution rates were computed on the reference strand (Section I.1.2).

convenient to decompose substitution rates into symmetrical and asymmetrical parts under strand exchange symmetry. As shown in Fig. 1B, symmetrical substitution rates are lower in genic regions than in intergenic regions, this is particularly true for the strong to weak substitutions $(C \to T)^s$ and $(G \to T)^s$. In the perspective of the model proposed in Chapter I (Eq. (166) in the summary), it implies negative $\tau^s_T < 0$ coefficients. Substitutional asymmetries are opposed and significantly different from 0 in genic (+) and (−) regions in Fig. 1C, clearly demonstrating the existence of a transcription-associated substitutional asymmetry. The substitutional asymmetries are one order of magnitude lower than symmetrical substitution rates (and thus than substitution rates), which a posteriori explains the ten fold difference between the y-axis units in Figs. 1B and 1C. The transcription-associated $(A \to G)^a_T > 0$ asymmetry is the highest one, followed by the $(C \to G)^a_T > 0$ and $(G \to T)^a_T > 0$ asymmetries. The sign of the transcription-associated asymmetries are consistent with previous findings (Green et al. 2003; Polak and Arndt 2008; Mugal et al. 2009; Chen et al. 2011), with the exception of the $(C \to T)^a_T$ asymmetry. We found along with (Green et al. 2003; Polak and Arndt 2008; Chen et al. 2011) a negative but weak $(C \to T)^a_T < 0$ asymmetry, in apparent contradiction with (Mugal et al. 2009). However the substitutional pattern determined by Mugal et al. (2009) was estimated further in the past than the human-chimpanzee divergence, hence the discrepancy might be explained by a time dependency of the $(C \to T)^a_T$ asymmetry (Mugal et al. 2009). We point out that the $(C \to T)^a_T$ asymmetry also varies along the transcript, which could hamper the determination of $(C \to T)^a_T$ when computed on the whole transcript. As observed by (Polak and Arndt 2008), a strong localized $(C \to T)^a_T > 0$ asymmetry is found restricted to the first two kbp downstream of the TSS, opposite to the weak $(C \to T)^a_T < 0$ on the remaining transcript.

The substitutional asymmetry decomposes into transcription- and replication-associated components

In the model proposed in Chapter I, both transcription and replication can generate strand asymmetry, and the replication-associated strand asymmetry depends on the replication fork polarity. We choose here to take advantage of the theoretical result of Chapter II (Eq. (68) in the summary), relating the replication fork polarity to the derivative of the mean replication timing. We use for this purpose replication timing data obtained in several human cell lines (Chen et al. 2010; Hansen et al. 2010). As described in (Baker et al. 2011), we computed from these data the Mean Replication Timing (MRT) and its derivative dMRT/dx. Note that the MRT profile is expressed as a fraction of S-phase and has therefore no dimension. If we multiply the MRT by the duration of S-phase we get a reasonable proxy for the MRT expressed in time, although the conversion between S-phase fraction and time is not strictly linear (Blumenthal et al. 1974). According to Eq. (68) of Chapter II, under the approximation of constant replication fork velocity, the replication fork polarity is proportional to the dMRT/dx profile:

$$p(x) = v\, T_S\, \mathrm{dMRT/dx}, \tag{1}$$

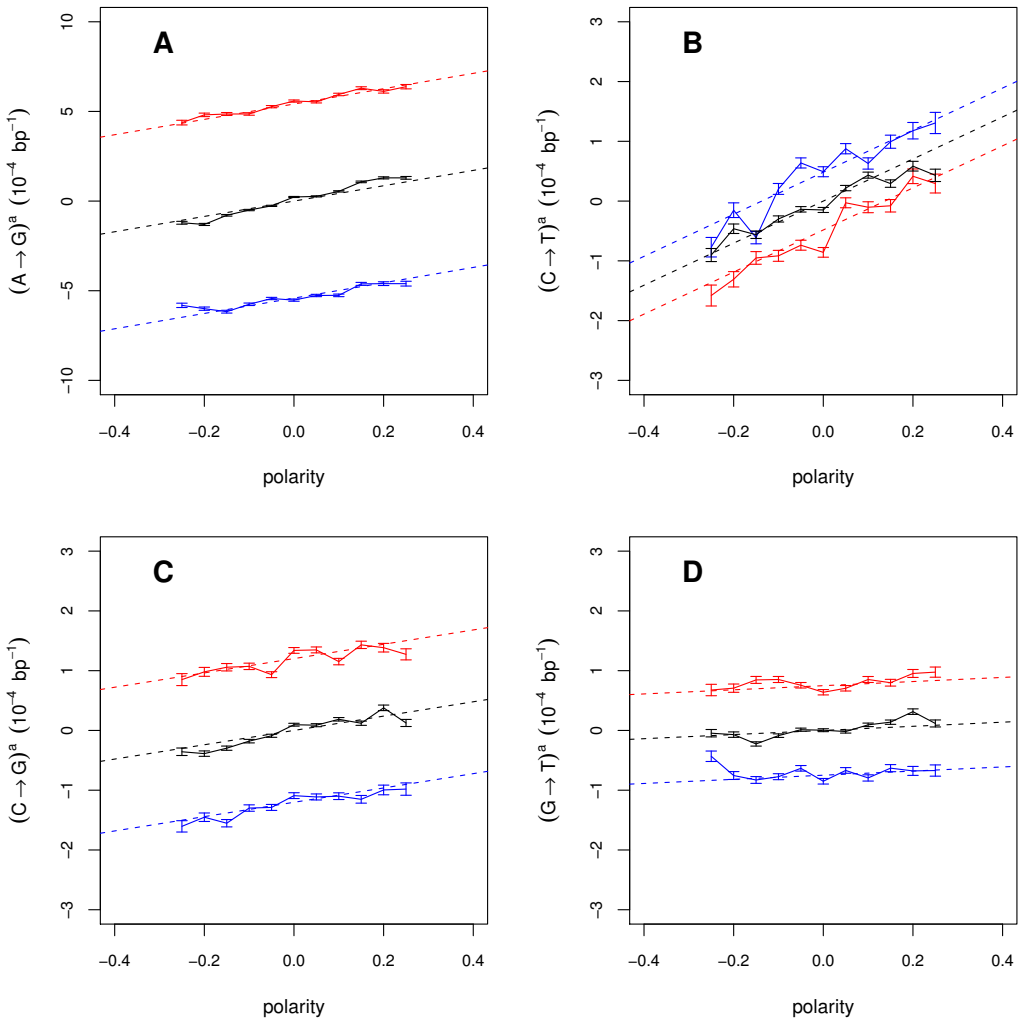

Figure 2: **The substitutional asymmetry decomposes into transcription- and replication-associated components.** Substitutional asymmetry versus replication fork polarity (determined in HeLa cell line using Eq. (1) and replication timing data from (Rappailles et al. 2011)) in genic sense (red), intergenic (black), and genic antisense (blue) regions, for (A) the $A \to G$ substitution, (B) the $C \to T$ substitution, (C) the $C \to G$ substitution, and (D) the $G \to T$ substitution. Substitution rates, replication fork polarity, and the gene orientation were computed on the reference strand. The dashed lines correspond to the least-squares fits to a line, following the linear model Eq. (2). The linear regression coefficients are reported in Table 1.

	$(A \to G)^a$	$(C \to T)^a$	$(C \to G)^a$	$(G \to T)^a$
τ_T^a (10^{-4} bp^{-1})	5.41 ± 0.05	-0.48 ± 0.05	1.20 ± 0.02	0.75 ± 0.02
τ_R^a (10^{-4} bp^{-1})	4.28 ± 0.26	3.52 ± 0.23	1.20 ± 0.12	0.35 ± 0.13

Table 1: **Transcription- and replication-associated substitutional asymmetries.** Coefficients τ_R^a, τ_T^a of the linear model Eq. (2), obtained by least-squares fits to a line in Fig. 2.

where v is the replication fork velocity and T_S the duration of S-phase. It is important to note that the replication program, and consequently the MRT profile and the replication fork polarity, are cell type specific. We would like to have access to the replication fork polarity in the germline, as only mutations occurring in germline cells are transmitted to the descendants, but unfortunately no experimental data in the germline is available today. As a substitute to germline replication fork polarity, we used the replication fork polarity determined in HeLa cell line, where the replication fork velocity $v = 0.64$ kbp/min has been measured by DNA combing and where the S-phase duration was estimated to be $T_S \sim 7$ h (Rappailles et al. 2011). The conservation of the replication fork polarity profile across differentiation will be addressed in Section III.3. Substitution rates were computed separately in genic (+), intergenic and genic (−) regions of given HeLa replication fork polarity values. As shown in Fig. 2, PR1 is not only broken in genic regions (red and blue) but also in intergenic regions (black). Furthermore the substitutional asymmetry in intergenic region is proportional to the HeLa replication fork polarity. In genic (+) (resp. (−)) regions, we recover the same linear behaviour adding up (resp. subtracting down) a constant corresponding to the transcription-associated asymmetry evidenced in Fig 1C. The substitutional asymmetry τ^a is therefore consistent with the following model:

$$\tau^a = \begin{cases} p\tau_R^a + \tau_T^a & \text{genic } (+) \\ p\tau_R^a & \text{intergenic} \\ p\tau_R^a - \tau_T^a & \text{genic } (-) \end{cases}, \tag{2}$$

in agreement with the minimal model for substitutional asymmetry proposed in Chapter I (Eq. (167) in the summary). The coefficients τ_T^a and τ_R^a, estimated by least-squares fits to a line (dashed lines in Fig. 2), are reported in Table 1. These results clearly support (i) that a replication-associated substitutional asymmetry does exist, (ii) that this replication-associated asymmetry is found in intergenic as well as in genic regions, and (iii) that the replication-associated asymmetry is proportional to the replication fork polarity (determined in the HeLa cell line). Furthermore, as reported in Table 2, the substitutional asymmetries correlate significantly with the replication fork polarity (and thus dMRT/dx) in intergenic regions, even though the replication fork polarity was determined in HeLa and not in the germline. Interestingly, the substitutional asymmetries do not correlate with the MRT ($R < 0.02$, p-value > 0.5), which is a strand-symmetric variable, while they do correlate with dMRT/dx which is a strand-asymmetric variable. On the

	$(A \to G)^a$	$(C \to T)^a$	$(C \to G)^a$	$(G \to T)^a$
p (HeLa)	0.30	0.17	0.19	0.09

Table 2: **Substitutional asymmetry correlates with the replication fork polarity.** Pearson correlation (R values) between the substitutional asymmetries and the replication fork polarity p in HeLa cell line. Substitutional asymmetries and p were calculated in non-overlapping 1Mbp windows genome wide. For substitution rates we only retained intergenic nucleotides. Only 1Mbp windows containing at least 100 kbp of aligned (intergenic) sequence were retained (N=2123). All p-values are $< 10^{-15}$ except for $(G \to T)^a$ (p-value $= 3\,10^{-5}$).

opposite the symmetrical substitution rates highly correlate with the MRT (Stamatoyannopoulos et al. 2009; Chen et al. 2010), but not with dMRT/dx ($R < 0.02$, p-value > 0.5). Therefore, as those correlations highlight, it is relevant to distinguish between strand-symmetric and strand-asymmetric variables. Mugal et al. (2009; 2010) reported that the substitutional asymmetry correlates strongly with the relative distance to skew N-domains borders (presented in Chapter IV). In our current perspective (Chapters IV and V), the relative distance to N-domains borders is directly related to the replication fork polarity in the germline. So far the substitutional asymmetry follows closely the model proposed in Chapter I (Eq. (167) in the summary): a replication-associated asymmetry proportional to the replication fork polarity, and a transcription-associated which adds to it. The estimates obtained for $(A \to G)^a_R > 0$, $(C \to T)^a_R > 0$, $(C \to G)^a_R > 0$, and $(G \to T)^a_R > 0$ replication-associated asymmetries (Table 1) are in agreement with previous studies (Polak and Arndt 2009; Mugal et al. 2009, 2010; Chen et al. 2011). We finally note that the $(C \to T)^a_R > 0$ replication-associated asymmetry is stronger than, and opposite to, the $(C \to T)^a_T < 0$ transcription-associated one (Table 1).

Symmetrical substitution rates are lower in genic region than in their flanking intergenic regions

In Fig. 1B, the average symmetrical substitution rates in genic regions were found to be lower than the corresponding rates in intergenic regions. However the genic and intergenic nucleotides could belong to genomic regions that do not share, even on average, the same characteristics. As substitution rates may depend on many variables, *e.g.* replication timing (Stamatoyannopoulos et al. 2009; Chen et al. 2010), the lower rates in genic region may not be directly associated to transcription, but could simply reflect that genes tend to belong to early replicating genomic regions. Therefore to further test if the lower symmetrical substitution rates could be attributed to transcription, we performed a regional analysis of substitution rates along large (> 100 kbp) human genes. In Fig. 3, a given substitution rate computed on the coding strand is displayed in purple, while the same substitution rate computed on the transcribed strand is displayed in orange. Equivalently, the orange curve is also equal to the reverse complementary substitution rate computed on the coding strand. We first note on Fig. 3A that there is a strong $(A \to G)^a_T > 0$ asymmetry

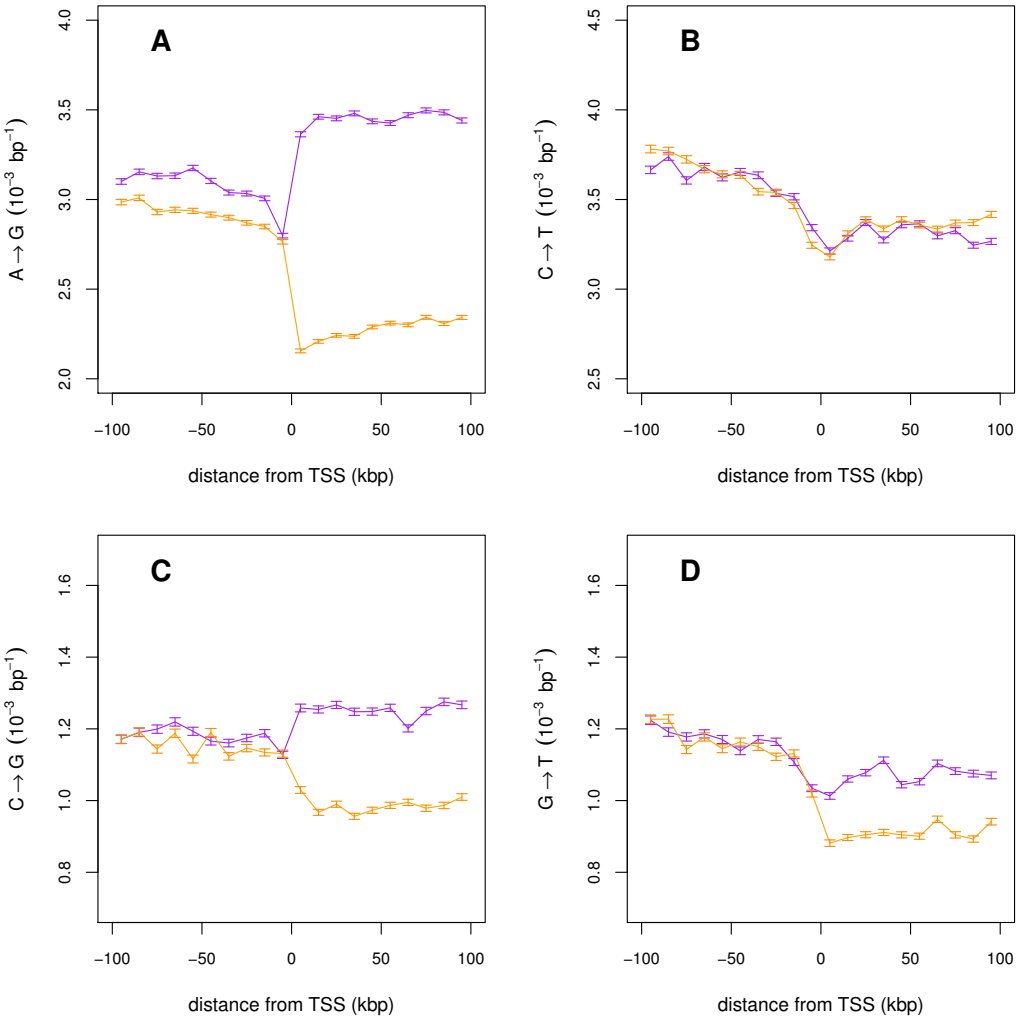

Figure 3: **Substitution rates along large (>100 kbp) human genes.** Average substitution rates in large human genes were computed every 10 kbp from 100 kbp upstream to 100 kbp downstream of the TSS. As genes are larger than 100 kbp, data points at 0 kbp $<$ distance to TSS $<$ 100 kbp correspond to the interior of the gene. For data points in the flanking intergenic region -100 kbp $<$ distance to TSS $<$ 0 kbp, we only retained intergenic nucleotides (as defined by the RefGene table). The substitution rates, and the distance to TSS are defined with respect to the coding strand of the gene (see Section I.1.3). (A) $A \to G$ substitution rate (purple) and the reverse complementary $T \to C$ substitution rate (orange), computed on the coding strand. Equivalently the orange curve corresponds to the $A \to G$ substitution rate computed on the transcribed strand (see Section I.1.3). (B) Same as in (A) but for the $C \to T$ substitution rate. (C) Same as in (A) but for the $C \to G$ substitution rate. (D) Same as in (A) but for the $G \to T$ substitution rate.

(the purple curve is above the orange one), which extends on the whole transcript, as previously observed (Green et al. 2003; Polak and Arndt 2008). Similarly significant $(C \to G)^a_T > 0$ and $(G \to T)^a_T > 0$ asymmetries are observed along the whole transcript (Figs. 3C and 3D). In contrast, no significant asymmetry is observed for the $C \to T$ substitution rate (Fig. 3B). Note that each data point corresponds to a 10 kbp bin, therefore our scale of analysis is too coarse to resolve the strong but localized $(C \to T)^a_T > 0$ asymmetry, restricted to the first 2 kbp downstream of the TSS (Polak and Arndt 2008). The regional variation of substitution rates observed in Fig. 3 thus confirms the transcription-associated substitutional asymmetries reported in Fig. 1C. We then note that for the $C \to T$ and $G \to T$ substitutions, the rates both in the transcribed and coding strands inside the gene are lower than the rate observed in the flanking intergenic region. Therefore for the $C \to T$ and $G \to T$ substitutions, the symmetrical part is clearly lower inside the gene than in the flanking intergenic region. This observation confirms the significant $(C \to T)^s_T < 0$ and $(G \to T)^s_T < 0$ reported in Fig. 1B. The variation of the $A \to G$ rate (Fig. 3A) and the $C \to G$ rate (Fig. 3C) are compliant with, but not demonstrative of, the weak $(A \to G)^s_T < 0$ and $(C \to G)^s_T < 0$ reported in Fig. 1B. Indeed, the symmetrical part of the $A \to G$ substitution rate is not significantly different in the genic and the flanking intergenic region, as previously observed in (Green et al. 2003). Finally we note a residual $(A \to G)^a_T > 0$ asymmetry in the flanking intergenic region in Fig. 3A. This is likely due to unannotated transcripts in the RefGene gene annotation table. The flanking "intergenic" region probably contains some unannotated transcripts, co-oriented with the gene. Note however that the $(A \to G)^a_T > 0$ asymmetry observed in the flanking intergenic region is ten fold lower than the asymmetry observed inside the gene, which suggests that unannotated transcripts are not numerous enough to affect our previous observations.

☛ *The substitutional asymmetry is compliant with the model proposed in Chapter I. The replication-associated asymmetry is proportional to the replication fork polarity determined in HeLa cell line (Fig. 2). The transcription-associated asymmetry adds to the replication-associated one and changes sign with gene orientation (Fig. 2). The symmetrical substitution rate is smaller in genic region than in intergenic region (Figs. 1B and 3), especially for the $C \to T$ and $G \to T$ substitutions.*

III.2 From substitutional to compositional asymmetry

If the substitutional asymmetry follows the decomposition observed in Fig. 2 and formalized in Eq. (2), we expect in turn the same decomposition for the compositional asymmetry, as measured by the GC and TA skews (Chapter I, Eq. (165) in the summary). In Section I.4.2 we formally studied the DNA composition evolution of a sequence submitted to substitutional asymmetries following Eq. (2). Such substitutional asymmetries give theoretically rise to transcription- and replication-associated skews in the DNA sequence, the latter being at all time proportional to

	S^*_{GC}	S^*_{TA}	S_{GC}	S_{TA}
S_T (%)	7.02 ± 0.16	10.80 ± 0.16	3.12 ± 0.05	4.23 ± 0.06
S_R (%)	10.54 ± 0.82	13.64 ± 0.85	6.06 ± 0.27	6.09 ± 0.31

Table 3: **Transcription- and replication-associated compositional asymmetries.** Coefficients S_R, S_T of the linear model Eq. (3), obtained by least-squares fits to a line in Fig. 4.

the replication fork polarity. In this Section, we check that the substitutional asymmetries found in Section III.1 generate transcription- and replication-associated GC and TA skews.

The compositional asymmetry decomposes into transcription- and replication-associated components

The equilibrium GC and TA skews, which are directly computed from the substitution rate matrix, can be interpreted as the current direction of evolution of the skews. As shown in Figs. 4A and 4B, the equilibrium skews S^*_{GC} and S^*_{TA} indeed decompose into transcription- and replication-associated components, consistent with the formal derivations made in Chapter I (Eq. (108)). If the current substitutional pattern is representative of the substitutional patterns that have shaped our genome, we expect the GC and TA compositional skews observed presently to follow the same decomposition. This is verified in Figs. 4C and 4D, where the compositional skews S_{GC} and S_{TA} are shown to decompose into transcription- and replication-associated components. Importantly, both equilibrium and compositional skews are proportional to the replication fork polarity. The compositional asymmetry S (where S denotes generically S^*_{GC}, S^*_{TA}, S_{GC}, or S_{TA}) is therefore consistent with the following model:

$$S = \begin{cases} pS_R + S_T & \text{genic } (+) \\ pS_R & \text{intergenic} \\ pS_R - S_T & \text{genic } (-) \end{cases}, \tag{3}$$

in agreement with the minimal model for the compositional asymmetry proposed in Chapter I (Eq. (168) in the summary). The coefficients S_T and S_R, estimated by least-squares fits to a line (dashed lines in Fig. 4), are reported in Table 3. We found positive $S_{TA,T}$ and $S_{GC,T}$ skews associated to transcription, as well as positive $S_{TA,R}$ and $S_{GC,R}$ skews associated to replication, in agreement with previous analyses (Touchon et al. 2003, 2004; Brodie of Brodie et al. 2005; Touchon et al. 2005). As reported in Table 4, both equilibrium and compositional skews correlate significantly with the replication fork polarity, even though the replication fork polarity was determined in HeLa and not in the germline. By contrast, the equilibrium and observed skews do not correlate with the MRT ($R < 0.02$ and $p > 0.45$), which is strand-symmetric. If we compare the numerical values in Figs. 4A and 4B, the observed compositional skews are two fold lower than the equilibrium skews shown in Figs. 4C and 4D. The compositional skews have clearly not reach equilibrium yet.

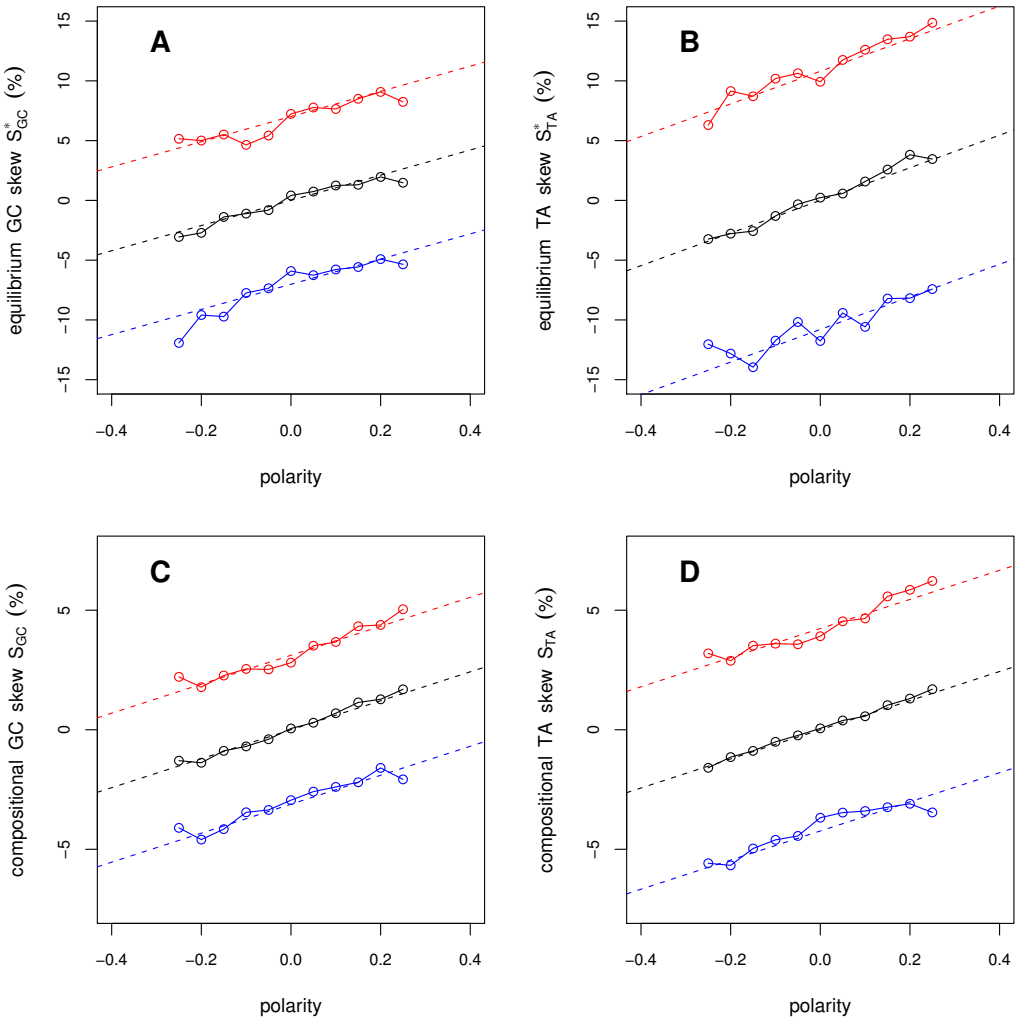

Figure 4: **The compositional asymmetry decomposes into transcription- and replication-associated components.** Compositional asymmetry versus the replication fork polarity (determined in HeLa cell line) in genic sense (red), intergenic (black), and genic antisense (blue) regions, for (A) the equilibrium GC skew, (B) the equilibrium TA skew, (C) the compositional GC skew, and (D) the compositional TA skew. The equilibrium composition was directly computed from the substitution rate matrix (Section I.3.1). For the compositional skews we only retained repeat-masked sequences. Equilibrium and compositional skews, replication fork polarity, and gene orientation were computed on the reference strand. The dashed lines correspond to the least-squares fits to a line, following the linear model Eq. (3). The linear regression coefficients are reported in Table 3.

	S^*_{GC}	S^*_{TA}	S_{TA}	S_{GC}
p (HeLa)	0.22	0.30	0.47	0.49

Table 4: **The compositional asymmetry correlates with the replication fork polarity.** Pearson correlation (R values) of equilibrium and observed compositional skews with the replication fork polarity. S^*_{TA}, S^*_{GC}, S_{TA}, S_{GC}, and p were calculated in non-overlapping 1Mbp windows genome wide. For substitution rates and sequence composition we only retained intergenic nucleotides. Only 1 Mbp windows containing at least 100 kbp of aligned (intergenic) sequences and at least 100 kbp of repeat-masked (intergenic) sequences were retained (N=1982). All p-values are $< 10^{-16}$.

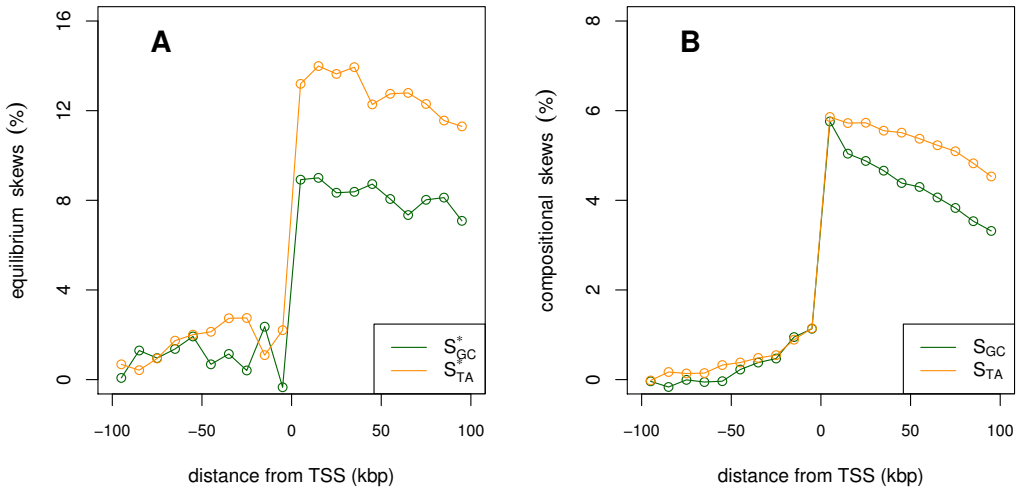

Figure 5: **Equilibrium and compositional skews along large (> 100 kbp) human genes.** Equilibrium skews (A) and compositional skews (B) versus the distance to the TSS. Average substitution rates and nucleotide composition in large human genes were computed very 10 kbp from 100 kbp upstream (distance to TSS = -100 kbp) to 100 kbp downstream of the TSS (distance to TSS = +100 kbp). As genes are larger than 100 kbp, data points at 0 kbp < distance to TSS < 100 kbp correspond to the interior of the gene. For data points in the flanking intergenic region -100 kbp < distance to TSS < 0 kbp, we only retained intergenic nucleotides (as defined by the RefGene table). For the nucleotide composition we only retained repeat-masked sequences. The substitution rates, the nucleotide composition, and the distance to TSS are defined with respect to the coding strand of the gene.

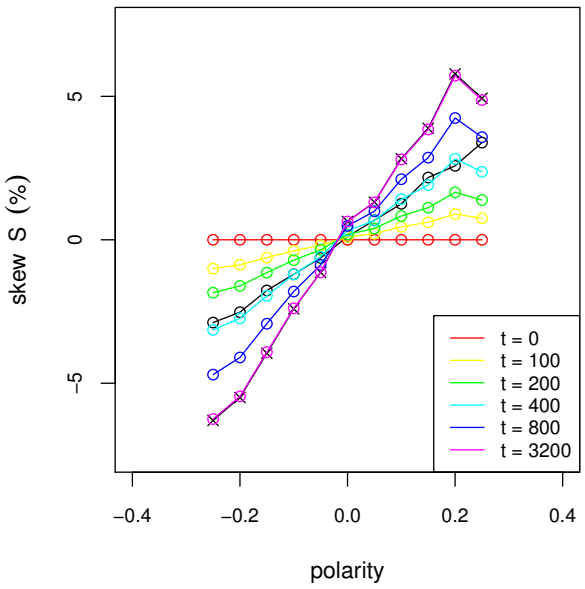

Figure 6: **Establishment of the compositional skew is a very slow process.** Time evolution of the total skew $S = S_{TA} + S_{GC}$, from initially null skews, under the current substitutional pattern obtained in intergenic regions for different replication fork polarity values in HeLa (Fig. 2). Time is indicated in Myrs; the time evolution was computed according to the neighbor-independent and time-homogeneous model of DNA composition evolution presented in Section I.3.1. Note that the skew S obtained at $t = 400$ Myrs (light blue curve) matches the observed compositional skew (black circles), whereas the equilibrium skew (black cross curve) is almost reached at $t = 3200$ Myrs (magenta curve).

As shown in Fig. 5 along large (> 100 kbp) human genes, the GC and TA skews extend on the whole transcript (Touchon et al. 2003, 2004).

The observed compositional skews were generated over several hundreds Myrs

In this paragraph we investigate the dynamics of compositional skew evolution. The convergence of the compositional skews towards equilibrium is governed by the time-scales τ_A, $\tau_D^{(1)}$, and $\tau_D^{(2)}$ introduced in Chapter I (Eqs. (73) and (75)). For the current substitution rates, these time-scales are of several hundreds Myrs. We can give however a more illustrative time-scale, defined as the time necessary to generate the observed compositional skews in a sequence exposed to the current substitutional pattern. As shown in Fig. 6, if we start from initial null skews, and if the sequence is submitted to the substitution rates found in intergenic region at given replication fork polarity (Fig. 2), the compositional skews increase over time. It is equal to the observed compositional skew (black circles in Fig. 6) after 400 Myrs. It almost reaches equilibrium after three billion years (black crosses in Fig. 6). Interestingly,

the estimated time to reach the observed skew (400 Myrs) is a bit larger than the last common ancestor of amniotes ($\sim$ 350 Myrs). Note that this estimation is somehow qualitative. Indeed the substitutional pattern that has generated the observed skew might have changed over time, and the current substitutional pattern may not faithfully reflect the substitutional pattern of these past 400 Myrs. For instance, the excellent correlation found by Mugal et al. (2009) between the substitutional asymmetry and the compositional skew implies that their substitutional pattern, determined further in the past than the human-chimpanzee divergence we considered in this manuscript, reflects more faithfully the substitutional pattern that has generated the skew. Let us also mention that the substitution rates might have been higher in the past, which would have transiently accelerated the skew evolution. Nonetheless, these observations clearly indicate that the skew evolution is a very slow process. The current and quite high value of the compositional skew requires a persistent direction of skew evolution, over several hundreds Myrs. Interestingly, the substitutional asymmetry is well conserved between human and mouse (Mugal et al. 2010), which consistently indicates that the substitutional asymmetry is well conserved on evolutionary time scales. In turn this suggests that the determinants of the substitutional asymmetry (the replication fork polarity for instance), which determine the direction of skew evolution, must have been well conserved over such time-scales. Indeed the replication timing, which determines the replication fork polarity, has been well conserved at least since the human-mouse divergence (Ryba et al. 2010; Yaffe et al. 2010).

Taking into account the $CpG \to TpG$ substitution

The neighbor-dependent $CpG \to TpG$ substitution is 13 fold more frequent than $C \to T$, the most frequent single nucleotide substitution according to Fig. 1A. As shown in Fig. 7A, the $(CpG \to TpG)^a$ asymmetry decomposes into transcription- and replication-associated components. There is a strong replication-associated asymmetry $(CpG \to TpG)^a_R = 13.2$ kbp^{-1} and a very weak transcription-associated asymmetry $(CpG \to TpG)^a_T = 0.9$ kbp^{-1}. In Section I.4.3, we studied the evolution of the DNA composition when taking into account the $CpG \to TpG$ substitution rate, using the neighbor-dependent model proposed in (Arndt et al. 2003). The neighbor-dependency does not affect the decomposition of the skew into transcription- and replication-associated components (Eqs. (138) and (141)), as confirmed for the equilibrium skew in Fig. 7B. We did not explore the time evolution in the neighbor-dependent model. A limiting issue is the great number of unknown parameters (the initial dinucleotide frequencies), even if we impose PR2 for the initial composition.

☛ *The compositional asymmetry is compliant with the model proposed in Chapter I. The replication-associated asymmetry is proportional to the replication fork polarity determined in HeLa cell line (Fig. 4). The transcription-associated asymmetry adds to the replication-associated one and changes sign with gene orientation (Fig. 4). This is true both for the equilibrium skew that gives the current direction of the skew*

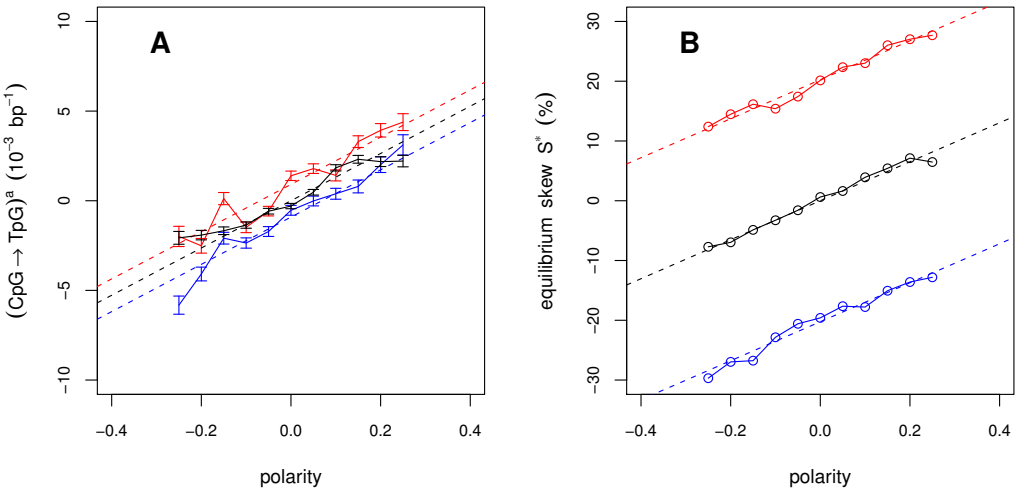

Figure 7: **Taking into account the neighbor-dependent $CpG \rightarrow TpG$ substitution rate.** $(CpG \rightarrow TpG)^a$ substitutional asymmetry (A) and equilibrium skew S^* (B) versus replication fork polarity (determined in HeLa cell line as in Fig. 2) in genic sense (red), intergenic (black), and genic antisense (blue) regions. The equilibrium composition was computed taking into account, besides the single-nucleotide substitution rate matrix, the neighbor-dependent $CpG \rightarrow TpG$ substitution rate (Section I.4.3), using the neighbor-dependent model proposed in (Arndt et al. 2003). Substitution rates, equilibrium skew, replication fork polarity, and gene orientation were computed on the reference strand.

evolution, and for the observed compositional skew. The compositional skew is not at equilibrium, and the skew evolution is an extremely slow process (time-scales of several hundreds Myrs). With the current substitutional pattern, it requires 400 *Myrs to generate the current values of the skew from initially null skews (Fig. 6).*

III.3 Discussion

Effect of gene expression on substitution rates

In the model proposed in Chapter I, the transcription-associated asymmetry increases with the transcription rate α. It is understood that the asymmetry should increase with the germline transcription rate, as only mutations occurring in the germline are transmitted to the descendants. The strand asymmetry could, or not, correlate with gene expression in somatic cells depending on the conservation of gene expression over differentiation. Various analyses already support the link between strand asymmetry and germline expression level. As reported in (McVicker and Green 2010), the $(A \rightarrow G)^a$ asymmetry and the $G + T$ content on the coding strand strongly correlate with germline expression level. Note that these correlations are higher than those previously reported between the $G + T$ content and housekeeping genes expression (Majewski 2003), expression in testis (Comeron 2004), and

breath of expression (Duret 2002), used as indirect estimators of the expression in the germline. During the male germline, the time spent as a spermatogonia cell is probably the longest, thus the gene expression in spermatogonia is expected to have the greatest impact on the transcription-associated strand asymmetry. Interestingly, the $(A \rightarrow G)^a$ asymmetry and the $G + T$ content most strongly correlate with the expression in spermatogonia (McVicker and Green 2010). On the opposite, the $(C \rightarrow T)^a$ asymmetry does not correlate significantly with the expression in germline cells (McVicker and Green 2010). Our results suggest that the $(C \rightarrow T)^a$ asymmetry, in transcribed and non-transcribed regions, is mainly driven by the replication fork polarity (Fig. 2B), which could explain the poor correlation observed in (McVicker and Green 2010). The correlation could also be affected by the variation of the $(C \rightarrow T)^a$ asymmetry along transcripts, as observed in (Polak and Arndt 2008). Note that the poor correlation between the $(C \rightarrow T)^a$ asymmetry and gene expression only applies to the most recent substitutional pattern, as estimated since the human-chimpanzee divergence. In contrast, with substitution rates estimated further in the past, Mugal et al. (2010) reported a strong correlation between the $(C \rightarrow T)^a$ asymmetry and gene expression. Interestingly, the substitutional asymmetry in genes correlates with both germline expression and the relative distance to skew N-domain borders (estimator of replication fork polarity, see Chapter IV) (Mugal et al. 2009, 2010). According to (Mugal et al. 2010), a linear model based on these two predictors has the best explanatory power which strongly supports the model proposed in Chapter I for substitutional asymmetry.

To my knowledge, the lower symmetrical substitution rates in genes were not previously reported. Moreover, the relationship with germline gene expression has neither been investigated. We noted that the strong to weak $C \rightarrow T$ and $G \rightarrow T$ substitutions were the most affected (Fig. 1B). We argue that the reduced rates are most likely due to some repair mechanism associated with transcription. A higher selective pressure in genes introns could induce a lower total substitution rate, but a priori, there is no reason to disfavor systematically the strong to weak substitutions. Biased-gene conversion (BGC), a neutral process which favors the fixation of GC rich alleles, can neither be invoked, as it impacts on weak to strong substitution rates, but not on strong to weak substitution rates (Duret and Arndt 2008). However BGC, along with reduced recombination rates observed in genes (McVicker and Green 2010), could explain the weakly reduced weak to strong $(A \rightarrow G)^s$ symmetrical substitution rate (Fig. 1B). We conjecture that if the rates are reduced in genes due to some repair mechanism associated with transcription, the reduction should be greater for the most expressed genes.

Are ORIs location relevant for the study of strand asymmetry in eukaryotes?

First studies on strand asymmetry and replication in human were contradictory and inconclusive. A replication-related strand asymmetry was reported around the

replication origin in the β-globin locus (Wu and Maeda 1987), but was not confirmed by further analyses (Bulmer 1991; Francino and Ochman 2000). Until recently very few (~ 30) human replication origins were known experimentally (Aladjem 2007). This is considerably less than the number (about $\sim 10^4$) of initiation sites needed each cell cycle to ensure complete duplication of the human genome. It was thus difficult to make the same analysis as in bacteria, defining locally leading and lagging strands in between unknown replication origins (Wu 1991). In our current perspective, it would have been even incorrect to make the same analysis as in bacteria. In contrast to bacteria, genomic regions in eukaryotes cannot be unambiguously assigned as leading or lagging strands (replication fork polarity $p = +1$ or $p = -1$). In eukaryotes, the replication program is more complex: several replication origins, firing at different times during the S-phase, and not always at the same positions and times over cell cycles. A loci is in general replicated by a proportions $p_{(\pm)}$ of $(\pm)$ forks over cell cycles, and the replication fork polarity $p = p_{(+)} - p_{(-)}$ can take values in the whole interval $[-1, 1]$.

Let us point out that it would be mathematically intractable to infer the replication fork polarity profile only from the locations of replication origins. Indeed the replication fork polarity at a locus depends on the locations of neighboring replication origins, on their efficiencies, on their firing time distributions, and on the combinatorial usage of replication origins. Most of these quantities are out of reach by current experimental abilities. In Chapter II, the replication fork polarity was shown to be proportional to the derivative of the mean replication timing (Eq. (68)). We pointed out that this result is exact provided that (i) the replication fork velocity is constant, and (ii) that replication origins are bidirectional, which seems rather reasonable assumptions given the information at hand. For comparison, Necsulea et al (2009) proposed to segment the human genome into leading and lagging strands as in bacteria, in between replication origins experimentally determined by (Cadoret et al. 2008). This segmentation not only assumes (i) and (ii), but further requires that all experimental origins are 100% efficient and that they all fire synchronously at the onset of S-phase, which looks rather drastic and unrealistic assumptions in higher eukaryotes. Using this segmentation into leading and lagging strands, Necsulea et al (2009) did not observe any strand asymmetry around experimental replication origins, and concluded to the nonexistence of replication-associated strand asymmetry.

However, a specific set of replication origins can be used to reveal replication-associated strand asymmetry. Actually we expect to observe a significant replication-associated strand asymmetry for very efficient and well-positioned replication origins, active in the germline, and well evolutionary conserved. Using either these very efficient replication origins or the relationship between replication fork polarity and the mean replication timing (Chapter II, Eq. (68)) thus appears as two different strategies, likely operative at different scales. The strand asymmetry directly asso-

ciated to a very efficient origin is expected to be localized around the origin, and not to extend beyond the closest termination sites. Loci too far upstream or downstream are replicated by forks coming from other initiation sites. On the opposite, the mean replication timing permits to study strand asymmetry at larger scales. Note that, due to the current resolution of replication timing data in human cell lines, we determined the dMRT/dx profile at the 100 kbp scale. At this scale, comparable to the typical replicon size (Berezney et al. 2000), the proximity to an initiation site is no longer relevant: each data point contains on average one initiation site. Even at the 100 kbp scale, many genomic regions have a significantly non null replication fork polarity. Indeed we observed in the human genome large scale gradients of replication timing, from hundreds of kbp to several Mbp (Rappailles et al. 2011). In these regions and at such scales, the strand asymmetry is presumably not generated by the proximity to a replication origin, but more likely due the average temporal order of replication origin firings.

Finally we note that replication timing data are now available in an increasing number of organisms (yeast (Raghuraman et al. 2001), drosophila (Schübeler et al. 2002), mouse (Hiratani et al. 2008), human (Woodfine et al. 2005; Hansen et al. 2010)). Hence, as regards future work, the relationship between replication fork polarity and replication timing (Chapter II, Eq. (68)) can further be used to study replication-associated strand asymmetry in various eukaryotic genomes.

The good conservation of dMRT/dx across differentiation ensures the robustness of our analysis

In Section III.2, we analysed strand asymmetry using the replication fork polarity determined in the HeLa cell line, as a substitute to germline replication fork polarity. In other cell lines (data from Hansen et al. 2010), in contrasts to HeLa (data from Rappailles et al. 2011), we had not access to both the replication fork velocity v and duration of S-phase T_S, and we were therefore not able to convert the dMRT/dx profile into a replication fork polarity profile using Eq. (1). Nonetheless in any examined cell line, we robustly observed that the substitutional and compositional asymmetries decompose into transcription- and replication-associated components, the latter being proportional to dMRT/dx, as exemplified in Fig. 8 for the equilibrium skew S^* and the compositional skew S in the BG02 embryonic stem cell line and in the GM06990 lymphoblastoid cell line. In intergenic regions, both the equilibrium and compositional skews correlate significantly with the dMRT/dx profile in any examined cell line (Table 5). We infer from the good correlation obtained between the dMRT/dx profiles of the different cell lines that they all correlate with the dMRT/dx profile (and consequently the replication fork polarity) in the germline. This explains, *a posteriori*, why we were able to measure replication-associated asymmetries, even with replication fork polarity profiles not estimated in the germline. Interestingly the correlation between the compositional skew and the different dMRT/dx profiles is as high as between the dMRT/dx profiles them-

	BG02	GM06990	K562	BJ	HeLa
BG02	1	0.59	0.62	0.52	0.57
GM06990	0.59	1	0.73	0.61	0.64
K562	0.62	0.73	1	0.57	0.63
BJ	0.52	0.61	0.57	1	0.73
HeLa	0.57	0.64	0.63	0.73	1
S	0.61	0.60	0.62	0.49	0.52
S^*	0.41	0.41	0.44	0.33	0.35

Table 5: **Conservation of dMRT/dx across differentiation.** Pearson correlation (R values) between dMRT/dx profiles from various cell lines: B0G2 embryonic stem cell, GM06990 lymphoblastoid, K562 erythroid, BJ fibroblast, and HeLa cell lines. For comparison, are also reported for each of these cell lines the Pearson correlation between the compositional skew S and the equilibrium skew S^* and dMRT/dx. S, S^*, and dMRT/dx were calculated genome wide in non-overlapping 1Mbp windows using replication timing data from (Hansen et al. 2010; Rappailles et al. 2011). For substitution rates and sequence composition we only retained intergenic nucleotides. Only 1 Mbp windows containing at least 100 kbp of aligned (intergenic) sequences and at least 100 kbp of repeat-masked (intergenic) sequences were retained (N=1982). All p-values are $< 10^{-16}$.

selves (Table 5).

Estimation of τ_R^a and τ_T^a coefficients for different cell lines

In each cell line, we performed the least-squares fits to a line of the substitutional and compositional asymmetries according to the minimal model Eqs. (2) and (3), as previously done for the HeLa cell line in Figs. 2 and 4. As reported in Table 6, the estimated transcription-associated asymmetries for stem cells, somatic cells and HeLa cells are in remarkable quantitative agreement. Unfortunately, for the replication-associated asymmetry, the linear regression versus the dMRT/dx values does not directly give access to τ_R^a or S_R, but only to $vT_S\,\tau_R^a$ or $vT_S\,S_R$. In all cell lines with the exception of HeLa (where v and T_s are known), we could only estimate $vT_S\,\tau_R^a$ and $vT_S\,S_R$ (Table 7). The replication fork velocity and the duration of S-phase are likely cell type specific, so the factor vT_S likely depends on the cell line considered. Consistently, the replication-associated asymmetries $vT_S\,\tau_R^a$ and $vT_S\,S_R$ determined in each considered cell line were found to be proportional to τ_R^a and S_R determined in the HeLa cell line (Fig. 9). Unfortunately we cannot assert (see next paragraph) that the linear regression versus the replication fork polarity, in each cell line, would yield the same estimates of τ_R^a and S_R as found in HeLa cells, and as possibly found in the germline. If we make however this strong assumption, the coefficients of proportionality reported in Fig. 9 would imply that $vT_S = 0.24$ Mbp in BG02 embryonic stem cells and $vT_S = 0.46$ Mbp in GM06990 lymphoblastoid cells, as compared to $vT_S = 0.27$ Mbp in HeLa cells as measured in (Rappailles et al. 2011).

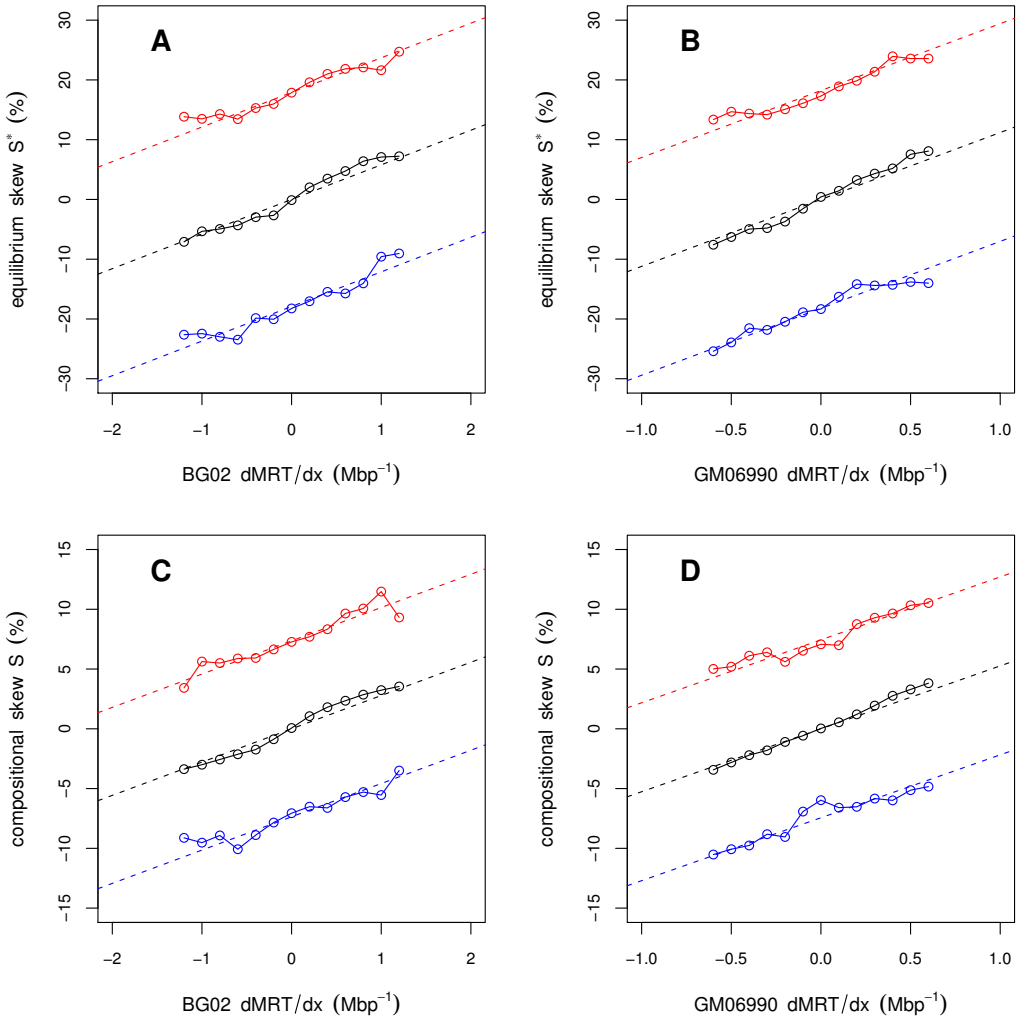

Figure 8: **The decomposition of S and S^* into transcription- and replication-associated components is observed for all examined cell lines.** (A) Equilibrium skew S^* versus dMRT/dx in BG02 embryonic stem cell line for genic sense (red), intergenic (black), and genic antisense (blue) regions. (B) Same as in (A) but using dMRT/dx determined in the GM06990 lymphoblastoid cell line. (C) Compositional skew S versus dMRT/dx in BG02 embryonic stem cell line for genic sense (red), intergenic (black), and genic antisense (blue) regions. (D) Same as in (C) but using dMRT/dx determined in the GM06990 lymphoblastoid cell line. Equilibrium and observed compositional skews, dMRT/dx, and gene orientation were computed on the reference strand. The dashed lines correspond to the least-squares fits to a line, following the linear model Eq. (3). The replication timing data were retrieved from (Hansen et al. 2010).

	BG02	GM06990	K562	BJ	HeLa
$(A \to G)^a_T$	5.47 ± 0.06	5.47 ± 0.07	5.44 ± 0.09	5.46 ± 0.05	5.41 ± 0.05
$(C \to T)^a_T$	-0.50 ± 0.04	-0.44 ± 0.05	-0.50 ± 0.04	-0.45 ± 0.04	-0.48 ± 0.05
$(C \to G)^a_T$	1.20 ± 0.03	1.20 ± 0.02	1.19 ± 0.03	1.19 ± 0.02	1.20 ± 0.02
$(G \to T)^a_T$	0.75 ± 0.03	0.76 ± 0.02	0.75 ± 0.03	0.74 ± 0.02	0.75 ± 0.02
$S^*_{GC,T}$	7.03 ± 0.14	7.10 ± 0.11	6.98 ± 0.15	7.06 ± 0.14	7.02 ± 0.16
$S^*_{TA,T}$	10.88 ± 0.21	11.12 ± 0.16	10.93 ± 0.18	11.08 ± 0.16	10.80 ± 0.16
$S_{GC,T}$	3.12 ± 0.07	3.19 ± 0.05	3.15 ± 0.05	3.21 ± 0.05	3.12 ± 0.05
$S_{TA,T}$	4.24 ± 0.06	4.26 ± 0.06	4.24 ± 0.06	4.31 ± 0.06	4.23 ± 0.06

Table 6: **Estimates of transcription-associated asymmetries using dMRT/dx in different human cell lines.** Transcription-associated asymmetries estimated by least-square fits to a line of the substitutional and compositional asymmetries, according to the linear model Eqs. (2) and (3), in BG02 embryonic stem cell, GM06990 lymphoblastoid cell, K562 erythroid cell, BJ fibroblast cell (Hansen et al. 2010), and HeLa cell (Rappailles et al. 2011) lines. For all cell lines, the linear regression was actually performed versus dMRT/dx values. The substitutional asymmetries are expressed in 10^{-4} bp^{-1} and the compositional asymmetries in %.

	BG02	GM06990	K562	BJ	HeLa
$vT_S\,(A \to G)^a_R$	0.97 ± 0.07	2.13 ± 0.16	2.18 ± 0.16	1.22 ± 0.08	1.15 ± 0.07
$vT_S\,(C \to T)^a_R$	0.98 ± 0.05	1.49 ± 0.1	1.42 ± 0.08	0.94 ± 0.07	0.95 ± 0.06
$vT_S\,(C \to G)^a_R$	0.30 ± 0.03	0.59 ± 0.05	0.58 ± 0.07	0.42 ± 0.04	0.32 ± 0.03
$vT_S\,(G \to T)^a_R$	0.08 ± 0.03	0.14 ± 0.05	0.22 ± 0.05	0.11 ± 0.04	0.09 ± 0.03
$vT_S\,S^*_{GC,R}$	2.56 ± 0.15	5.09 ± 0.25	4.85 ± 0.29	3.24 ± 0.23	2.83 ± 0.22
$vT_S\,S^*_{TA,R}$	3.23 ± 0.23	6.11 ± 0.35	6.34 ± 0.35	3.79 ± 0.26	3.67 ± 0.23
$vT_S\,S_{GC,R}$	1.47 ± 0.08	2.60 ± 0.11	2.79 ± 0.09	1.84 ± 0.08	1.63 ± 0.07
$vT_S\,S_{GC,R}$	1.31 ± 0.07	2.66 ± 0.12	2.66 ± 0.11	1.72 ± 0.1	1.64 ± 0.08

Table 7: **Estimates of replication-associated asymmetries using dMRT/dx in different human cell lines.** Replication-associated asymmetries estimated by least-squares fits to a line of the substitutional and compositional asymmetries, according to the linear model Eqs. (2) and (3), in BG02 embryonic stem cell, GM06990 lymphoblastoid cell, K562 erythroid cell, BJ fibroblast cell (Hansen et al. 2010), and HeLa cell (Rappailles et al. 2011) lines. For all cell lines, the linear regression was actually performed versus dMRT/dx values. The substitutional asymmetries are expressed in 10^{-4} bp^{-1} and the compositional asymmetries in %.

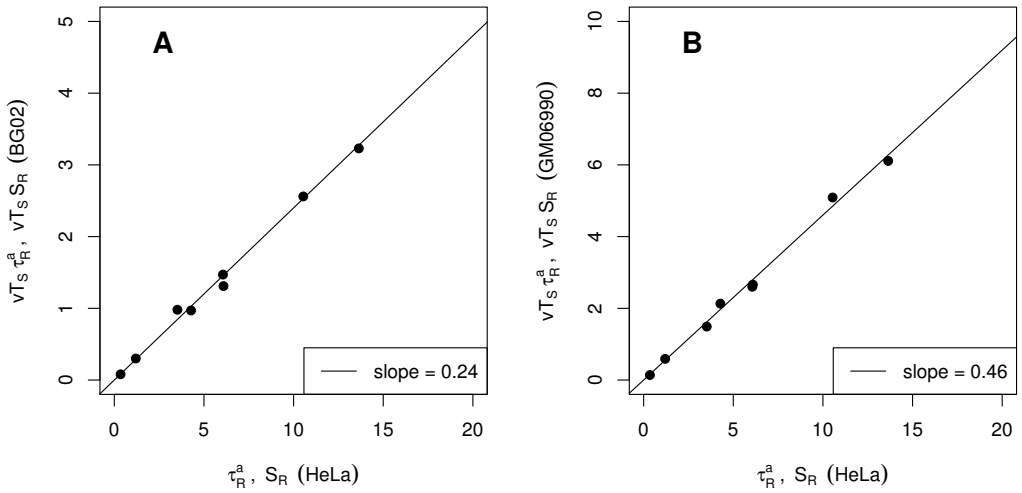

Figure 9: **Consistent estimates of replication-associated asymmetries when using different cell lines** Replication-associated asymmetries $vT_S\,\tau_R^a$ and $vT_S\,S_R$ estimated in BG02 embryonic stem cell line (A) and GM06990 lymphoblastoid cell line (B) versus the corresponding replication-associated asymmetries τ_R^a and S_R determined in HeLa cell line (Tables 1 and 3). The substitutional asymmetries are expressed in 10^{-4} bp^{-1} and the compositional asymmetries in %.

Are the replication-associated asymmetries overestimated?

The replication-associated asymmetries τ_R^a and S_R found using HeLa replication fork polarity are unexpectedly high, they are comparable and sometimes greater than the corresponding transcription-associated asymmetries (Tables 1 and 3). The τ_R^a and S_R asymmetries are theoretically the maximal replication-associated asymmetries observable in the human genome when the replication fork polarity $p = \pm 1$. Note that only a few genomic regions, if any, are expected to have $p = \pm 1$ replication fork polarity. Such genomic regions would have to be, at each cell cycle in the germline and on evolutionary time scales, always replicated by forks of the same directionality. Interestingly, as reported in Chapter IV, the replication-associated asymmetries observed at compositional skew upward jumps (S-jumps) in the human genome (Brodie of Brodie et al. 2005; Touchon et al. 2005) are about three fold lower than the coefficients τ_R^a and S_R obtained in the present study from HeLa cell replication timing data. For example, in intergenic regions downstream of S-jumps, the equilibrium and compositional skews are respectively equal to $S^* = 7.69\%$ and $S = 3.72\%$ (Chen et al. 2011), while the corresponding coefficients reported in Table 3 are equal to $S_R^* = 24.18\%$ and $S_R = 12.15\%$. This suggests that only a few genomic regions have a replication polarity (in the germline and integrated over evolutionary time scale) larger than $\sim 1/3$, provided that the coefficients τ_R^a and S_R have not been overestimated. We see two causes leading to a possible overestimation of the τ_R^a and S_R coefficients: (i) the underestimation of HeLa replication fork polar-

ity, and (ii) the non-conservation of replication fork polarity between HeLa and the germline. The replication fork polarity in HeLa was measured according to Eq. (1), and thus directly depends on the replication fork velocity v and duration of S-phase T_S. Thus an underestimation of v or T_S might have led to an underestimation of the replication fork polarity, and in turn to the overestimation the coefficients τ_R^a and S_R obtained by linear regression. For instance, if v would be equal to twice its value measured by DNA combing in HeLa cells (Rappailles et al. 2011), then the τ_R^a and S_R coefficients would be divided by two. The τ_R^a and S_R coefficients might also be overestimated if the germline replication fork polarity was, on average, larger than HeLa replication fork polarity. In Section III.2, we measured substitutional and compositional asymmetries in regions of fixed replication fork polarity in HeLa cells (p_{HeLa}). As the correlations reported in Table 5 suggest, in regions of given p_{HeLa} values, the average replication fork polarity in the germline (p_{germline}) is likely proportional to p_{HeLa}:

$$p_{\text{germline}} = K p_{\text{HeLa}}. \tag{4}$$

According to our minimal model (Chapter I, Eq. (167) in the summary), we expect to observe the following substitutional asymmetries:

$$\tau^a = \begin{cases} p_{\text{germline}}\tau_R^a + \tau_T^a & = p_{\text{HeLa}}\left(K\tau_R^a\right) + \tau_T^a & \text{genic } (+) \\ p_{\text{germline}}\tau_R^a & = p_{\text{HeLa}}\left(K\tau_R^a\right) & \text{intergenic} \\ p_{\text{germline}}\tau_R^a - \tau_T^a & = p_{\text{HeLa}}\left(K\tau_R^a\right) - \tau_T^a & \text{genic } (-) \end{cases} \tag{5}$$

Hence the coefficient $\tau_{R,\text{HeLa}}^a = K\tau_R^a$, as estimated by the linear regression versus p_{HeLa}, is expected to be proportional to τ_R^a. If $K > 1$ (resp. $K < 1$), the coefficients reported in Table 7 would actually overestimate (resp. underestimate) the replication-associated asymmetries.

The $(A \to G)_R^a$ asymmetry may require an extension of the minimal model

Although the decomposition of the substitutional and compositional asymmetries into transcription- and replication-associated components was robustly observed for all examined cell lines, we observed for many cell lines a deviation of the $(A \to G)^a$ asymmetry from the minimal model Eq. (2). As shown in Fig. 10 for the K562 erythroid and GM06990 lymphoblastoid cell lines, the replication-associated asymmetry is more pronounced in intergenic regions than in genic regions. The $(A \to G)^a$ is better described by the following extension of the minimal model:

$$(A \to G)^a = \begin{cases} p\,(A \to G)_{R,\text{genic}}^a + (A \to G)_T^a & \text{genic } (+) \\ p\,(A \to G)_{R,\text{intergenic}}^a & \text{intergenic} \\ p\,(A \to G)_{R,\text{genic}}^a - (A \to G)_T^a & \text{genic } (-) \end{cases}, \tag{6}$$

allowing for different slopes in intergenic and genic regions. Note that the slopes have to be the same in genic $(+)$ and $(-)$ regions due to strand-exchange symmetry (Chapter I, Eq. (4)). The least-squares fit of the $(A \to G)_R^a$ asymmetry according to Eq. (6) (dashed lines in Fig. 10) yields in all cell lines, a slope larger in intergenic

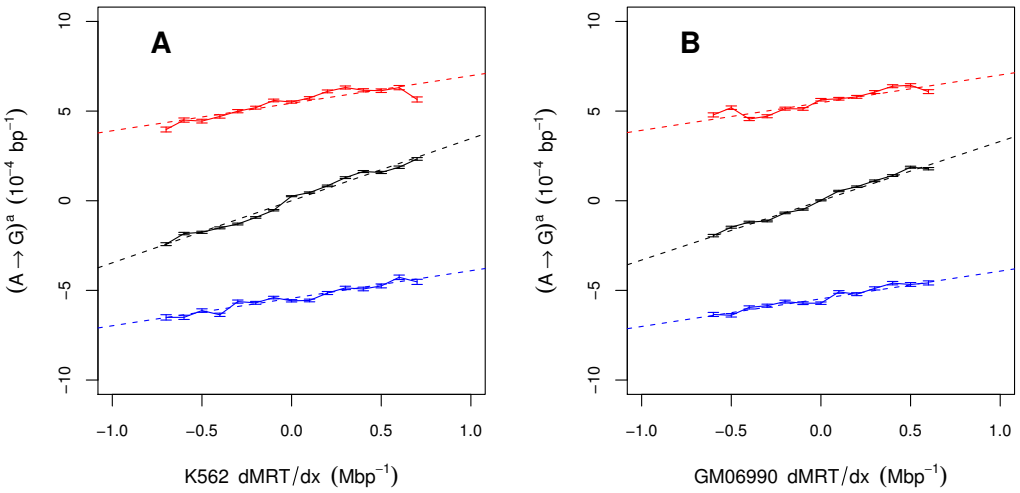

Figure 10: **The replication-associated $(A \to G)^a_R$ asymmetry is more pronounced in intergenic regions.** $(A \to G)^a$ asymmetry versus dMRT/dx in genic sense (red), intergenic (black), and genic antisense (blue) regions, where dMRT/dx was determined in the K562 erythroid cell (A), and GM06990 lymphoblastoid cell (B) lines. Substitution rates, dMRT/dx, and the gene orientation were computed on the reference strand. The dashed lines correspond to the least-squares fits to a line, following the linear model Eq. (6).

regions, from 1.57 fold higher in HeLa to 2.15 fold higher in GM06990 and 2.26 fold higher in K562. Finally, we note that a fixation bias due to biased gene conversion (BCG) can be proposed to explain such a trend. BCG, in regions of high recombination rate, favors the fixation of weak to strong substitution (Duret and Arndt 2008). The reduced crossover rates observed in human genes (McVicker and Green 2010) suggest that the weak to strong $A \to G$ and $T \to C$ substitutions have a higher fixation probability in intergenic regions. A higher fixation probability would amplify the substitutional asymmetry in intergenic regions.

Summary of Chapter III

In this Chapter, we provided clear evidence for replication-associated strand asymmetry in the human genome. The substitutional asymmetry (Fig 2) and the compositional asymmetry (Fig 4) follow closely the minimal model proposed in Chapter I: both decompose into transcription- and replication-associated components, the former changes sign with gene orientation, and the latter is proportional to the replication fork polarity. The replication fork polarity profiles correlate in different cell lines (Table 5), which explains *a posteriori* why we were able to provide evidence for replication-associated asymmetry, even if our replication fork polarity

estimates were not determined in the germline. Of course, we expect our analyses to be even more conclusive when using germline replication fork polarity. Interestingly, the establishment of the compositional skews is an extremely slow process. The compositional skews have not reached equilibrium yet. With the current substitutional pattern, it would require 400 Myrs to generate the observed compositional skews from initially null skews (Fig. 6). In turn this is an indication that the determinant of the skew evolution in intergenic region, namely the replication fork polarity, has been well evolutionary conserved over several hundreds Myrs. In our analyses of substitution rates, we also observed lower symmetrical rates in genic regions for the strong to weak $C \to T$ and $G \to T$ substitutions (Fig. 3). If further analyses confirm this trend, this will suggest that some repair mechanisms coupled to transcription have an important impact on the substitutional pattern found in genes.

Chapter IV

Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes

As proposed theoretically (Chapter I), and verified experimentally in the human genome (Chapter III), the compositional asymmetry decomposes into transcription- and replication-associated components, the latter being proportional to the replication fork polarity. In this Chapter, we develop a wavelet-based methodology to analyze the DNA strand asymmetry profiles with the specific goal to extract the contributions associated with replication and transcription. In a first step, we use an adapted N-shaped analyzing wavelet to perform a multi-scale pattern recognition analysis of the sum of the TA and GC skews along human chromosomes. This method provides an objective segmentation of the human genome in skew domains of $\simeq 1$ Mbp characteristic size, bordered by two putative replication origins recognized as large amplitude upward jumps in the noisy skew profile. In between the two upwards jumps, the skew decreases rather linearly as the signature of the progressive inversion of the replication fork polarity. In a second step, we use a least-square fitting procedure to disentangle, in these skew domains, the small-scale (the mean human gene size $\simeq 30$ kbp) square-like transcription component from the global N-shaped component induced by replication. When applying this procedure to the 22 human autosomes, we delineate 678 replication domains of mean length $\bar{L} = 1.2 \pm 0.6$ Mbp spanning 33.8% of the human genome and we predict 1062 replication origins. As a comparative analysis, we further apply our wavelet-based methodology to skew profiles along the mouse chromosomes. The striking similarity of the results in human and mouse indicates that skew N-domains are likely to be a general feature

of mammalian genomes. The transcription- and replication-associated components estimated by our disentangling methodology are shown to respectively correlate to the expression in the germline and to the replication fork polarity in various cell lines.

IV.1 Introduction: zooming in genomic sequences with the wavelet-transform "microscope"

The wavelet-transform: a mathematical microscope to track singularities and study multifractal distributions.

Since the pioneering works of J. Morlet and A. Grossmann in the early 1980's (Goupillaud et al. 1984; Grossmann and Morlet 1984; Grossmann and Morlet 1985), the continuous wavelet transform (WT) has been the subject of considerable theoretical developments and practical applications in a wide variety of fields (Combes et al. 1989; Meyer 1992; Daubechies 1992; Ruskai et al. 1992; Meyer and Roques 1993; Farge et al. 1993; Arneodo et al. 1995a; Erlebacher et al. 1996; Mallat 1998; Torresani 1998; Silverman and Vassilicos 2000; Jaffard et al. 2001). Originally introduced to perform time-frequency analysis, the WT has been early recognized as a mathematical microscope that is well adapted to characterize the scale-invariance properties of fractal objects and to reveal the hierarchy that governs the spatial distribution of the singularities of multifractal measures and functions (Arneodo et al. 1988; Holschneider 1988; Arneodo et al. 1989; Jaffard 1989; Holschneider and Tchamitchian 1990; Jaffard 1991; Mallat and Hwang 1992; Mallat and Zhong 1992; Arneodo et al. 1992a). This has led A. Arneodo and collaborators (Muzy et al. 1991, 1993, 1994; Bacry et al. 1993; Arneodo et al. 1995c) to elaborate on a unified statistical (thermodynamic) description of multifractal distributions including measures and functions, the so-called Wavelet Transform Modulus Maxima (WTMM) method. This method relies on the computation of partition functions from the WT skeleton defined by the wavelet transform modulus maxima. This skeleton provides an adaptive space-scale partition of the fractal distribution under study, from which one can extract the $D(h)$ singularity spectrum of Hölder exponent values as the equivalent of a thermodynamic potential (entropy) (Arneodo et al. 1995c). We refer the reader to Bacry et al. (1993), Jaffard (1997a; 1997b) and collaborators (Jaffard et al. 2006) for rigorous mathematical results and to Hentschel (1994) for the theoretical treatment of random multifractal functions. Applications of the WTMM method to 1D signals (Arneodo et al. 2002a) and its generalization in 2D for image analysis (Arneodo et al. 2000; Decoster et al. 2000; Arneodo et al. 2003) and in 3D for scalar and vector fields analysis (Kestener and Arneodo 2003, 2004, 2007) have already provided insights into a wide variety of problems (Arneodo et al. 2002a), in domains as different as the fully-developed turbulence (Arneodo et al. 1998b, 1999b; Roux et al. 1999; Delour et al. 2001; Mordant et al. 2002, 2003), hydrology (Venugopal et al. 2006a,b; Roux et al. 2009), astrophysics (Khalil et al. 2006; Kestener et al. 2010; McAteer et al. 2010), geophysics (Arrault et al. 1997; Arneodo et al.

1999a; Roux et al. 2000), econophysics (Arneodo et al. 1998c; Muzy et al. 2001), fractal growth phenomena (Arneodo et al. 1992b,c,d; Kuhn et al. 1994), medical time series analysis (Ivanov et al. 1996, 1999) and medical and biological image processing (Arneodo et al. 2003; Kestener et al. 2001; Khalil et al. 2007; Caddle et al. 2007). Surprisingly, among these applications, there is one that turns out to be quite fruitful and very promising in regards to its unexpected appropriateness, namely the multi-scale wavelet-based analysis of genomic sequences (Arneodo et al. 1995b, 1996, 2002a; Audit et al. 2001, 2002).

Long-range correlations in genomic sequences.

The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences has been the subject of increasing interest (Stanley et al. 1993; Li et al. 1994; Arneodo et al. 2002a). During the past twenty years or so, there has been intense discussion about the existence, the nature and origin of the long-range correlations (LRC) observed in DNA sequences (Li 1992; Peng et al. 1992; Voss 1992; Nee 1992; Borštnik et al. 1993; Chatzidimitriou-Dreismann and Larhammar 1993; Karlin and Brendel 1993; Larhammar and Chatzidimitriou-Dreismann 1993; Peng et al. 1993; Stanley et al. 1993; Voss 1994; Azbel' 1995; Herzel and Große 1995; Mantegna et al. 1995; Arneodo et al. 1996; Li 1997; Viswanathan et al. 1998; Arneodo et al. 2002a). One of the main obstacles to LRC analysis in DNA sequences is the genuine mosaic structure of these sequences that are well known to be formed of patches of different underlying composition (Gardiner 1996; Li et al. 1998; Bernardi 2000). When using the DNA walk representation, these patches appear as trends in the DNA walk landscapes that are likely to break scale-invariance (Nee 1992; Peng et al. 1992; Borštnik et al. 1993; Chatzidimitriou-Dreismann and Larhammar 1993; Larhammar and Chatzidimitriou-Dreismann 1993; Karlin and Brendel 1993; Stanley et al. 1993; Viswanathan et al. 1998; Arneodo et al. 2002a). Indeed, most of the techniques, *e.g.* the variance method, used in the early studies for characterizing the presence of LRC, were not well-adapted to study non-stationary sequences. There have been some phenomenological attempts to differentiate local patchiness from LRC using *ad hoc* methods such as the so-called "min-max method" (Peng et al. 1992) and the "detrended fluctuations analysis" (Peng et al. 1994). In that context the WT has been early recognized as a well-suited technique that overcomes this difficulty (Arneodo et al. 1995b, 1996, 2002a). By considering analyzing wavelets that make the WT microscope blind to low-frequency trends, any bias in the DNA walk can be removed and the existence of power-law correlations with specific scale invariance properties can be revealed accurately. As a first important result, from a systematic WT analysis of human exons, CDSs and introns, LRC were found in non-coding sequences as well as in regions coding for proteins somehow hidden in their inner codon structure (Arneodo et al. 1998a). This observation made rather questionable the model based on genome plasticity proposed at that time to account for the reported absence of LRC in coding sequences (Li 1992; Peng et al. 1992; Stanley et al. 1993; Arneodo et al. 1995b; Buldyrev et al. 1995; Arneodo et al.

1996). An alternative structural interpretation of these LRC has emerged from a comparative multifractal analysis of DNA sequences using structural coding tables based on nucleosome positioning data (Audit et al. 2001, 2002). The application of the WTMM method has revealed that the corresponding DNA chain bending profiles are monofractal (homogeneous) as characterized by a unique Hölder exponent $h = H$ and that there exists two LRC regimes. In the 10 bp - 200 bp range, LRC are observed for eukaryotic sequences as quantified by a Hurst exponent value $H \simeq 0.6$ as the signature of the nucleosomal structure. In contrast, for eubacterial sequences, the uncorrelated $H = 0.5$ value is systematically obtained. These LRC were further shown to favor the autonomous formation of small (a few hundred bps) 2D DNA loops and in turn the propensity of eukaryotic DNA to interact with histones to form nucleosomes (Vaillant et al. 2005, 2006). In addition these LRC might induce some local hyper-diffusion of these loops which would be a very attractive interpretation of the nucleosome repositioning dynamics. Over larger distances ($\geq$ 200 bp), stronger LRC with $H \simeq 0.8$ seem to exist in any sequence (Audit et al. 2001, 2002) as experimentally confirmed by atomic force microscopy imaging of naked DNA molecules deposited onto a mica surface under 2D thermodynamic equilibrium conditions (Moukhtar et al. 2007, 2010). Furthermore these LRC were recently observed in *S.cerevisiae* nucleosome positioning *in vivo* data suggesting that they are involved in the collective nucleosome organization of the so-called 30 nm chromatin fiber (Vaillant et al. 2007; Arneodo et al. 2008). The fact that this second regime of LRC is also present in eubacterial sequences shows that it is likely to be a possible key to the understanding of the structure and dynamics of both eukaryotic and prokaryotic chromatin fibers.

Bifractality of the compositional asymmetry.

The increasing availability of new fully sequenced genomes has provided an unprecedented opportunity to generalize the application of the WTMM method to genome-wide multifractal sequence analysis when using alternative codings that have a clear functional meaning. Among these codings, the TA and GC skews permit to study the strand asymmetry generated by the transcription and replication processes (Chapter I). Genome-wide multi-scale analysis of mammalian genomes has clearly shown compositional asymmetry in intergenic regions and further confirmed the existence of replication-associated strand asymmetries (Nicolay et al. 2004; Brodie of Brodie et al. 2005; Touchon et al. 2005; Chen et al. 2011). In particular the application of the WTMM method to the skew $S = S_{TA} + S_{GC}$ in the human genome (Nicolay et al. 2007; Arneodo et al. 2009) has revealed the bifractal nature of the corresponding DNA walk landscape which involves two competing scale invariant (from repeat masked distances of 1 kbp up to 40 kbp) components characterized by Hölder exponent $h_1 = 0.78$ and $h_2 = 1$ respectively. The former corresponds to the long-range correlated homogeneous fluctuations previously observed in DNA walks generated with structural codings (Audit et al. 2001, 2002). The latter is associated with **the presence of jumps in the skew profile**. A majority of the detected

upward (resp. downward) jumps were shown to co-locate with gene transcription start sites (TSS) (resp. transcription termination sites (TTS)). However, about a third of the detected upward jumps are still observed at scale $\geq$ 200 kbp, larger than the typical gene size, as bordering large-scale (mean size $\simeq$ 1 Mbp) **N-shaped skew domains**. The N-domains borders were hypothesized to be replication origins active in the germline (Brodie of Brodie et al. 2005; Touchon et al. 2005; Huvet et al. 2007), possibly specified by an open chromatin structure favorable to early replication initiation and permissive to transcription (Audit et al. 2009). A preliminary analysis of human replication timing data confirmed that a significant number of N-domain borders are initiation zones that replicate earlier in the S-phase than their surrounding regions, whereas central regions replicate late (Audit et al. 2007).

☛ *Our aim in this Chapter is to use the WT transform to objectively delineate skew N-domains thereby predicting (at their edges) replications origins directly from the DNA sequence. In our current perspective, skew N-domains actually define replication domains in the germline, characterized by a global N-shape of the replication fork polarity. With an adequate choice of the analyzing wavelet, we show that the proposed method can be further used to disentangle, in the so-identified replication domains, the contribution coming from transcription (local square-like genic skew component) from the one associated with replication (global N-shaped skew component).*

IV.2 Review of transcription- and/or replication- coupled strand asymmetries in mammalian genomes

☛ *A square-like mean skew profile is observed in mammalian genes (see Fig. 1). In prokaryotes, where the replication program follows the replicon model, a square-like skew profile is also observed associated to replication. In particular, the upward and downward jumps of the skew profile remarkably colocalize with the replication origin and terminus. In contrast, we observe in mammalian genomes a recurrent N-shape pattern in the skew profile, likely associated to replication (see Fig. 3). We propose that the serrated "factory roof" skew profile observed in these N-domains results from the superimposition of the genic square-like components and the global N-shape component induced by replication (see Fig. 4).*

IV.2.1 Definitions of the compositional skews

We will mainly use in this study the TA and GC skews computed in non-overlapping 1 kbp windows (Eq. 158 of Chapter I) (Arneodo et al. 2007, 2009):

$$S_{TA} = \frac{[T]-[A]}{[T]+[A]}, \qquad S_{GC} = \frac{[G]-[C]}{[G]+[C]}, \tag{1}$$

where $[T], [A], [G]$ and $[C]$ are respectively the frequencies of T, A, G and C in the windows. Because of the observed correlation between TA and GC skews (Touchon et al. 2003) we will mainly consider the total skew:

$$S = S_{TA} + S_{GC}. \tag{2}$$

Sequences and gene annotation data ("knownGene") were retrieved from the UCSC Genome Browser (May 2004). We used RepeatMasker to exclude repetitive elements (SINEs, LINEs, ...) that might have been inserted recently in the genome and would not reflect long-term evolutionary patterns.

IV.2.2 Transcription-induced square-like skew profiles in mammalian genomes

Substitutional and compositional asymmetries associated to transcription have been observed in organisms across the whole life tree (Section I.1.3). In a previous study, Touchon et al. (2003; 2004) reported definite evidence for compositional asymmetry in transcribed regions of human sequences. The distributions of the TA and GC skews, computed on the 14 854 intron-containing genes, present positive mean values for "sense" (+) genes (7058), namely $\bar{S}_{TA} = 4.49 \pm 0.01\%$ and $\bar{S}_{GC} = 3.29 \pm 0.01\%$, and nearly opposed values for "antisense" (−) genes (7346). In Fig. 1 are reported the mean values of S_{TA} and S_{GC} for all genes, computed on the coding strand, as a function of the distance to the 5' or 3' gene ends. At the 5' gene extremity (Fig. 1(a)), a sharp transition of both skews is observed from close to zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for S_{TA} and between 3 and 5% for S_{GC}. At the 3' gene extremity (Fig. 1(b)), the TA and GC skews also exhibit a transition from significantly large positive values inside the gene to very small values in untranscribed regions. However, in comparison to the steep transition observed at 5' end, the 3' end mean profile presents a slightly smoother transition pattern extending over ~ 5 kbp and including regions downstream of the 3' end likely reflecting the fact that transcription continues to some extent downstream of the polyadenylation site. In pluricellular organisms, mutations responsible for the observed biases have occurred in germline cells. It could happen that gene 3' ends annotated in the databank differ from the poly-A sites effectively used in germline cells. Such differences would then lead to some broadening of the skew profiles. Overall, the results reported in Fig. 1 suggest that the S_{TA} and S_{GC} are constant along introns. Since introns amount for about 80% of gene sequences, this means that **skew profiles induced by transcription processes have a characteristic square-like shape** (Touchon et al. 2003, 2004; Arneodo et al. 2007, 2009). However, the absence of asymmetries in intergenic regions does not exclude the possibility of additional replication associated biases. Such biases would present opposite signs in regions of opposite replication fork polarity that would cancel each other in our statistical analysis.

If there is not doubt that the mean TA and GC skew profiles are different from

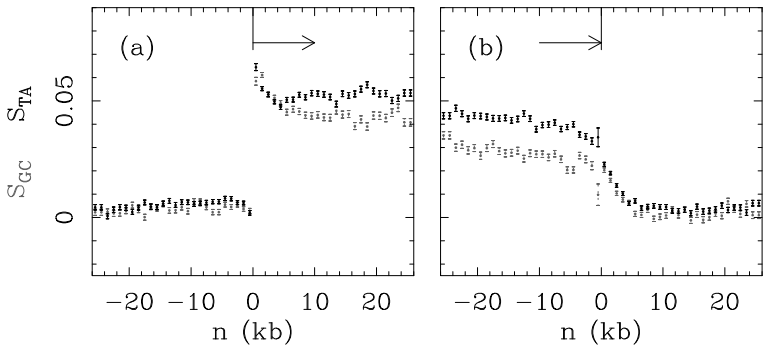

Figure 1: **Transcription square-like skew component in human genes.** TA (•) and GC (•) skew profiles in the regions surrounding 5' and 3' gene extremities (Touchon et al. 2003). S_{TA} and S_{GC} were calculated in 1 kbp windows, on the coding strand, starting from each gene extremities in both directions. Only intronic and repeat-masked nucleotides were retained. In abscissa is reported the distance (n) of each 1 kbp window to the indicated gene extremity; zero values of abscissa correspond to 5' (a) or 3' (b) gene extremities. In ordinate is reported the mean value of the skews over 14 854 intron-containing genes for all 1 kbp windows at the corresponding abscissa. Error bars represent the standard error of the means.

zero inside the genes likely resulting from transcription-coupled processes, how many genes actually contribute to these mean profiles or in other words, how many genes have biased sequences? Since each square-like skew pattern is edged by one upward and one downward jump, the set of human genes that are significantly biased is expected to contribute to an equal number of $\Delta S > 0$ and $\Delta S < 0$ jumps. This is exactly what we observed when using the WT microscope to detect jumps in the noisy total skew profile S when exploring the range of scales $10 \leq a \leq 40$ kbp, typical of human gene size (Nicolay et al. 2007). Out of 20 023 TSS, 36% (7228) were found to be delineated within 2 kbp by an upward jump of amplitude $\Delta S > 0.1$. This percentage of biased genes provides a very reasonable estimate of the number of genes expressed in germ-line cells as compared to the 31.9% recently experimentally found to be bound to PolII in human embryonic stem cells (Lee et al. 2006).

Remark. This study of strand asymmetries in intronic sequences has been further extended to evolutionary distant eukaryotes (Touchon et al. 2004). When appropriately examined, all genomes present transcription-coupled excess of T over A ($S_{TA} > 0$) in the coding strand. In contrast, GC skew is found positive in mammals and plants but negative in invertebrates suggesting different repair mechanisms associated with transcription in vertebrates and invertebrates (Touchon et al. 2004; Touchon 2005).

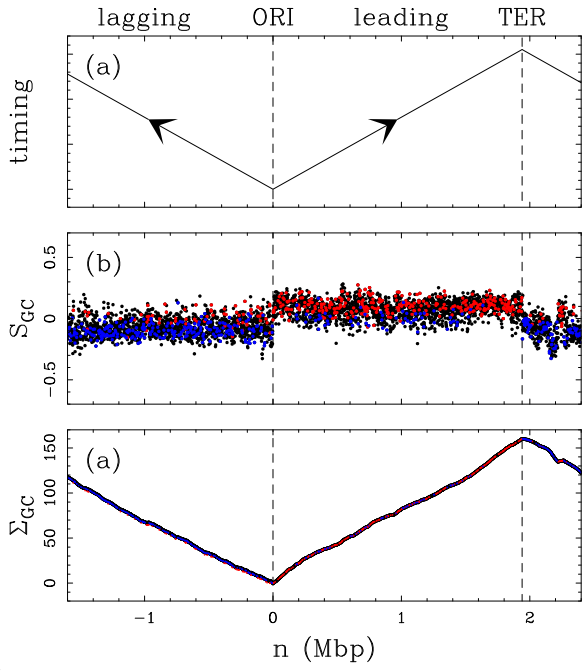

Figure 2: **Replication square-like skew component in Bacillus subtilis.** (a) Schematic representation of the divergent bidirectional progression of the two replication forks from the replication origin. The leading (resp. lagging) strand is replicated by a sense (+) (resp. antisense (−)) fork and has replication fork polarity $p = +1$ (resp. $p = -1$). (b) S_{GC} calculated in 1 kbp windows along the genomic sequence of *Bacillus subtilis*. (c) Cumulated skew Σ_{GC}. The vertical lines correspond respectively to the replication origin (ORI) and termination (TER) positions. In (b) and (c), red (resp. blue) points correspond to (+) (resp. (−)) genes.

IV.2.3 Replication-induced N-shaped skew profiles in mammalian genomes

Substitutional and compositional asymmetry associated to replication has been observed in organisms across the whole life tree (Section I.1.2). The existence of replication-associated strand asymmetries has been mainly established in bacterial genomes (Lobry 1995; Mrázek and Karlin 1998; Frank and Lobry 1999; Rocha et al. 1999; Tillier and Collins 2000). As illustrated in Fig. 2, the GC and TA skews abruptly switch sign (over a few kbp) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This step-like profile is characteristic of the replicon model (Jacob et al. 1963). In *Bacillus subtilis*, as in most bacteria, the leading (resp. lagging) strand (Fig. 2(a)) is generally richer (resp. poorer) in G than in C (Fig. 2(b)), and to a lesser extent in T than in A (data not shown). This typical pattern is particularly clear when plotting the cumulated skews Σ_{GC} (Fig. 2(c)) and Σ_{TA}; both

present decreasing (or increasing) profiles in regions situated 5' (or 3') to the origin, displaying a characteristic $\vee$-shape pointing to the replication origin position (similarly a characteristic $\wedge$-shape is observed at the terminus position). The research of $\vee$ patterns in the cumulated skews has been extensively used as a strategy to detect the position of the (unique) replication origin in (generally circular) bacterial genomes (Mrázek and Karlin 1998; Frank and Lobry 1999; Rocha et al. 1999; Tillier and Collins 2000). We note that the replication-associated compositional asymmetry observed in bacteria is a particular case of the model proposed in Chapter I. In the replicon model, the circular bacterial chromosome is divided into two halves (the leading and lagging strands) of opposed replication fork polarity. The replication fork polarity has therefore the same square-like pattern as the S_{GC} skew. Furthermore, the cumulative skew Σ_{GC} profile is expected to give qualitatively the replication timing pattern, as the replication fork polarity is related to the derivative of the replication timing. Interestingly, the Σ_{GC} profile (Fig. 2(c)) indeed mimics the replication timing pattern expected in the replicon model (Fig. 2(a)) (see also Fig. 1 in Chapter II).

In previous works, Brodie of Brodie et al (2005) and Touchon et al (2005) have investigated the behavior of the skew profiles around 9 replication origins experimentally identified in the human genome. As shown in Fig. 3(a) for TOP1 replication origin, most of these origins also correspond to rather sharp (over several kbp) transitions from negative to positive S (S_{TA} as well as S_{GC}) skew values that clearly emerge from the noisy background. As shown in Fig. 3(b-d), sharp upward jumps of amplitude $\Delta S \geq 15\%$, similar to the ones observed for the known replication origins (Fig. 3(a)), seem to exist at many other locations along the human chromosomes. This observation led A. Arneodo and collaborators to develop an upward jump detection methodology based on the WT microscope (Brodie of Brodie et al. 2005; Touchon et al. 2005). By selecting in the WT skeleton, the maxima lines that still exist at scales $a \geq 200$ kbp, *i.e.* scales larger than the typical gene size (~ 30 kbp), one not only reduces the effect of the noise but also the contribution of the upward (5' extremity) and downward (3' extremity) jumps associated to the square-like skew pattern induced by transcription (Fig. 1). When applying this wavelet-based method to the 22 human autosomes, retaining as putative replication origins upward jumps with $\Delta S \geq 12\%$, *i.e.* with an amplitude much larger than the one induced by transcription at the TSS (Fig. 1(a)), A. Arneodo and collaborators got a set of 1012 candidates mainly located in regions with G+C $\leq 42\%$ (as seen in Fig. 3(d), in G+C rich regions the required scale separation between the characteristic replicon and gene sizes was no longer tractable to the multi-scale wavelet-based methodology) (Brodie of Brodie et al. 2005; Touchon et al. 2005).

But when examining the behavior of the skews at large distance from the replication origins, one does not observe a square-like pattern with upward and downward jumps at the origin and termination positions as expected from the replicon model.

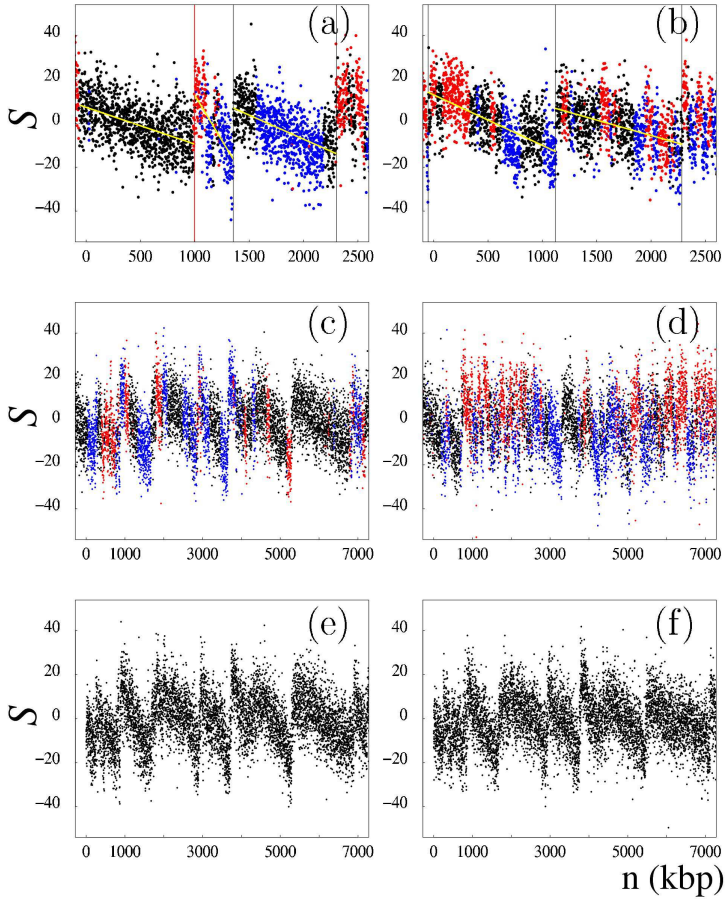

Figure 3: **Replication N-shaped skew component in mammalian genomes.** S profiles along mammalian genome fragments (without repeats) (Touchon et al. 2005; Arneodo et al. 2007). (a) Fragment of human chromosome 20 including the TOP1 origin (red vertical line). (b and c) Human chromosome 4 and chromosome 9 fragments, respectively, with low GC content (36%). (d) Human chromosome 22 fragment with larger GC content (48%). In (a) and (b), vertical lines correspond to selected putative origins; yellow lines are linear fits of the S values between successive putative origins. Black, intergenic regions; red, (+) sense genes; blue, (−) anti-sense genes. Note the fully intergenic regions upstream of TOP1 in (a) and from positions 5290-6850∼kbp in (c). (e) Fragment of mouse chromosome 4 homologous to the human fragment shown in (c). (f) Fragment of dog chromosome 5 homologous to the human fragment shown in (c). In (e) and (f), genes are not represented.

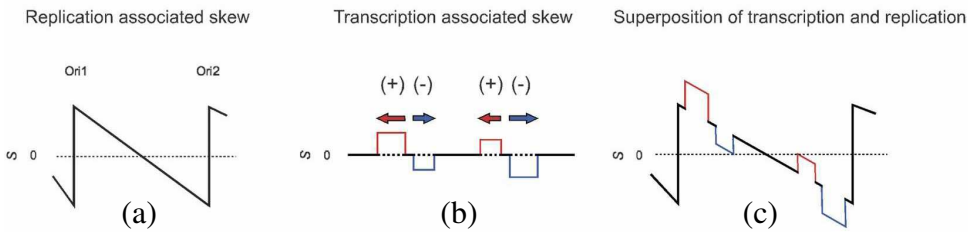

Figure 4: **Factory roof model of mammalian skew profiles.** (a) N-shaped replication-associated skew profile. (b) Transcription-associated skew profile showing positive square blocks at (+) gene positions and negative square blocks at (−) gene positions. (c) Superimposition of the replication- and transcription- associated skew profiles producing the final factory-roof pattern that defines "N-domains" (Huvet et al. 2007).

Indeed the most striking feature is the fact that in between two neighboring major upward jumps, not only the noisy S profile does not present any comparable downward sharp transition, but it displays a remarkable decreasing linear behavior. At chromosome scale, one thus gets jagged S profiles that have the aspect of "factory roofs" (Touchon et al. 2005; Brodie of Brodie et al. 2005; Arneodo et al. 2007; Huvet et al. 2007). Note that the jagged S profiles shown in Fig. 3(a-d) look somehow disordered because of the extreme variability in the distance between two successive upward jumps, from spacing ∼50-100 kbp (∼100-200 kbp for the native sequences) mainly in GC rich regions (Fig. 3(d)), up to 1-2 Mbp (∼2-3 Mbp for native sequences) (Fig. 3(c)). But what is important to notice is that some of these segments between two successive skew upward jumps are entirely intergenic (Fig. 3(a,c)), clearly suggesting that the observed peculiar **N-shape skew profile is characteristic of a strand bias resulting solely from replication** (Brodie of Brodie et al. 2005; Touchon et al. 2005; Huvet et al. 2007). Importantly, as illustrated in Fig. 3(e,f), the factory-roof pattern is not specific to human sequences but is also conserved in homologous regions of the mouse and dog genomes (Touchon et al. 2005). Hence, the presence of strand asymmetry in regions that have strongly diverged during evolution further supports the existence of compositional bias associated with replication in mammalian germline cells (Touchon et al. 2005; Brodie of Brodie et al. 2005; Arneodo et al. 2007; Huvet et al. 2007; Chen et al. 2011).

IV.2.4 A working model of mammalian "factory roof" skew profiles

As discussed in Chapter I (Eq. (168) in the summary), the compositional asymmetry can be decomposed into transcription- and replication-associated components. The transcription-associated asymmetry changes sign with gene orientation and increases in magnitude with the transcription rate (in the germline). The replication-associated is proportional to the replication fork polarity (in the germline). Under the assumption of constant replication fork velocity, the replication fork polarity

is also proportional to the derivative of the mean replication timing (Chapter II, Eq. (68) in the summary). According to the results reported just above, we will use as a working model that the overall factory roof profile observed for mammalian genomes actually results from the superposition of two patterns (Fig. 4) (Huvet et al. 2007). One decreases steadily from the 5' to the 3' direction and would be attributable to replication in germline cells (Fig. 4(a)), and would reflect the progressive inversion of the replication fork polarity. In turn, the linear decrease of the replication fork polarity would indicate that the germline replication timing has a parabolic U-shape (the derivative of a parabolic U-shape is a linear N-shape). The global N-shaped skew component, attributed to replication, thus defines **skew N-domains as replication domains characterized by a U-shaped replication timing profile in the germline**. The other pattern would result from transcription-associated strand asymmetries that generate square-like profiles corresponding to (+) sense and (−) antisense genes (Fig. 4(b)). When the two profiles are superimposed, this leads to the factory roof pattern (Fig. 4(c)) (Huvet et al. 2007). Because the typical gene size ($\sim$ 30 kbp) is much smaller than the characteristic distance ($\sim$ 1 Mbp) between two adjacent putative replication origins, these skew domains were named "N-domains" in regards of their overall qualitative N shape (Huvet et al. 2007).

IV.3 Detecting replication N-domains with the continuous wavelet transform

☛ *The wavelet transform, using an adapted N-shaped wavelet, provides a robust way to perform a multi-scale pattern recognition of the letter N in a noisy skew profile. In this section, we summarize the wavelet-based methodology previously developed by Nicolay in (Nicolay 2006; Baker et al. 2010) to detect skew N-domains. The efficiency of this methodology is verified on synthetic skew signals.*

IV.3.1 The continuous N-let transform

The continuous wavelet transform (WT) is a space-scale analysis which consists in expanding a signal S in terms of wavelets that are constructed from a single function, the analyzing wavelet Ψ, by means of dilations and translations (Goupillaud et al. 1984; Grossmann and Morlet 1984; Grossmann and Morlet 1985; Combes et al. 1989; Meyer 1992; Daubechies 1992):

$$W_\Psi[S](b,a) = \frac{1}{a}\int_{-\infty}^{+\infty} S(y)\Psi\left(\frac{y-b}{a}\right)dy, \tag{3}$$

where b and a (>0) are the space and scale parameters respectively. The analysing wavelet Ψ is generally chosen to be well localized in both space and frequency. Usually Ψ is required to be of zero mean for the WT to be invertible. For the particular purpose of segmenting the human genome, and more generally mammalian

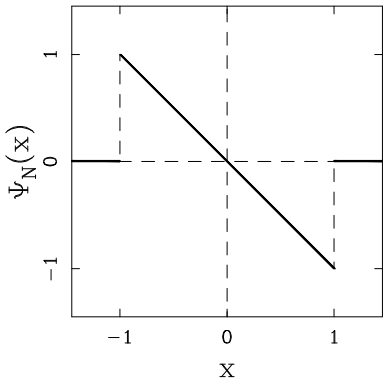

Figure 5: **N-shaped analyzing wavelet.** $\Psi_N(x)$ defined in Eq. (4).

genomes, according to the working model described in Fig. 4, we will use in the following an adapted analyzing wavelet called N-let because of its shape that looks like the letter N (Fig. 5) (Audit et al. 2007; Huvet et al. 2007; Arneodo et al. 2009):

$$\Psi_N(x) = -x\chi_{[-1,1]}(x), \tag{4}$$

where $\chi_{[-1,1]}(x)$ is the characteristic function of the interval $[-1, 1]$.

IV.3.2 Multi-scale pattern recognition with the continuous N-let transform

Along the line of the working model of mammalian "factory roof" skew profiles defined in Fig. 4, let us compute the N-let transform of the following theoretical skew profile (Nicolay 2006):

$$S(x) = (-\theta(x - r^*) + h)\chi_{[r_1,r_2]}(x), \tag{5}$$

where r_1 and r_2 are the bordering replication origin positions and $r^* = (r_1 + r_2)/2$. The N-let transform of Eq. (5) takes the following analytical expressions:

$$W_{\Psi_N}[S](b,a) = \begin{cases} 0 \\ \frac{\theta(b-r^*)-h}{2}\left(1 - \frac{(r_1-b)^2}{a^2}\right) + \frac{\theta a}{3}\left(1 - \frac{(r_1-b)^3}{a^3}\right) \\ \frac{2\theta a}{3} \\ \frac{\theta(b-r^*)-h}{2}\left(\frac{(r_2-b)^2}{a^2} - 1\right) + \frac{\theta a}{3}\left(\frac{(r_2-b)^3}{a^3} + 1\right) \\ \frac{\theta(b-r^*)-h}{2a^2}\left((r_2-b)^2 - (r_1-b)^2\right) + \frac{\theta}{3a^2}\left((r_2-b)^3 - (r_1-b)^3\right) \end{cases}$$

$$\text{for} \begin{cases} b \leq r_1 - a \text{ or } b \geq r_2 + a \\ r_1 - a < b \leq r_1 + a \text{ and } b < r_2 - a \\ b \geq r_1 + a \text{ and } b \leq r_2 - a \\ b > r_1 + a \text{ and } r_2 - a < b < r_2 + a \\ b < r_1 + a \text{ and } b > r_2 - a \end{cases} \tag{6}$$

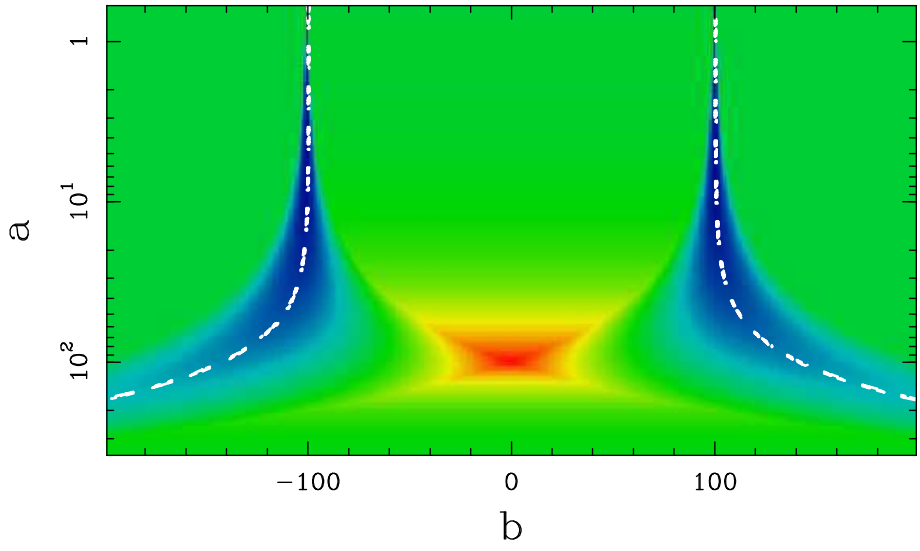

Figure 6: **Space-scale representation of the N-let transform.** N-let transform of the skew function $S(x) = -10^{-2}x\chi_{[-100,100]}(x)$ (see Eq. (5)). $W_{\Psi_N}[S](b,a)$ is maximum at $(b^*,a^*) = (0,100)$. $W_{\Psi_N}[S](b,a)$ is color coded from blue (minimum) to red (maximum) through green (intermediate values). The white dashed-lines correspond to the WTMM lines (local negative minimum of $W_{\Psi_N}[S](b,a)$) $\mathcal{L}_L$ and $\mathcal{L}_R$ that point to the extremities of the theoretical N-domains in the limit $a \to 0^+$.

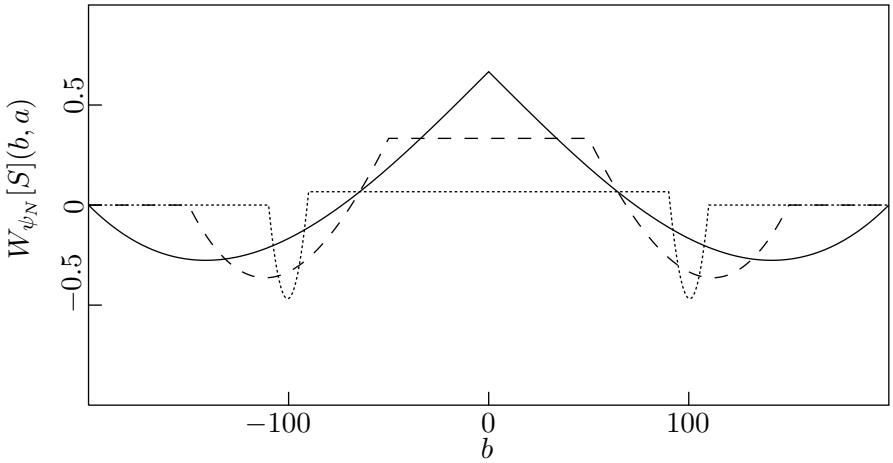

Figure 7: **1D cut of the N-let transform.** Horizontal 1D cuts of the N-let transform represented in Fig. 6 for the scales $a = 10$ (dotted line), 50 (dashed line) and 100 (solid line).

As an illustration, a space-scale representation of the N-let transform of $S(x)$ for the following parameters values $\theta = 10^{-2}$, $h = 0$, $r_1 = -100$ and $r_2 = 100$, is shown in Fig. 6. Some 1D cuts corresponding to different scale values $a \leq a^*$ are shown in Fig. 7. For a given scale a, $W_{\Psi_N}[S](b, a)$ exhibits a plateau for $b \in [b_1^*(a), b_2^*(a)]$, where $b_1^*(a) = r_1 + a$ and $b_2^*(a) = r_2 - a$. When increasing a, this plateau increases linearly with a while the interval $[b_1^*(a), b_2^*(a)]$ shrinks to zero so that $W_{\Psi_N}[S](b, a)$ presents a local maximum in the (b, a) half-plane at the point (b^*, a^*) :

$$W_{\Psi_N}[S](b^*, a^*) = \frac{2}{3}\theta a^*, \tag{7}$$

where

$$b^* = r^* = \frac{r_1 + r_2}{2}, \qquad a^* = \frac{r_2 - r_1}{2}. \tag{8}$$

The determination of this N-let transform local maximum (the red spot in Fig. 6) therefore provides an estimate of the mid-point and the size of the support of the function S via Eq. (8) and of its linear slope $-\theta$ via Eq. (7). Note that these results extend to non-zero values of the offset parameter h in Eq. (5) provided it remains small as compared to the jump amplitude θa^*.

IV.3.3 Numerical method

The method we propose to identify replication N-domains in noisy factory roof like profiles (Fig. 3) involves several steps.

* Step 1: Detecting potential N-domain borders.

  We smooth the skew S with a square-like filter of size 20 kbp:

$$\tilde{S}(x) = \frac{1}{20} S * \chi_{[-10,10]}(x). \tag{9}$$

  The amplitude $\Delta S(x)$ at a point x is defined as:

$$\Delta S(x) = \tilde{S}(x + 20) - \tilde{S}(x - 20). \tag{10}$$

  On purpose, we do not take into account the nucleotides that are closer than 10 kbp to the jump location for which a high variability of skew values is observed. Then, along the line of previous investigation of replication origins in mammalian genomes (Audit et al. 2007; Huvet et al. 2007; Arneodo et al. 2009), the position x_n of an upward jump will be considered as a good candidate for the location of a putative replication origin if it is a local maximum of $\Delta S(x)$ and satisfies the condition:

$$\tilde{S}(x_n - 20) \leq -\epsilon \quad \text{and} \quad \tilde{S}(x_n + 20) \geq \epsilon. \tag{11}$$

  This condition not only fixes a threshold ($= 2\epsilon$) in the jump amplitude but it also imposes the fact that a putative replication origin must correspond to a

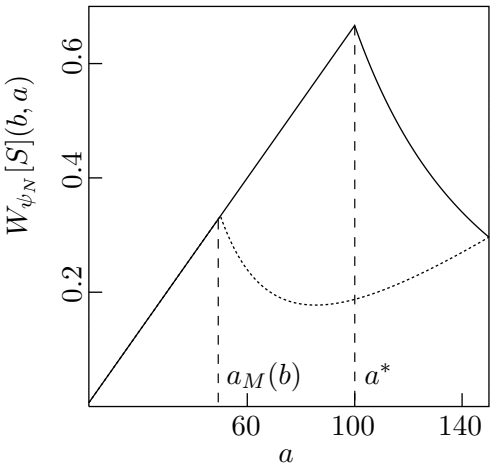

Figure 8: **1D cut of the N-let transform.** Vertical 1D cuts of the N-let transform represented in Fig.6 for $b = 50$ (dotted line) and $b = b^* = 0$ (solid line). The largest value of $a_M(b)$ is obtained for $b = b^*$, $a = a^*$ (Eq. (8)), *i.e.* when the analyzing N-let is positioned at the center of the support of S.

jump from negative to positive skew values. In the present work, according to the histograms of S values obtained from the human and mouse genomes, we will fix the threshold parameter ϵ to:

$$\epsilon = 3\,10^{-2}. \tag{12}$$

In this way, for each human and mouse chromosomes, we obtain a dictionary of upward jump locations as potential candidates for N-domain borders.

∗ Step 2: Associating pairs of borders to define N-domains.

For each pair of selected upward jumps (x_1, x_2), we determine the position (b_M, a_M) of the local maximum of the N-let transform (*i.e.* red spot in Fig. 6). For each $b \in [x_1, x_2]$, we estimate the range of scales $[a_1(b), a_2(b)]$ over which $W_{\Psi_N}[S](b, a)$ behaves linearly with the scale a as predicted by Eq. (6). We impose $a_1(b)$ to be larger than 40 (kbp) to minimize bias induced by the noise and the N-let sampling and, using a classical least square fit procedure, we estimate $a_2(b)$. As illustrated in the Fig. 8, for the theoretical example defined in Eq. (5), in the absence of noise, $a_2(b)$ corresponds to the scale $a_M(b)$ for which $W_{\Psi_N}[S](b, a)$ starts decreasing. For the genomic noisy skew profile, the sharp maximum of $W_{\Psi_N}[S](b, a)$ *vs* a turns out to be smoothed out so that $a_2(b) \leq a_M(b)$. We estimate $a_M(b)$ as the first N-let transform maximum above $a_2(b)$, provided $a_M(b) - a_2(b) \leq a_2(b)/10$ (if not we set $a_M(b) = a_2(b)$). Then, b_M is the position corresponding to the maximal value of $a_M(b)$ for $b \in [x_1, x_2]$. Finally, we check the consistency of associating (x_1, x_2) into a N-domain by comparing (b_M, a_M) with (b^*, a^*) (Eq. (8)). The interval $[x_1, x_2]$ is considered as a N-domain candidate if (b_M, a_M) corresponds to the expectation (b^*, a^*)

within 10% accuracy and allowing for a maximum of 30 kbp error on domain length. It is retained only if the χ^2 obtained by a linear regression fit of S over $[x_1, x_2]$ is smaller than a critical threshold. We select between overlapping intervals according to their χ^2. For example, if the intervals $I_1 = [x_1, x_2]$, $I_2 = [x_2, x_3]$ and $I_3 = [x_1, x_3]$ are all three good candidates, we will retain either I_1 and I_2 or I_3 according to whether $\chi^2_{I_1} + \chi^2_{I_2} < \chi^2_{I_3}$ or the opposite.

* Step 3: Refining N-domain border locations.

  The dictionary of upward jumps generated in Step 1 contains jumps identified at rather small scales ($\leq$ 20 kbp) as compared to the mean replication N-domain size we want to detect (see Section IV.4) and also to the characteristic size of mammalian genes that were shown to induce transcription-associated upward jumps in the skew profile at their promoter (Section IV.2.2). Consistently with the strategy of detecting putative replication origins pioneered in (Brodie of Brodie et al. 2005; Touchon et al. 2005), we take advantage of the space-scale representation provided by the WT to follow from large scales $a \leq a_M\left(\frac{x_1+x_2}{2}\right)$ to small scales, the blue tongues like the ones shown in Fig. 6 that are likely to point to the jump positions at the N-domain extremities. Practically, we identify in the WT skeleton the two nearest WTMM lines that exist at scale $a_M\left(\frac{x_1+x_2}{2}\right)$ immediately on the left $\mathcal{L}_L$ and the right $\mathcal{L}_R$ of the central point $\frac{(x_1+x_2)}{2}$, and that correspond to the two local negative minima of $W_{\Psi_N}[S](b, a)$ when represented versus b in Fig. 7. In regards to the noise amplitude and the mean gene size, we reallocate the N-domain extremities x_1 and x_2, to b_1 and b_2 respectively, where

$$b_1 \in \mathcal{L}_L(b, a) \quad \text{and} \quad b_2 \in \mathcal{L}_R(b, a) \quad \text{for } a = 40 \text{ kbp}. \tag{13}$$

  Then, we check that the new domain $[b_1, b_2]$ still satisfy the consistency condition of Step 2; if not we reject the interval $[b_1, b_2]$ as possible N-domain candidate.

IV.3.4 Test application on synthetic skew signals

To test our methodology, we generated synthetic factory roof like profiles of the following simple form (Nicolay 2006; Baker et al. 2010):

$$S(x) = \sum_j S_j(x) + g(x), \tag{14}$$

where $S_j(x)$ are functions similar to the one defined in Eq. (5):

$$S_j(x) = \left(-\theta_j \left(x - (r_j + \rho_j/2)\right) + h_j\right) \chi_{[r_j, r_j+\rho_j]}(x), \tag{15}$$

where $r_j = \sum_{i=1}^{j-1} \rho_i$ and $g(x)$ is a centered white noise. We chose the parameters to get a numerical S profile similar to the ones observed in the human and mouse

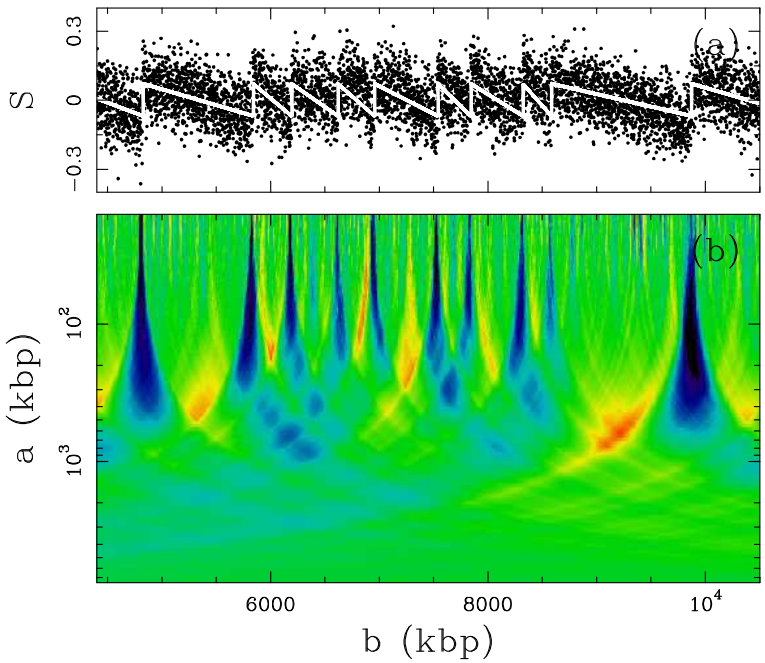

Figure 9: **N-let transform of synthetic skew profile.** (a) Synthetic skew profile (Eq. (14)) generated as explained in the text; the gray line represents the succession of deterministic N-shape functions $S_j(x)$. (b) Space-scale representation of the N-let transform of $S(x)$ using the same color coding as in Fig. 6 (Baker et al. 2010).

genomes (see Section IV.4). We fixed $\theta_j \rho_j = 0.14$, $h_j = 0$ an the noise standard deviation $\sigma_g = 0.08$, consistently with the corresponding genomic mean values (see Sections IV.4 and IV.5). We generated the length ρ_j of the synthetic N-domains according to a normal law of mean $\bar{\rho} = 550$ kbp and standard deviation σ_ρ =300 kbp consistently with the size statistics of the human and mouse masked N-domains (see Table 1). Square-like skew profiles of mean length 30 kbp and amplitude $\Delta S = 0.06$ (resp. $\Delta S = -0.06$) for sense (resp. anti-sense) genes were finally added to mimic the contribution to the skew associated to transcription (Section IV.2.2).

In Fig. 9 is shown the space-scale representation of a portion of the generated synthetic skew signal provided by the N-let transform. This representation illustrates the essence of our methodology, namely a multi-scale pattern recognition in the N-let transform half-plane. As reported in Fig. 10, out of the generated 2201 N-domains, 1997 were identified by our methodology, *i.e.* more than 90%. When examining the N-domains that were missed, they mainly correspond to small domains of length $\rho \leq 200$ kbp for which our method failed because of the lack of scale separation between the three components contributing to the total skew, namely the replication- and transcription-associated skews and the noise. When further analyzing the 1997

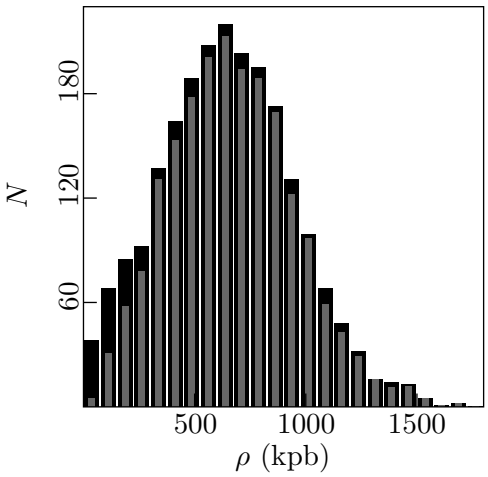

Figure 10: **Testing the performance of our multi-scale N-domain detection methodology.** Histogram of skew N-domain length generated as explained in the text and illustrated in Fig. 9(a): (black) theoretical histogram corresponding to a synthetic skew signal containing 2201 N-domains; (gray) histogram obtained from the 1997 N-domains detected by our N-let based methodology (Baker et al. 2010).

detected N-domains, we realized that the extremities of these domains were predicted with an accuracy of ~ 15 kbp which is reasonable in regards to the noise amplitude. As explored in the pioneering study of (Brodie of Brodie et al. 2005; Touchon et al. 2005), a better accuracy in upward jump detection can be obtained if one uses a smoother analyzing wavelet than the N-let like the first derivative of the Gaussian function. This leads us to modify Step 3 in our method.

- Step 3: In Step 3, the N-domain extremities b_1 and b_2 in Eq. (13) are now determined from the corresponding WTMM lines $\mathcal{L}_L$ and $\mathcal{L}_R$ in the WT skeleton computed with the first derivative of the Gaussian function $G^{(1)}(x) = -x \exp\left(-x^2/2\right)$.

When reproducing the test application with this new Step 3, similar efficiency was obtained but with a better accuracy in determining the N-domains edges, the mean error being reduced to 5 kbp (Nicolay 2006; Baker et al. 2010).

IV.4 Identifying replication N-domains in the human and mouse genomes

☛ *Using our wavelet-based methodology we detected in the human and mouse genomes respectively 678 and 587 skew N-domains. These skew N-domains correspond to large-scale genomic structures (from 200 kbp up to 3 Mbp) and recover a significant proportion of the human and mouse genomes (*33.8% *and* 22.3% *respectively).*

	N	L (Mbp)	C_N (%)	$\bar{n}_{\text{genes}}$	C_{inter} (%)	GC (%)
Human	678	0.63 ± 0.33 (masked) 1.19 ± 0.62 (native)	33.8	4.93	57.3 (masked)	40.8
Mouse	587	0.54 ± 0.34 (masked) 0.91 ± 0.62 (native)	22.3	4.13	58.2 (masked)	42.4

Table 1: **Replication N-domains detected in the human and mouse genomes**. For the human and mouse genomes we indicate the number N of detected skew N-domains, their characteristic length L (mean length ± standard deviation), the genome coverage C_N by N-domains, the mean number $\bar{n}_{\text{genes}}$ of genes found per N-domain, the intergenic coverage C_{inter} of N-domains, and the mean GC content.

IV.4.1 Human autosomes

When applying the wavelet-based method described in Section IV.3 to the skew profiles along the 22 human autosomes, we detected 759 N-domain candidates (Fig. 11). Among these domains, we discarded 17 domains that contain stretches of N (unknown nucleotides) longer than 100 kbp. We also removed from the remaining N-domains those (64) of length $L > 2.8$ Mbp whose shape is reminiscent of an N but split in half, leaving in the center a region of null skew whose length increases with domain size. As recently discovered (Zaghloul 2009), these split-N-domains have a central region corresponding to large heterochromatic gene deserts of homogeneous composition, *i.e.* both a null skew and a constant and low GC content. We ended with a library of 678 N-domains bordered by 1062 putative replication origins spanning 33.8% of the genome (Table 1). As shown in Fig. 12, the size of these N-domains ranges from $\sim$ 200 kbp (resp. 100 kbp when masked) to 2.8 Mbp (resp. 1.6 Mbp when masked) with a mean length $\bar{L} = 1.19$ Mbp (resp 0.63 Mbp when masked). Most of the 1062 putative replication origins at the extremities of the detected replication domains are intergenic (78%) and are located near to a gene promoter more often than would be expected by chance (data not shown). These N-domains contain approximately equal numbers of genes oriented in each direction (1653 (+) genes and 1690 (−) genes) with a mean gene number per domain of 4.93. As observed in (Huvet et al. 2007), gene distributions in the 5' half of N-domains contain more (+) than (−) genes, and vice-versa for the 3' half of N-domains. Note that these N-domains have a high intergenic coverage where the skew S is likely to result from replication only. As reported in Table 1 and Fig. 13, most of the detected N-domains are mainly intergenic with a mean (masked) coverage of 57.3%. Indeed only a few N-domains (64/678) have a (masked) intergenic coverage less than 20% (Fig. 13(b)).

IV.4.2 Mouse autosomes

When reproducing the same wavelet-based analysis for the 19 mouse autosomes, 634 N-domains candidates were identified with no domain containing large stretches of

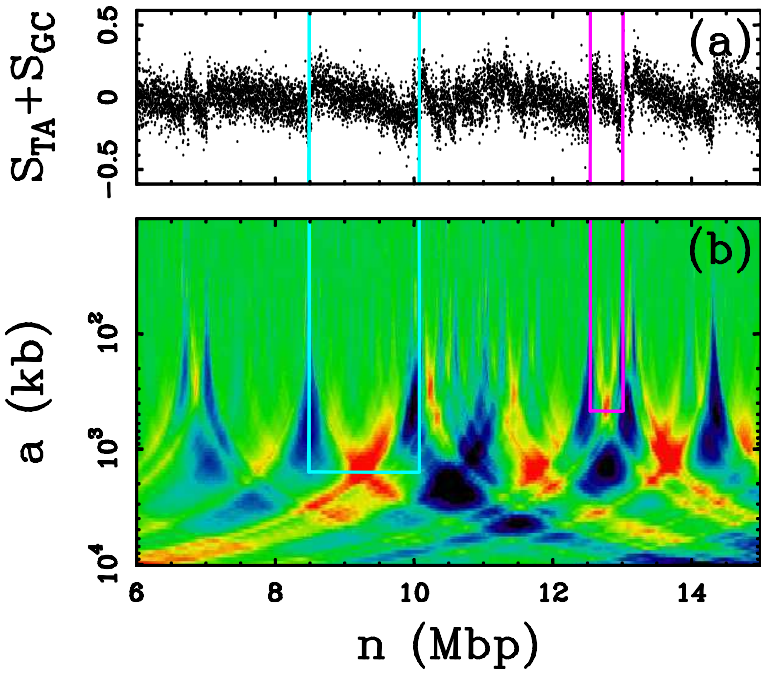

Figure 11: **Genome wide multiscale detection of the skew N-domains using the N-let transform**. (a) Skew profile S of a 9 Mbp repeat-masked fragment of human chromosome 21. (b) N-let transform of S using Ψ_N (Fig. 5); $W_{\Psi_N}[S](n,a)$ is color-coded from dark-blue (min; negative values) to red (max; positive values) through green (null values). Light blue and purple lines illustrate the detection of two replication domains of significantly different sizes. Note than in (b), blue cone-shape areas signing upward jumps point at small scale (top) towards the putative replication origins and that the vertical positions of the WT maxima (red areas) corresponding to the two indicated replications domains match the distance between the putative replication origins (1.6 Mbp and 470 kbp respectively) (Baker et al. 2010).

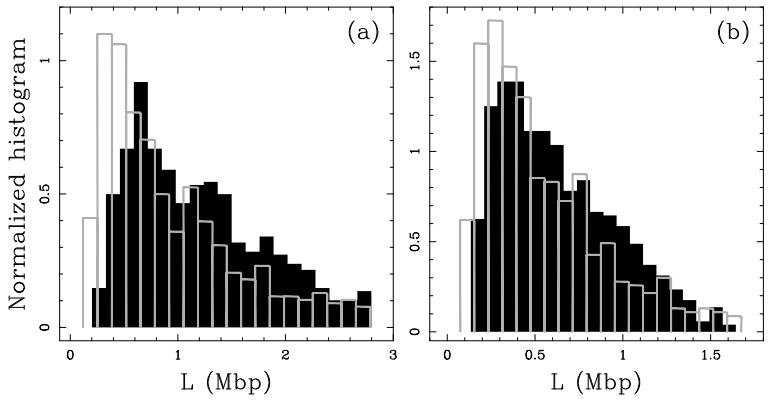

Figure 12: **Statistics of N-domain length.** Normalized histograms of N-domain length detected in the human (black) and mouse (gray) genomes: (a) native sequences; (b) masked sequences (Baker et al. 2010).

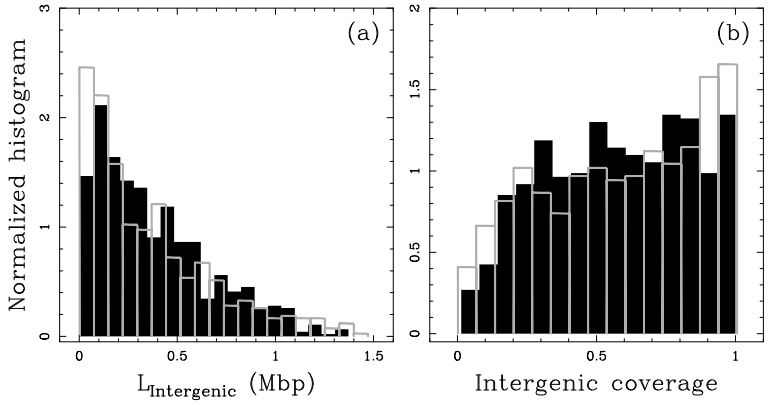

Figure 13: **Statistics of intergenic regions in N-domains.** Normalized histograms of intergenic regions in N-domains detected in human (black) and mouse (gray) genomes: (a) masked intergenic length; (b) coverage by masked intergenic regions per N-domain (Baker et al. 2010).

unsequenced nucleotides (N). After discarding 47 detected N-domains of length $L >$ 2.8 Mbp, we ended with a library of 587 N-domains that cover 22.3% of the genome, *i.e.* a percentage slightly smaller than previously obtained for the human genome. This results from the fact that more small domains are detected in the mouse genome (Fig. 12) with a mean native (resp. masked) length of 0.91 Mbp (resp. 0.54 Mbp). The mean GC content of the mouse N-domains (42.4%) is also slightly higher than in the human N-domains (40.8%). Despite these slight quantitative differences, most of the features concerning gene organization in human N-domains are also observed in mouse N-domains with a mean number of genes of 4.13 and globally a similar number of genes oriented in each direction (1220 (+) genes and 1204 (−) genes). Like in the human autosomes a majority of the 994 putative replication origins that border the 587 mouse N-domains are intergenic (71%) . Importantly the relative coverage of these N-domains by intergenic regions is important (58.2%) and statistically very similar to what is observed with the human autosomes (Fig. 13).

IV.5 Disentangling transcription- and replication-associated strand asymmetries

☛ *Following the factory roof model (Fig. 4) of the skew profile inside N-domains, we proceed to disentangling the transcription- and replication-associated strand asymmetries in the human and mouse genomes. The replication bias is shown to significantly correlate with the replication fork polarity in various cell lines, while the transcription bias is shown to be driven by gene expression in mitotic spermatogonia.*

IV.5.1 Method

Our method to disentangle transcription- and replication-associated skews is based on the working model shown in Fig. 4. When superimposing the N-shaped replication profile and the transcription square-like skew profiles, we get the following theoretical skew profile in a replication N-domain :

$$S(x') = S_R(x') + S_T(x') = -2\delta\left(x' - \frac{1}{2}\right) + h + \sum_{genes} c_g \chi_g(x'), \qquad (16)$$

where position x' within the domain has been rescaled between 0 and 1, h and δ (> 0) are parameters that define the replication bias ($S_R^{5'} = h+\delta$ at the 5' N-domain extremity and $S_R^{3'} = h - \delta$ at the 3' extremity), χ_g is the characteristic function for the g^{th} gene that belongs to the N-domain (1 when the points is in the gene and 0 elsewhere) and c_g is its transcriptional bias calculated on the reference strand (likely to be positive for (+) genes and negative for (−) genes). For each N-domain detected as explained in Section IV.3 (see Fig. 11), we used a general least-square fit procedure to estimate, from the noisy S profile the parameters δ, h and each of the c_g's. The resulting χ^2 value was then used to select the N-domain candidates where

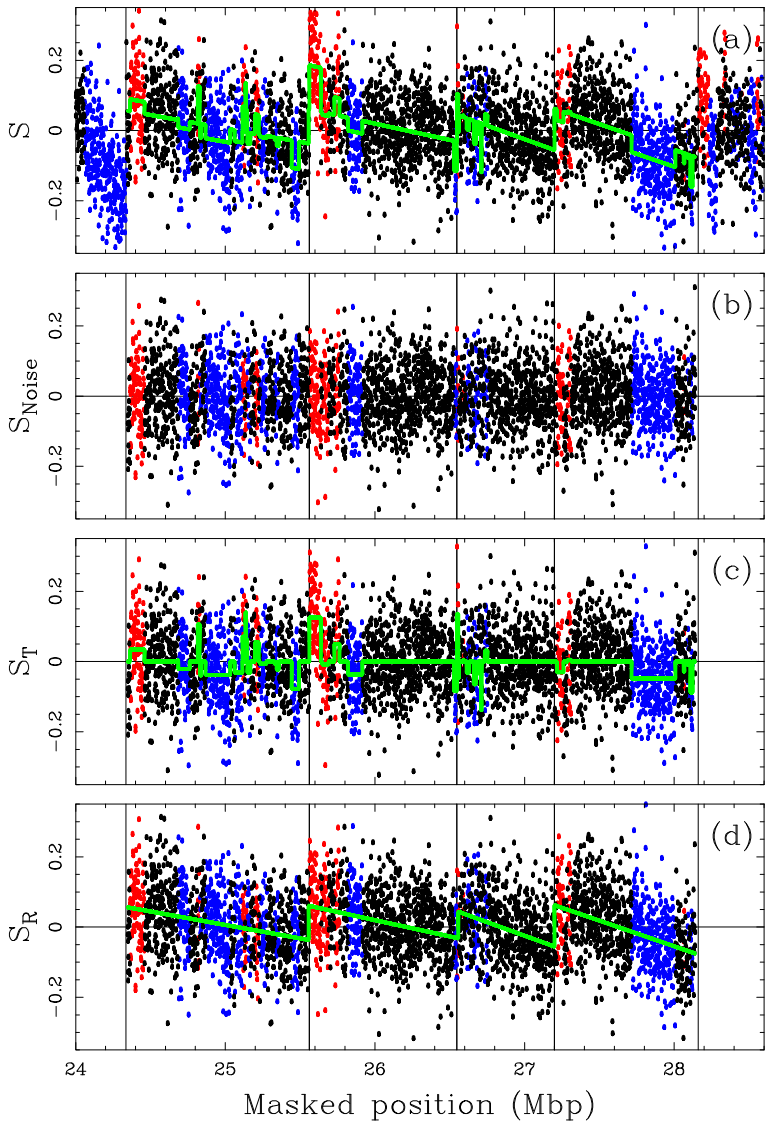

Figure 14: **Disentangling transcription- and replication-associated skew components.** (a) Skew profile S of a 4.3 Mbp repeat-masked fragment of human chromosome 6; each point corresponds to a 1 kbp window: red, (+) genes; blue, (−) genes; black, intergenic regions (the color is defined by majority rule); the estimated skew profile (Eq. (16)) is shown in green; vertical lines corresponds to the locations of 5 putative replication origins that delimit 4 adjacent domains identified by the wavelet-based methodology. (b) Noise component S_{Noise} obtained by subtracting the estimated total skew (green line in (a)) from the original skew profile in (a). (c) Transcription-associated skew S_T obtained by subtracting the estimated replication-associated profile (green lines in (d)) from the original S profile in (a); the estimated transcription step-like profile (third term of the *rhs* of (Eq. (16)) is shown in green. (d) Replication-associated skew S_R obtained by subtracting the estimated transcription step-like profile (green lines in (c)) from the original S profile in (a); the estimated replication serrated profile (first two terms in the *rhs* of (Eq. (16)) is shown in green (Baker et al. 2010).

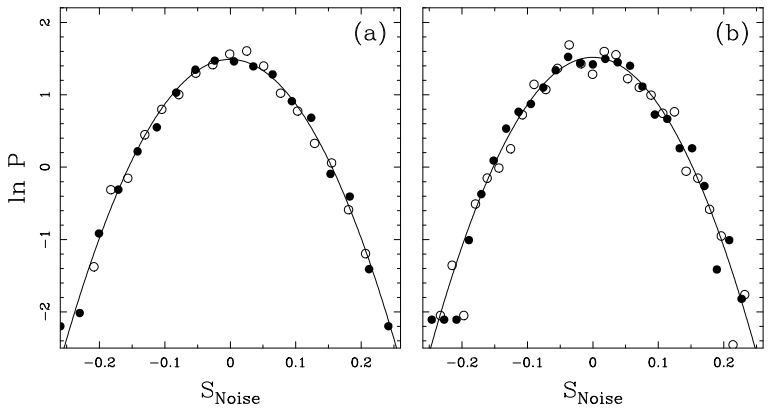

Figure 15: **Characterization of the noise component of the skew.** Semi-log representation of normalized histograms of skew values in: (a) a N-domain of length $L = 2.6$ Mbp (1.5 Mbp masked) in the human genome and (b) a N-domain of length $L = 2.3$ Mbp (1.3 Mbp masked) in the mouse genome. The symbols correspond to the histogram of the skew component S_{Noise} (•) and of the total skew increments $\delta S/\sqrt{2}$ (◦). The continuous parabola corresponds to Gaussian distributions of standard deviation $\sigma = 0.090$ (a) and 0.087 (b) (Baker et al. 2010).

the skew was well described by Eq. (16). As illustrated in Fig. 14 for a fragment of human chromosome 6 that contains 4 adjacent replication domains (Fig. 14(a)), this method provides a very efficient way to disentangle the square-like transcription skew component (Fig. 14(c)) from the N-shaped component induced by replication (Fig. 14(d)).

Remark. In the least-square fit procedure, we fixed the variance σ^2 of the Gaussian noise to the variance $\sigma^2 = \frac{1}{2}\sigma^2_{\delta S}$ computed in each N-domain from the probability distribution function of the skew increments $\delta S(n)/\sqrt{2} = [S(n+1) - S(n)]/\sqrt{2}$. As quantitatively verified *a posteriori* (Fig. 15), this variance directly estimated from the total skew S (Fig. 14(a)) is a very good approximation of the noise component of the skew once subtracted our model skew profile (Fig. 14(b)).

IV.5.2 Human autosomes

Among the 678 N-domains detected in the 22 human autosomes, our disentangling methodology failed to provide satisfactory results (prohibitive too large χ^2 values) for 14 domains only. We checked that the main reason for which our working hypothesis Eq. (16) did not apply was the fact that some regions presented anomalous high amplitude skew values. Hence, the results reported in the following correspond to a statistical analysis performed on 664 N-domains.

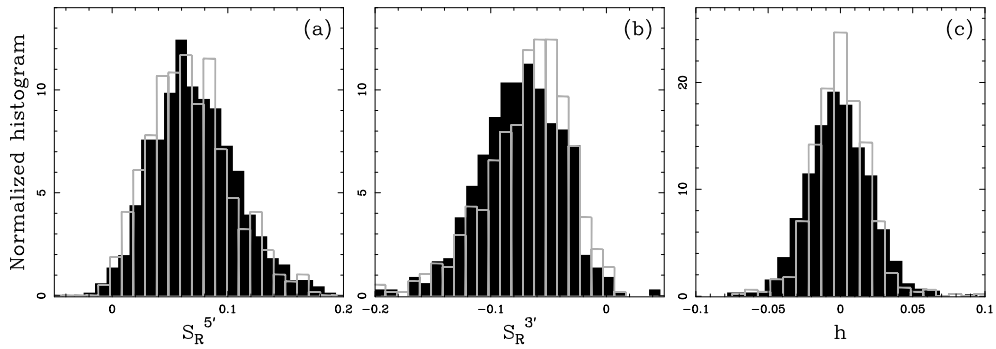

Figure 16: **Normalized histograms of replication parameters.** (a) $S_R^{5'}$, (b) $S_R^{3'}$ and (c) h estimated in 664 N-domains identified in the 22 human autosomes (black) and in 585 N-domains detected in the 19 mouse autosomes (gray) (see Table 2) (Baker et al. 2010).

	Number of N-domains	$\bar{S}_R^{5'}$	$\bar{S}_R^{3'}$	$\bar{h}$
Human	664	7.2 ± 0.1	-7.4 ± 0.1	$(-9.5 \pm 8.4)\,10^{-2}$
Mouse	585	6.8 ± 0.2	-6.8 ± 0.2	$(-1.8 \pm 7.9)\,10^{-2}$

Table 2: **Mean replication parameters.** $\bar{S}_R^{5'}$, $\bar{S}_R^{3'}$ and $\bar{h}$ (in percent $\pm$ SEM) computed with our wavelet-based disentangling method from human and mouse skew profiles (see Eq. (16)).

Replication bias.

In Fig. 16 are reported the results of our estimate of the replication parameters $S_R^{5'} = h + \delta$ (Fig. 16(a)), $S_R^{3'} = h - \delta$ (Fig. 16(b)), and h (Fig. 16(c)). The normalized histogram of the offset parameter h (vertical shift of the N profile) is symmetric (Fig. 16(c)), with a mean value $\bar{h} = (-9.5 \pm 8.4)\,10^{-4}$ (Table 2), that cannot be distinguished from zero. This means that the replication N-shaped profile is mainly observed centered at zero with equal statistical probability of upward and downward vertical shift by a few percent. The histograms of replication bias at the 5' (Fig. 16(a)) and 3' (Fig. 16(a)) N-domain extremities are quite symmetric one from each other with mean values $\bar{S}_R^{5'} = 7.2 \pm 0.1\%$ and $\bar{S}_R^{3'} = -7.4 \pm 0.1\%$, as expected from an antisymmetric N-shaped skew pattern of zero mean. Altogether these results provide some estimate of the mean replication bias $\bar{\delta} = 7.3 \pm 0.1\%$, which corresponds to an upward jump of mean amplitude $\simeq 14.6\%$ in the skew profile at the putative replication origins that border the replication N-domains. These estimations are consistent with those obtained in the pioneering study of large amplitude upward jumps in human skew profiles (Brodie of Brodie et al. 2005; Touchon et al. 2005).

As a test of the reliability of our methodology, we compare in Fig. 17 the results

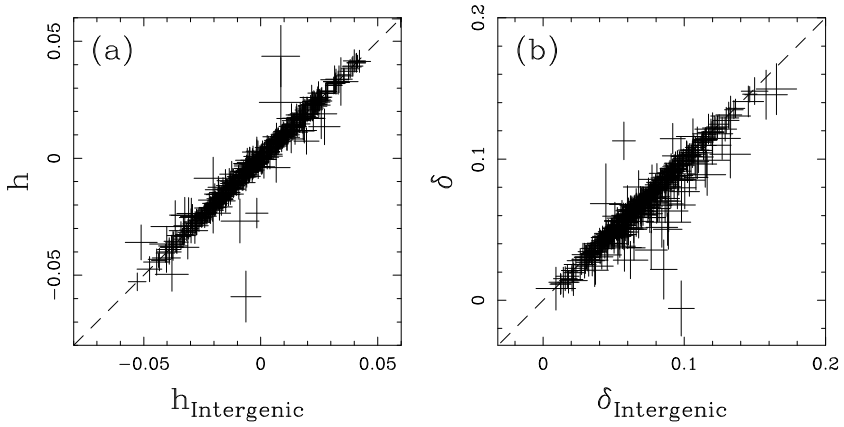

Figure 17: **Test of the consistency of our disentangling methodology.** For each N-domain detected in the 22 human autosomes, the replication parameters computed directly from the intergenic skew only are plotted versus the corresponding parameters derived from the method described in Section IV.5.1. (a) h *vs* $h_{intergenic}$; (b) δ *vs* $\delta_{intergenic}$. For clarity, only (437/664) N-domains containing more than 200 kbp of intergenic masked sequences are represented such that the error bars remain small enough (Baker et al. 2010).

of our estimates of the replication parameters h and δ for each N-domain to the corresponding values obtained directly from some fit of the S profile when considering only the intergenic regions where the observed skew is supposed to result from replication only. As observed in Fig. 17(a) for the parameter h and in Fig. 17(b) for the parameter δ, a large majority of points fall, up to the numerical uncertainty, on the diagonal. Thus, except for a small percentage of N-domains where intergenic coverage (Fig. 13) is too small to allow us to compute the replication parameters directly from the intergenic skew profile, the estimates reported in Table 2 are quite consistent with the skew values observed genome wide in intergenic sequences.

The replication bias correlates with the replication fork polarity

As discussed in the previous Chapters, the replication-associated $S_R(x)$ profile defined in Eq. (16) is predicted to be proportional to the replication fork polarity $p(x)$ (Chapter I, Eq. (168) in the summary), which itself is proportional to the derivative of the mean replication timing (Chapter II, Eq. (68) in the summary). These relations of proportionality can be written, using notations consistent with the previous Chapters, under the following form:

$$S_R(x) = S_R\, p(x) = v T_S\, S_R\, \mathrm{dMRT/dx}, \tag{17}$$

where v denotes the replication fork velocity, T_S the duration of the S-phase, MRT the mean replication timing and dMRT/dx its derivative. Note that the duration of the S-phase appears as experimental MRT profiles are expressed in S-phase fraction and not in unit of time. Of course Eq. (17) is expected to hold for the replication

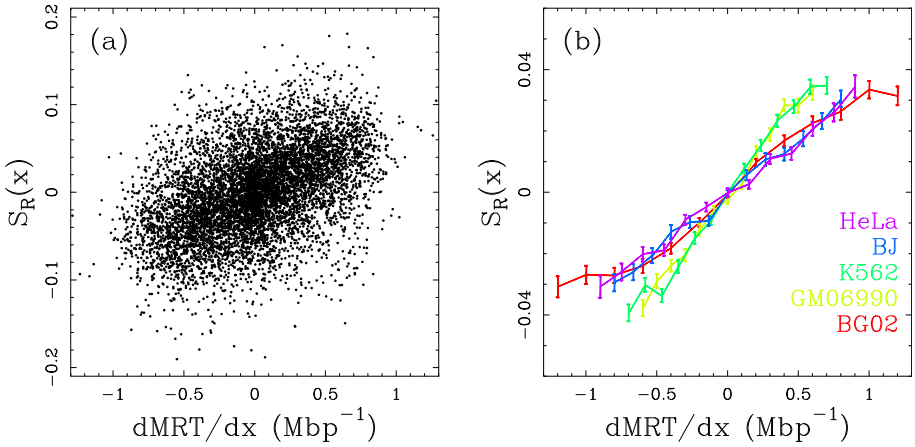

Figure 18: **The replication bias $S_R(x)$ is on average proportional to dMRT/dx.** Replication bias $S_R(x)$ (Eq. (16)) versus dMRT/dx in (a) GM6990 lymphoblastoid cell line (each point represents a 100 kbp window) and (b) in BG02 stem cell, GM06990 lymphoblastoid, K562 erythroid, BJ fibroblast and HeLa cell lines (mean values indicated, errors bars represent SEM).

	BG02	GM06990	K562	BJ	HeLa
BG02	1	0.38	0.41	0.36	0.34
GM06990	0.38	1	0.61	0.45	0.41
K562	0.41	0.61	1	0.43	0.41
BJ	0.36	0.45	0.43	1	0.52
HeLa	0.34	0.41	0.41	0.52	1
S_R	0.42	0.46	0.46	0.35	0.36

Table 3: **The replication bias $S_R(x)$ correlates with dMRT/dx.** Pearson correlation (R values) between replication-associated skew $S_R(x)$ (Eq. (16)) and dMRT/dx in BG02 embryonic stem cell, GM06990 lymphoblastoid, K562 erythroid, BJ fibroblast, and HeLa cell lines. $S_R(x)$ and dMRT/dx were calculated in non-overlapping 100 kbp windows located in skew N-domains (N=7751). All p-values are $< 10^{-16}$.

	$n^{(+)}$	$\bar{S}_T^{(+)}$	$n^{(-)}$	$\bar{S}_T^{(-)}$
Human	726	5.1 ± 0.2	731	-5.2 ± 0.2
Mouse	467	5.1 ± 0.2	481	-4.9 ± 0.2

Table 4: **Mean transcription bias** (in percent ± SEM) computed with our wavelet-based disentangling method from human and mouse skew profiles (see Eq. (16)). $\bar{S}_T^{(+)}$ and $\bar{S}_T^{(-)}$ are the mean transcription skews computed for (+) and (−) genes of length ≥ 20 kbp.

fork polarity determined in the germline. Unfortunately no replication timing data in germline cells are available to date. As a substitute to germline data, we used experimental replication timing data obtained in BG02 embryonic stem cell, GM06990 lymphoblastoid, K562 erythroid, BJ fibroblast, and HeLa cell line (Chen et al. 2010; Hansen et al. 2010). The $S_R(x)$ profile correlates significantly with dMRT/dx, as exemplified for the lymphoblastoid GM06990 cell line in Fig. 18(a). Interestingly the correlation between $S_R(x)$ and the dMRT/dx profiles in different cell lines is as high as between the dMRT/dx profiles themselves (Table3). In Fig. 18(b) we superimposed the average $S_R(x)$ profiles obtained in regions of given dMRT/dx values, for different indicated cell lines. The $S_R(x)$ profile is found to be proportional, on average, to the dMRT/dx profile in every cell line. The different slopes, equal to $vT_S\,S_R$ (Eq. (17)), can be attributed to cell specific values of v and T_S, but can also be attributed to the unequal conservation of the replication fork polarity among these cell lines and the germline. Indeed the replication fork polarity in one cell line needs not, even on average, be equal to the replication fork polarity in another cell line. For instance, in all examined cell lines, the regions of highest replication fork polarity have a replication bias of $\sim 4\%$, which is two fold lower than the highest replication bias observed at N-domain border $S_R^{5'} \sim 7\%$ (Table 2). Finally, we note that the coefficients $vT_S\,S_R$ obtained by linear regression of the $S_R(x)$ profile versus dMRT/dx are quite consistent with the values obtained in (Section III.3, Table 7). The linear regression of $S_R(x)$ versus dMRT/dx indeed yields $vT_S\,S_R = 2.76\%$, 5.69%, 5.20%, 3.61%, and 3.44% for respectively the BG02, GM06990, K562, BJ and HeLa cell lines while we obtained respectively $vT_S\,S_R = 2.78\%$, 5.26%, 5.45%, 3.56%, and 3.27% in (Section III.3, Table 7) .

Transcription bias.

To estimate the mean transcription bias of the genes belonging to a given N-domain, we considered the transcription-associated skew S_T obtained after subtracting the estimated N-shaped replication profile as illustrated in Fig. 14(c). Then as first proposed in (Touchon et al. 2004), the transcription skew of the genes was measured by averaging S_T over intron sequences after removing 490 bp at each intron extremities in order to get rid of the contribution to the skew coming from splicing signals. In Fig. 19 and Table 4 are reported the transcription bias so estimated for 726 sense (+) genes and 731 anti-sense genes (−) of length ≥ 20 kbp so that

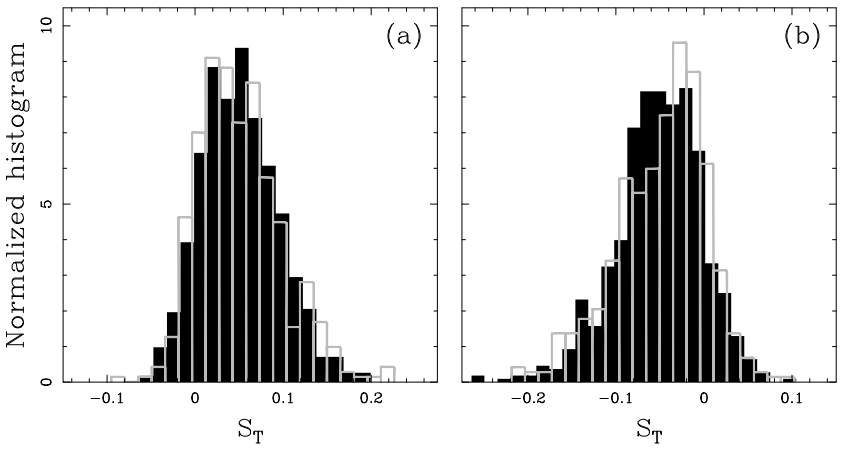

Figure 19: **Statistics of transcription bias.** Normalized histograms of transcription bias S_T computed as explained in the main text for human (black) and mouse (gray) genes of length ≥ 20 kbp. (a) sense (+) genes; (b) anti-sense (−) genes (see Table 4) (Baker et al. 2010).

	BG02	GM06990	K562	BJ	HeLa
S_R	0.36 (10^{-16})	0.41 (10^{-16})	0.40 (10^{-16})	0.30 (10^{-16})	0.31 (10^{-16})
S_T	0.09 ($2\,10^{-3}$)	0.13 ($8\,10^{-6}$)	0.09 ($2\,10^{-3}$)	0.14 ($8\,10^{-7}$)	0.13 ($6\,10^{-6}$)

Table 5: **The transcription bias S_T correlates poorly with dMRT/dx.** Pearson correlation between transcription bias S_T and replication bias S_R for human genes of length ≥ 20 kbp (N=1457) and dMRT/dx in B0G2 embryonic stem cell, GM06990 lymphoblastoid, K562 erythroid, BJ fibroblast, and HeLa cell lines (p-values are in parentheses). The skews S_T and S_R were computed on the coding strand.

the total intronic coverage is enough to ensure convergence in the estimate of the gene transcription skew. The histograms of S_T values for (+) and (−) genes are remarkably symmetric with means $\bar{S}_T^{(+)} = 5.1 \pm 0.2\%$ and $\bar{S}_T^{(-)} = -5.2 \pm 0.2\%$ respectively. This confirms that the local genic contribution of transcription to the total skew is of the same magnitude ($\sim 5\%$) than the contribution induced by replication ($\sim 7.5\%$) which can be seen *a posteriori* as a justification of the need to develop a methodology capable to disentangle these two skew components. Interestingly, the transcription-associated S_T correlates much lower with dMRT/dx than the replication-associated S_R (Table 5). Recently, the compositional asymmetry in human genes was shown to significantly correlate with the expression in germline cells, especially in spermatogonia (McVicker and Green 2010). As a perspective, it will be interesting to correlate the transcription-associated skew S_T with germline expression data in the human genome.

IV.5.3 Mouse autosomes

Among the 587 N-domains detected in the 19 mouse autosomes, only 2 were discarded by our disentangling methodology as incompatible with our working model Eq. (16). The results presented in this section thus correspond to a statistical analysis performed on 585 N-domains.

Replication and transcription biases.

As shown in Fig. 16, the histograms of replication parameters ($S_R^{5'}$, $S_R^{3'}$, h) values computed with our methodology cannot be statistically distinguished from the ones previously obtained for the human autosomes. The detected N-domains have a zero mean replication skew ($\bar{h} = (-1.8 \pm 7.9)\,10^{-4}$) and an antisymmetric shape with $\bar{S}_R^{5'} = 6.8 \pm 0.2\%$ and $\bar{S}_R^{3'} = -6.8 \pm 0.2\%$ (Table 2) corresponding to an upward jump in the mouse skew profile, at the detected putative replication origins, of characteristic amplitude $\sim 13.6\%$ similar to the $\sim 14.4\%$ previously observed for the human autosomes. Similarly, the estimates of the transcription bias for sense (Fig. 19(a)) and antisense (Fig. 19(b)) mouse genes are in remarkable agreement with those obtained for human genes. As reported in Table 4, we got the following mean values $\bar{S}_T^{(+)} = 5.1 \pm 0.2\%$, $\bar{S}_T^{(-)} = -4.9 \pm 0.2\%$ for sense and antisense genes respectively. To corroborate the analyses made in the human genome (Section IV.5.2), it would be interesting to correlate the replication-associated profile $S_R(x)$ with dMRT/dx, using replication timing data obtained in mouse (Farkash-Amar et al. 2008).

The transcription bias correlates with gene expression in spermatogonia cells.

We further focused on a key group of 315 protein-coding genes that were found to be differentially expressed between male germ cells and somatic controls and to display highly similar meiotic and post-meiotic patterns of transcriptional induction across human, mouse and rat populations (Chalmel et al. 2007). In the male germ line, the spermatogonia cell undergoes the greater number of mitoses. Thus, among male germline cells, the time spent as a spermatogonia cell is probably the longest, and the gene expression in spermatogonia is expected to have the greatest impact on the transcription bias. The transcription-induced component S_T of the skew correlates significantly with the expression level in mitotic spermatogonia ($R = 0.36$, p-value$= 6.9\,10^{-11}$, Fig. 20(a)), on the contrary to the replication-induced component S_R ($R = 0.17$, p-value$= 2\,10^{-3}$). Among germline cells, S_T increases with the expression level only in mitotic spermatogonia cells (Fig. 20(b)). In contrast with mitotic spermatogonia, no significant positive correlations were found between S_T and expression data in meiotic spermatocytes and post meiotic spermatids (Table 6). This result indicates that the transcription bias is indeed driven by the gene expression in spermatogonia, as observed in the human genome (McVicker and Green 2010).

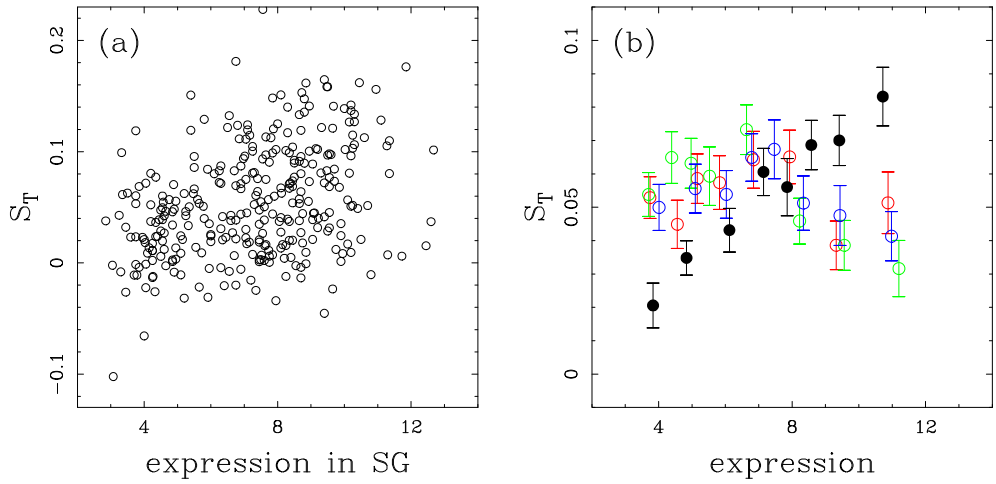

Figure 20: **The transcription bias S_T increases in magnitude with gene expression in mitotic spermatogonia.** Transcription skew S_T versus the expression in (a) mitotic spermatogonia (each circle represents a gene) and (b) in mitotic spermatogonia (black), meiotic spermatocytes (green), post-meiotic spermatids (red), and seminiferous tubules (blue) (mean values computed on 40 genes, errors bars represent SEM). The expression data correspond to the key group of 315 protein-coding genes with conserved transcriptome in human and rodent male gametogenesis identified in (Chalmel et al. 2007). Importantly these genes are *differentially expressed* in spermatogonia, spermatocytes, spermatids and tubules. The skew S_T was computed on the coding strand.

	SG	SC	ST	TU
S_T	0.36 ($6.9\,10^{-11}$)	-0.019 (0.73)	-0.190 ($7.7\,10^{-4}$)	-0.042 (0.5)
S_R	0.17 ($2.1\,10^{-3}$)	0.025 (0.65)	-0.058 (0.30)	0.051 (0.36)

Table 6: **The transcription bias S_T reflects gene expression in mitotic spermatogonia.** Pearson correlation between expression data in germline cells (SG=spermatogonia, SC= spermatocytes, ST= spermatids, TU=tubules) and the transcription bias S_T or the replication bias S_R (p-values are in parentheses). The expression data correspond to the key group of 315 protein-coding genes with conserved transcriptome in human and rodent male gametogenesis identified in (Chalmel et al. 2007). Importantly these genes are *differentially expressed* in spermatogonia, spermatocytes, spermatids and tubules. The skews S_T and S_R were computed on the coding strand.

Summary of Chapter IV

In this Chapter, we developed a multi-scale methodology based on the continuous wavelet transform with an adapted (N-shaped) analyzing wavelet. The implementation of this method allowed us to delineate 678 and 587 skew N-domains in the human and mouse genomes respectively. The skew N-domains correspond to large-scale genomic structures (from 300 kbp up to 3 Mbp size), and cover a significant proportion of the human and mouse genomes (22.3% and 33.8% respectively). The skew profile inside N-domain is well described by the factory roof model (Fig. 4): genes add transcription square-like components to the global skew N-shape attributed to replication. Skew N-domains are proposed to be replication domains in the germline, characterized by a N-shaped replication fork polarity, with putative replication origins at their edges. Following the factory roof model, we disentangled inside N-domains the contributions associated to transcription and replication in the skew profile. The confrontation with replication timing data and germline expression data strongly supports the relevance of this decomposition. In the human genome, the replication bias strongly correlates with the dMRT/dx profile (estimator of replication fork polarity) in various cell lines (Fig. 18 and Table 3), while the transcription bias correlates poorly (Table 5). In the mouse genome, the transcription bias strongly correlates with the expression in germline (Fig. 20(a)), whereas the replication bias correlates much less significantly (Table 6). Among germline cells, it is actually the expression in mitotic spermatogonia that drives the correlation with the transcription bias (Table 6 and Fig. 20(b)), as observed in the human genome (McVicker and Green 2010). Finally, as the replication fork polarity is proportional to the derivative of the mean replication timing, the putative N-shape replication fork polarity in skew N-domains suggests that the mean replication timing has a parabolic U-shape. We thus conclude that skew N-domains are replication domains in the germline, characterized by a U-shape mean replication timing, with initiation zones at their edges.

Chapter V

Replication domains are self-interacting chromatin structural units

In Chapter IV we showed that 30% of the human genome was covered by large-scale ($\sim$ 1 Mbp) N-shaped compositional skew domains, the so-called N-domains. According to the relationships established in Chapters I to IV between replication-associated strand asymmetry, replication fork polarity and replication timing, skew N-domains were proposed to be replication domains in the germline, characterized by a U-shape of the replication timing. Here, we show that replication timing U-domains are robustly observed in several cell lines as covering about half of the human genome. Significant numbers of U-domains coincide with skew N-domains, which indirectly supports their interpretation as germline replication domains. However a majority of U-domains are cell line specific and therefore belong to genomic regions of high replication timing plasticity. As previously observed for skew N-domains (Audit et al. 2009), U-domain borders stand out from their environment by a localized ($\sim$ 300 kbp) open chromatin structure. Long-range chromatin interaction data (Hi-C) further suggests that U-domains correspond to self-interacting chromatin structural units. The compartmentalization of the genome into replication U-domains provides new insights on the organization of the replication program in the human genome.

V.1 Introduction

Comprehensive knowledge of genetic inheritance at different development stages relies on elucidating the mechanisms that regulate the DNA spatio-temporal replication program and its possible conservation during evolution (Gilbert 2010). In multi-cellular organisms, there is no clear consensus sequence where initiation may occur (Berezney et al. 2000; Bell and Dutta 2002). Instead epigenetic mechanisms

may take part in the spatial and temporal control of replication initiation in higher eukaryotes in relation with gene expression (Bogan et al. 2000; Méchali 2001; McNairn and Gilbert 2003; Aladjem 2007; Courbet et al. 2008; Méchali 2010). For many years, understanding the determinants that specify replication origins has been hampered by the small number (approximately 30) of well-established replication origins in the human genome and more generally in mammalian genomes (Aladjem 2007; Hamlin et al. 2008; Gilbert 2010). Recently, nascent DNA strands synthesized at origins were purified by various methods (The ENCODE Project Consortium 2007; Cadoret et al. 2008; Karnani et al. 2009; Mesner et al. 2011) to map a few hundreds putative origins in 1% of the human genome. For unclear reasons, the concordance between the different studies is very low (from $< 5\%$ to $< 25\%$) (Cadoret et al. 2008; Karnani et al. 2009; Hamlin et al. 2010; Mesner et al. 2011).

In a completely different approach to map replication origins, previous *in silico* analyses of the nucleotide compositional skew $S = (T - A)/(T + A) + (G - C)/(G + C)$ of the human genome showed that the sign of S abruptly changed from $(-)$ to $(+)$ when crossing known replication initiation sites. This allowed A. Arneodo and collaborators to predict putative origins at more than a thousand sites of S sign inversion (S-jumps) along the human genome (Brodie of Brodie et al. 2005; Touchon et al. 2005) (see Fig. 3 of Chapter IV). Further analyses of S patterns identified 663 megabase-sized N-domains whose skew profile displays a N-like shape (Fig. 1A), with two abrupt S-jumps bordering a DNA segment whose skew linearly decreases between the two jumps (Brodie of Brodie et al. 2005; Touchon et al. 2005; Audit et al. 2007; Huvet et al. 2007; Baker et al. 2010; Arneodo et al. 2011). Skew N-domains have a mean length of 1.2 ± 0.6 Mbp and cover 29.2% of the human genome (Section IV.4). The initiation zones predicted at N-domains borders would be specified by an open chromatin structure favorable to early replication initiation and permissive to transcription (Audit et al. 2009; Arneodo et al. 2011). The determination of HeLa replication timing profile (Chen et al. 2010) and the analysis of available timing profiles in several human cell lines (Woodfine et al. 2005; Desprat et al. 2009; Hansen et al. 2010; Ryba et al. 2010; Yaffe et al. 2010) confirmed that significant numbers of N-domains borders colocalize with early initiation zones (Audit et al. 2007; Chen et al. 2011).

Recent studies have shown that replication induces different mutation rates on the leading and lagging replicating strands (Chen et al. 2011). This asymmetry of rates acting during evolution has generated the skew upward jumps that result from the abrupt inversion of replication fork polarity at N-domain extremities. Inside N-domains, the linear decrease of the skew (Fig. 1A) likely reflects a progressive inversion of the replication fork polarity. This organization of replication in a large proportion of the genome contrasts with the previously proposed segmentation of mammalian chromosomes in regions replicated either by multiple synchronous origins with equal proportion of forks coming from both directions (0.2-2.0 Mbp Con-

stant Timing Regions) or by unidirectional replication forks (0.1-0.6 Mbp Transition Timing Regions) (Farkash-Amar et al. 2008; Hiratani et al. 2008; Desprat et al. 2009; Ryba et al. 2010). According to Eq. (68) of Chapter II, we expect the derivative of the replication timing in the germline to be shaped as a N inside skew N-domains. In this Chapter, we show that the corresponding U-shape of the replication timing profile is not specific to the germline but is generally observed in all replication timing profiles examined, thus establishing these "U-domains" as a new type of replication domains. As previously observed with the early initiation zones bordering N-domain extremities, those specific to the U-domains are significantly enriched in open chromatin markers as well as insulator-binding proteins CTCF (Phillips and Corces 2009; Ohlsson et al. 2010) and are prone to gene activity.

DNA replication and transcription require great reproducibility and coordination, all this in the crowded environment of the cell nucleus. Regulation of these complex processes may partly rely on the conformation and dynamics of the chromatin fiber that ultimately condition DNA sequence accessibility. The chromatin fiber is a nucleoprotein filament with non homogeneous structural and mechanical properties (Wolffe 1998; Horn and Peterson 2002). This heterogeneity evidently affects how the fiber folds and organizes into higher order structures like loops, coil or chromonema. But, despite increasing experimental (Belmont 2001; Cremer and Cremer 2001; Gasser 2002; Dekker 2003; Gilbert et al. 2005; Branco and Pombo 2006; Shopland et al. 2006) and modeling (Cook 1995; Sachs et al. 1995; Ostashevsky 1998; Münkel et al. 1999) efforts, the so-called tertiary chromatin structure is still very controversial (Cook 2001; Müller et al. 2004) and the possible role, if any, of the DNA sequence at such large scales remains an enigma. Although the existence of chromatin loops, ranging in size from several kbp to 10 Mbp or more (Dekker 2003; Chambeyron and Bickmore 2004; Müller et al. 2004) has been extensively discussed in the literature, it has generally been inferred from indirect assays. Chromatin loops were proposed to result from the clustering of DNA and/or RNA polymerases (Cook 1995, 2002). In this model, non specific entropic forces between DNA and/or RNA polymerases, already engaged on the chromatin fiber, are supposed to drive the aggregation of these polymerases thereby promoting the genome compartmentalization into rosette-like multi-loop patterns containing several thousands (and sometimes millions) of base pairs. This dynamical multi-loop model has been emphasized as providing a very attractive description of replication foci and transcription factories (Cook 1999).

In this Chapter, we analyse recent Hi-C data (Lieberman-Aiden et al. 2009) and show that replication U-domains likely correspond to self-interacting structural chromatin units. These data actually suggest that the "islands" of open chromatin observed at U-domains borders are at the heart of a compartmentalization of chromosomes into chromatin units of independent replication and of coordinated gene transcription.

V.2 From compositional skew N-domains to replication timing U-domains

V.2.1 Linking replication fork polarity to nucleotide compositional skew profile and replication timing

In Chapter I (Eq. (168) in the summary), we argued that the skew S resulting from mutational asymmetries associated to replication was proportional to the replication fork polarity p:

$$S(x) \sim p(x). \tag{1}$$

The linear decrease of S in N-domains from positive ($5'$ end) to negative ($3'$ end) values would thus reflect a linear decrease of the replication fork polarity with a change of sign in the middle of the N-domains. This result strongly supports the interpretation of N-domains (Fig. 1A-C) as replication units in germline cells. In Chapter II (Eq. (68) in the summary), under the central hypotheses that the replication fork velocity is constant and that replication is bidirectional from each origin, we demonstrated that the replication fork polarity was proportional to the derivative of the mean replication timing (MRT):

$$p(x) \sim \mathrm{dMRT/dx}. \tag{2}$$

The fork polarity should therefore provide a direct link between the skew S and the derivative of the replication timing profile in germline cells.

To test this relationship, we used, as a substitute to germline MRT, the replication timing profiles of seven somatic cell lines (one embryonic stem cell, three lymphoblastoid, a fibroblast, an erythroid and HeLa cell lines) (Chen et al. 2010; Hansen et al. 2010) (Section V.4). We first correlated the skew S with dMRT/dx, in the BG02 embryonic stem cells, over the 22 human autosomes (Fig. 1D). The significant correlations observed in intergenic ($R = 0.40$, $P < 10^{-16}$), genic (+) ($R = 0.34$, $P < 10^{-16}$) and genic (−) ($R = 0.33$, $P < 10^{-16}$) regions are representative of the correlations observed in the other 6 cell lines (Table 1). These correlations are as important as those obtained between the dMRT/dx profiles in different cell lines (Table 2), as well as those previously reported between the replication timing data themselves (Hansen et al. 2010; Ryba et al. 2010; Yaffe et al. 2010). The correlations between S and dMRT/dx are even stronger when focusing on the 663 skew N-domains (Table 1). The correlations obtained in intergenic regions ($R = 0.45 \pm 0.06$) are recovered to a large extent in genic regions ($R = 0.34 \pm 0.03$) where the transcription-associated skew S_T was hypothesized to superimpose to the replication-associated skew S_R (Audit et al. 2007; Huvet et al. 2007; Baker et al. 2010) (Fig. 4 of Chapter IV and Section IV.2.4). Further evidence of this link between S and dMRT/dx was obtained when averaging, for the different cell lines, the dMRT/dx profiles inside the 663 skew N-domains after rescaling their length to

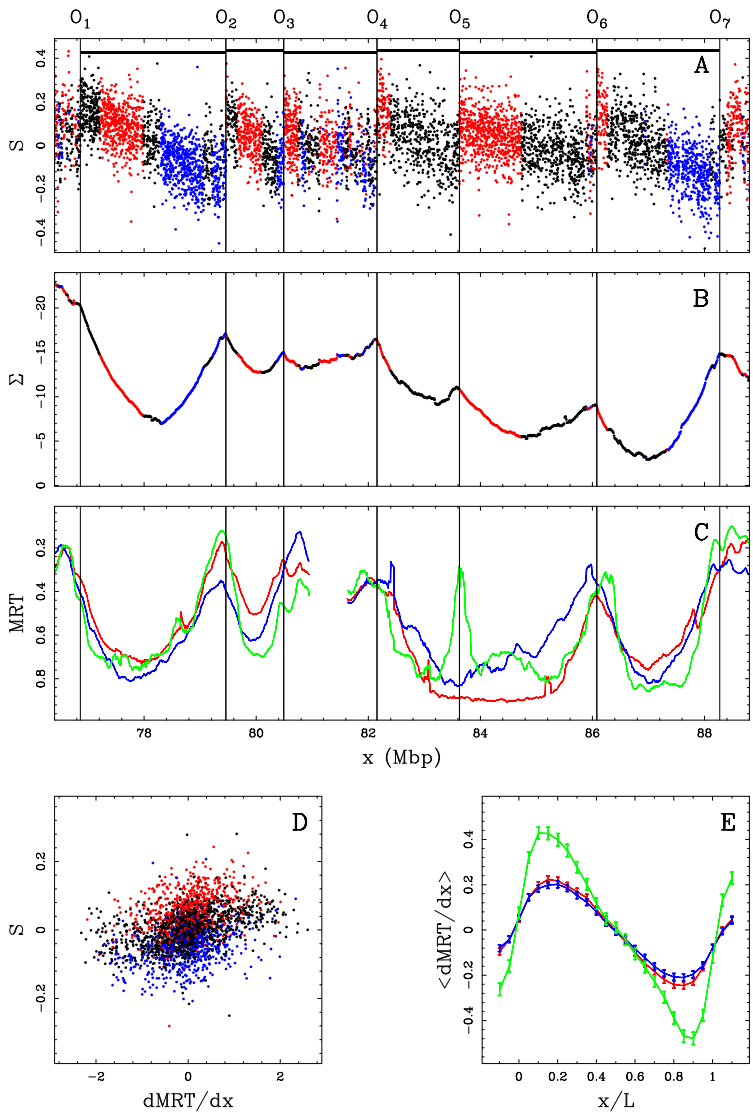

Figure 1: **Comparing skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ and mean replication timing (MRT).** (A) S profile along a 11.4 Mbp long fragment of human chromosome 10 that contains 6 skew N-domains (horizontal black bars) bordered by 7 putative replication origins O_1 to O_7. Each dot corresponds to the skew calculated for a window of 1 kbp of repeat-masked sequence. The colors correspond to intergenic (black), (+) genes (red) and (−) genes (blue). (B) Corresponding cumulative skew profile Σ obtained by cumulative addition of $S-$values along the sequence. (C) MRT profiles from early, 0 to late, 1 for BG02 (green), K562 (red) and GM06990 (blue) cell lines. (D) Correlations between S and dMRT/dx, in BG02 (100 kbp windows) along the 22 human autosomes; colors as in (A); the corresponding Pearson correlations are given in Table 1. (E) Average dMRT/dx profiles (± SEM) in the 663 skew N-domains after rescaling their length L to unity; colors as in (C).

R	BG02	K562	GM06990	H0287	TL010	BJ	HeLa
GW (+)	0.34	0.36	0.35	0.34	0.33	0.31	0.33
GW (i)	0.40	0.45	0.42	0.41	0.41	0.35	0.32
GW (−)	0.33	0.37	0.34	0.35	0.34	0.33	0.34
Ndom (+)	0.36	0.43	0.42	0.42	0.41	0.32	0.38
Ndom (i)	0.45	0.50	0.48	0.48	0.47	0.38	0.35
Ndom (−)	0.35	0.44	0.44	0.43	0.41	0.40	0.40

Table 1: **The compositional skew correlates with dMRT/dx.** Pearson correlation (R values) between the skew S and dMRT/dx, from different cell lines . S and dMRT/dx were calculated in non-overlapping 100 kbp windows genome wide (GW) and in the 663 skews N-domains (Ndom). Each 100 kbp window was classed as intergenic (i), genic (+) or genic (−) by majority rule. All p-values are $< 10^{-16}$.

R	BG02	K562	GM06990	H0287	TL010	BJ	HeLa
BG02	1	0.42	0.39	0.39	0.35	0.39	0.36
K562	0.42	1	0.58	0.57	0.56	0.43	0.39
GM06990	0.39	0.58	1	0.9	0.84	0.47	0.41
H0287	0.39	0.57	0.9	1	0.84	0.47	0.41
TL010	0.35	0.56	0.84	0.84	1	0.45	0.37
BJ	0.39	0.43	0.47	0.47	0.45	1	0.52
HeLa	0.36	0.39	0.41	0.41	0.37	0.52	1

Table 2: **Conservation of dMRT/dx across differentiation.** Pearson correlation (R values) of the derivative of MRT, dMRT/dx, between different pairs of human cell lines (Methods). dMRT/dx was calculated in non-overlapping 100 kb windows over the 22 human autosomes. All p-values are $< 10^{-16}$.

unity (Fig. 1E). These mean profiles are shaped as a N, suggesting that some properties of the germline replication program associated with the pattern of replication fork polarity are shared by somatic cells.

V.2.2 Replication timing U-domains are robustly observed in human cell lines

According to Eqs. (1) and (2), the integration of the skew S is expected to generate a profile rather similar to the replication timing profile. In segments of linearly changing skew, the integrated S function is thus expected to show a parabolic profile. The integrated S function when estimated by the cumulative skew Σ (Fig. 1B) along N-domains of a 11.4 Mbp long fragment of human chromosome 10, indeed displays a U-shaped (parabolic) profile likely corresponding the replication timing profile in the germline. Remarkably, the 6 N-domains effectively correspond to successive genome regions where the MRT in the BG02 embryonic stem cells is U-shaped (Fig. 1C). The 7 putative initiation zones (O_1 to O_7) corresponding to upward S-jumps (Fig. 1A),

	Ndom	BG02	K562	GM06990	H0287	TL010	BJ	HeLa
N	663	1534	876	882	830	664	1150	1422
L	1.19	1.09	1.42	1.52	1.57	1.62	1.19	1.06
G	29.2	61.9	46.1	49.5	48.1	39.6	50.5	55.7
GC	40.30	40.25	40.84	40.85	40.94	41.13	40.84	40.72

Table 3: **Replication timing U-domains detected in different cell lines using our wavelet-based methodology**. N = number, L = mean length (Mbp), G = genome coverage (%), GC = mean GC-content (%) of the replication timing U-domains found in the 22 human autosomes. Corresponding data for the skew N-domains (replication domains in the germline) are given for comparison.

co-locate (up to the $\sim$ 100 kbp resolution) with MRT local extrema which supports that they are highly active in BG02. These initiation zones can present cell specificity as exemplified by the putative replication origin O_5 which is inactive (or late) in both the K562 erythroid and GM06990 lymphoblastoid cell lines (Fig. 1C) resulting in domain "consolidation" (Hiratani et al. 2010). Two neighboring U-domains ($[O_4, O_5]$ and $[O_5, O_6]$) in BG02 merged into a larger U-domain in the K562 and GM06990 cell lines. Note that the other 3 N-domains ($[O_1, O_2]$,$[O_2, O_3]$, and $[O_6, O_7]$) are replication timing U-domains common to BG02, K562 and GM06990.

To detect U-domains in replication timing profiles at genome scale, we developed a wavelet-based methodology (see Section V.2.3) which allowed us to identify in the 7 human cell lines from 664 (TL010) up to 1534 (BG02) U-domains of mean size ranging from 1.06 Mbp (HeLa) up to 1.62 Mbp (TL010) and covering from 39.6% (TL010) to 61.9% (BG02) of the genome (Table 3). For each cell line, the average MRT profile of U-domains has an expected parabolic shape (Fig. 2A) representative of individual U-domains (Figs. 2C, 3A and 4A). Inside the U-domains, the derivative dMRT/dx is N-shaped (Figs. 2D, 3B and 4B) like the skew profile inside N-domains (Figs. 3F and 4F). When rescaling the size of each U-domains to unity for a given cell line, these profiles superimpose onto a common N-shaped curve well approximated by the average dMRT/dx profile (Fig. 2B).

To determine the amounts of U-domains conserved in different cell types, we computed for each cell type pair the mutual covering of the corresponding sets of U-domains (two U-domains are shared by two different cell lines if each domain covers more than 80% of the other domain (Table 4)). Taking as reference the matching obtained for the three lymphoblastoid cell lines (from $\sim$ 40 to 65%; Table 4), the matchings between the other cell lines were statistically significant. The number of U-domain shared by cell type pairs were all significantly larger than the number expected by chance ($P < 10^{-3}$; Section V.4). For example BG02 shares 197 and 189 U-domains with K562 and GM06990 respectively, when only 45 and 46 are expected by chance. This corresponds to a significant proportion ($\sim$20%) of the U-domains

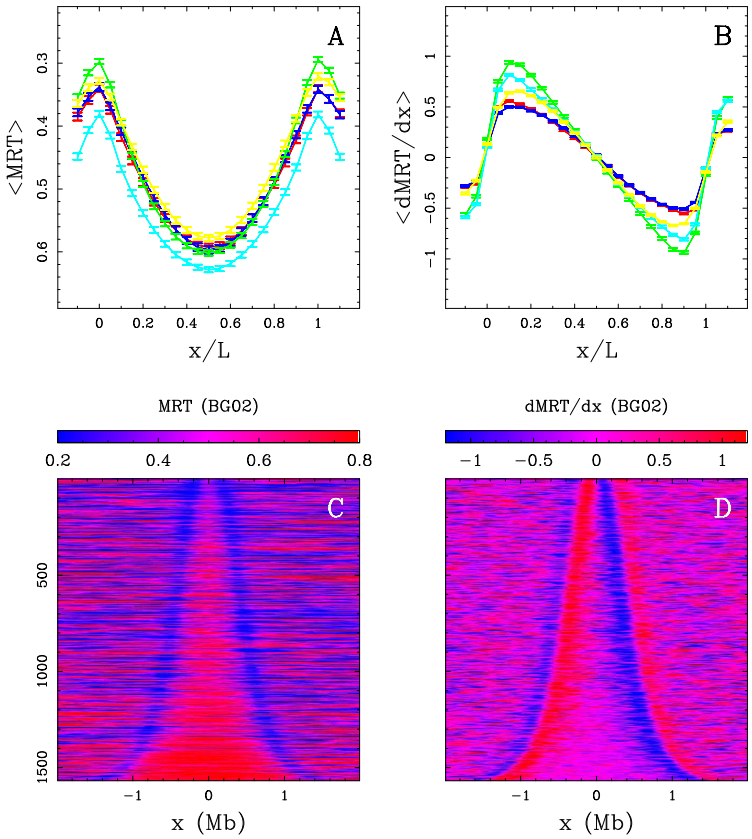

Figure 2: **Replication timing U-domains in different human cell lines:** BG02 (green), K562 (red), GM06990 (blue), BJ (magenta), and HeLa (cyan). (A) Average MRT profiles (± SEM) inside detected replication U-domains (Table 3). (B) Corresponding average dMRT/dx profiles (± SEM). (C) The 2534 BG02 U-domains were centered and ordered vertically from the smallest (top) to the longest (bottom). The MRT profile of each domain is figured along a horizontal line using the MRT (BG02) color map. (D) Same as in (C) but for dMRT/dx using the dMRT/dx (BG02) color map.

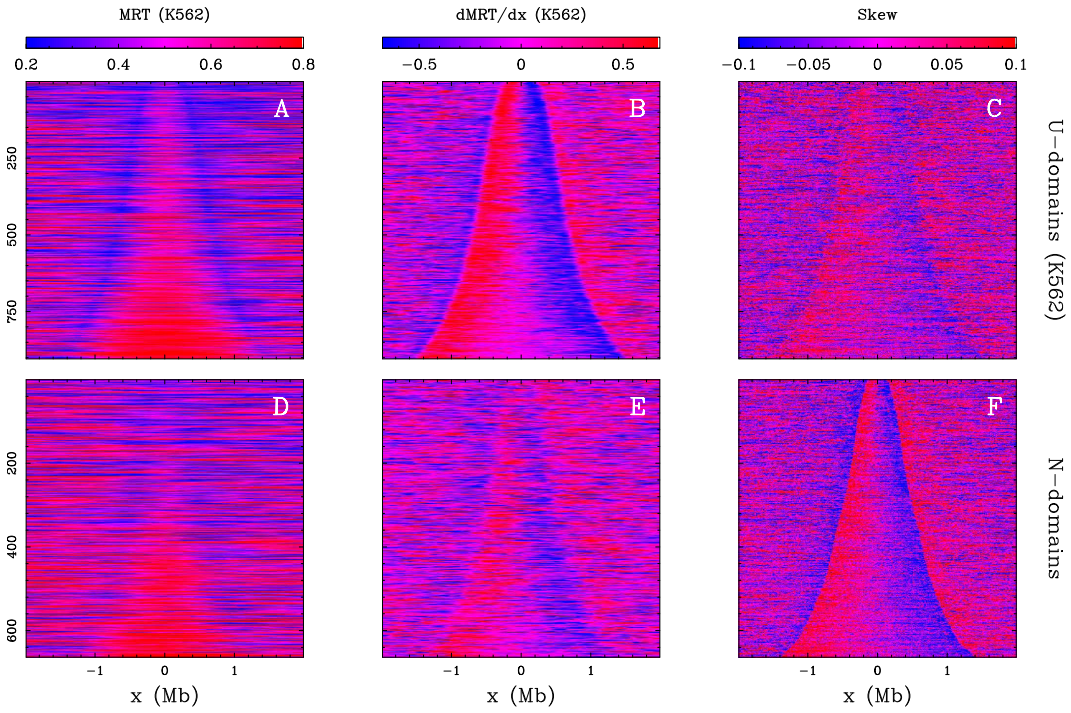

Figure 3: **The replication U-domains in the erythroid K562 cell line significantly match the skew N-domains.** The 876 replication timing U-domains detected in K562 cell line were centered and ordered vertically from the smallest (top) to the largest (bottom): the MRT (A), dMRT/dx (B), and skew S (C) profiles of each domain are figured along a horizontal line using the corresponding color maps. Same representation of the MRT (D), dMRT/dx (E), and S (F) profiles in the 663 skew N-domains.

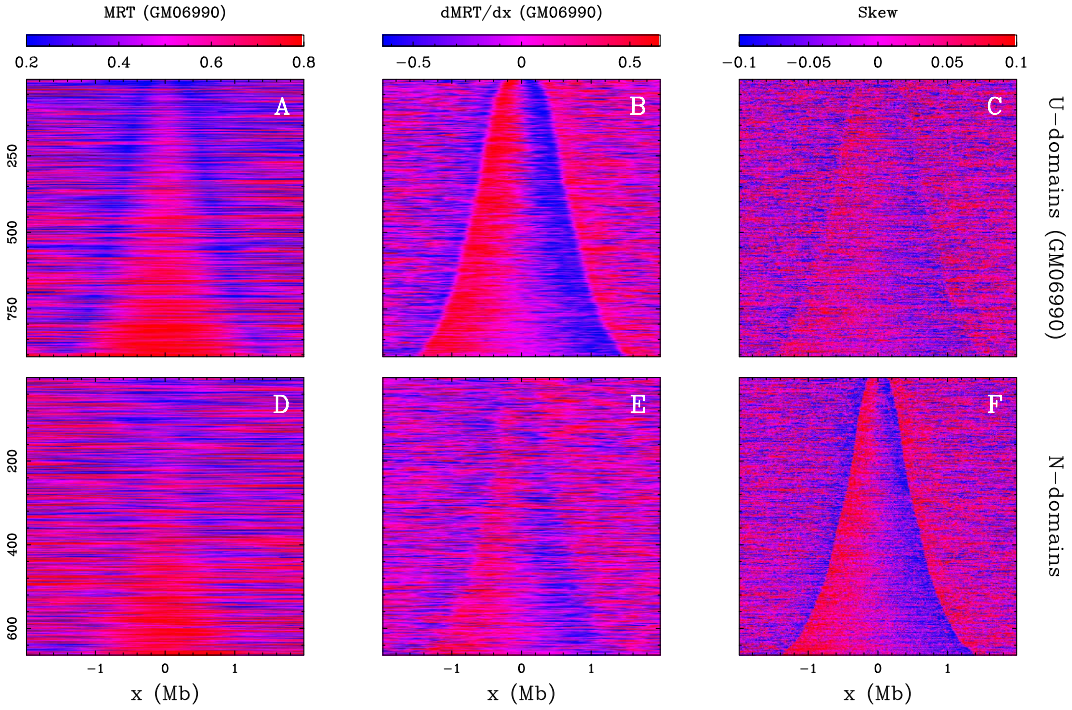

Figure 4: **The replication U-domains in the lymphoblastoid GM06990 cell line significantly match the skew N-domains.** Same as in Fig. 3 but for the lymphoblastoid GM06990 cell line (882 replication timing U-domains).

	Ndom	BG02	K562	GM06990	H0287	TL010	BJ	HeLa
Ndom	100	10.2	13.6	13.5	13.1	13	7.22	8.44
BG02	23.7	100	22.5	21.4	20.5	17.9	18	16.7
K562	17.9	12.8	100	28.5	28.8	30.9	16	13.9
GM06990	17.9	12.3	28.7	100	64.6	56.2	16	12.2
H0287	16.4	11.1	27.3	60.8	100	56.6	16.9	13
TL010	13	7.76	23.4	42.3	45.3	100	12.3	9.21
BJ	12.5	13.5	21	20.9	23.4	21.2	100	23.5
HeLa	18.1	15.5	22.5	19.6	22.3	19.7	29	100

Table 4: **Percentage of matchings between replication timing U-domains in different cell lines including skew N-domains in the germline**. A U-domain in a given cell line (column) was considered as matching a U-domain in another cell line (row) if more than 80% nucleotides of each of these U-domains were common to the two domains. For instance, 23.7% of skew N-domains match a BG02 U-domain, while 10.2% BG02 U-domains match a skew N-domain.

of the individual cell lines (Table 4), as compared to the matchings ($\lesssim 5\%$) expected by chance. A significant percentage of N-domains correspond to U-domains (*e.g.* from 12.5% in BJ up to 23.7% in BG02). This explains that when representing the MRT profile of K562 and GM06990 instead of the skew S, along the set of N-domains ordered according to their size, we can recognize the edges of many N-domains (Figs. 3D and 4D respectively). The same observation can be made when comparing the dMRT/dx profiles (Figs. 3E and 4E respectively) to the corresponding skew profiles (Figs. 3F and 4F). Note that the N-domains match only $7-14\%$ of the U-domains of various cell lines due to the very stringent N-domain selection criteria (Huvet et al. 2007; Baker et al. 2010) that yielded only 663 N-domains (29.2% of the genome) as compared to much larger U-domain numbers (Table 3). Replication timing U-domains are robustly observed in all cell lines, covering $\sim 50\%$ of the human genome. For each cell type, about half U-domains are shared by at least another cell line, namely BG02 (38.4%), K562 (61%), GM06990 (59.2%), BJ (51.6%), HeLa (44.7 %). This is also true for the skew N-domains (50.2%) that likely correspond to replication timing U-domains in the germline. Conversely this equally means that about half U-domains are specific to only one cell line: for instance 61.6% BG02 U-domains are only encountered in the BG02 cell line, and for the other cell line these percentage are K562 (39%), GM06990 (40.8%), BJ (48.4%), HeLa (55.3%). Therefore about half of the human genome that is covered by U-domains corresponds to regions of high replication timing plasticity where replication domains may (i) reorganize according to the so-called "consolidation" scenario (merging of two U-domains into a larger one) (Fig. 1C), (ii) experience some boundary shift and (iii) emerge in a late replicating region as previously observed in the mouse genome during differentiation (Hiratani et al. 2010).

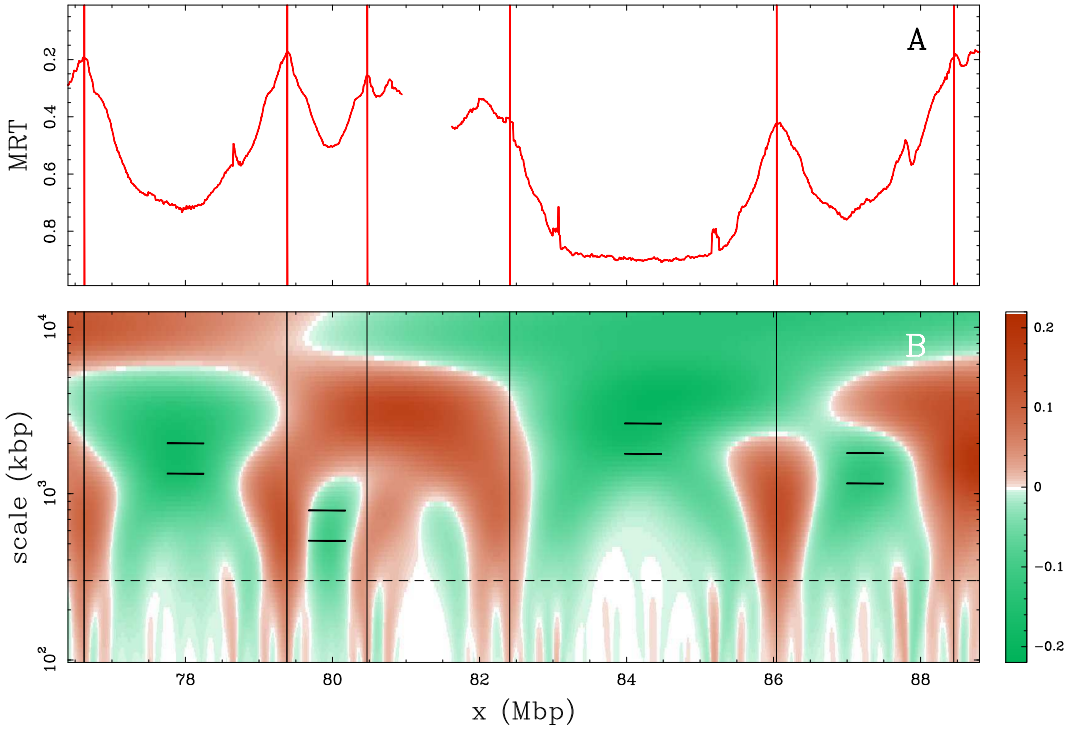

Figure 5: **Wavelet-based methodology to delineate U-shaped replication timing domains.** (A) MRT profile obtained in K562 cell line along a 11.4 Mbp long segment of human chromosome 10. (B) Space-scale representation of second-order variations for the MRT profile presented in (A); $T_{g^{(2)}}^{MRT}$ (Eq. (3)) values are color coded using green (resp. orange) shades for negative (resp. positive) curvature (note that MRT axis is going downwards). Horizontal dashed line marks scale 300 kbp used to detect regions of preferential replication initiation (vertical lines). Pairs of horizontal bars delineate the scale range where strong negative curvature is expected for parabolic U-shaped MRT profile. Regions delineated by two successive regions of preferential replication initiation are kept as U-domain if $T_{g^{(2)}}^{MRT} \leq -0.04$ at their midpoint for some scale value in this range.

V.2.3 Detection of U-domains along mean replication timing profiles

Within the approximation of constant fork velocity, the second derivative of MRT profiles is related to the average initiation site density minus the average termination site density (Chapter II, Eq. (69) in the summary). Here, we propose to segment MRT at points of maximal curvature *i.e.* regions that present on average more initiation than termination events.

This can be achieved using the continuous wavelet transform, which provides a powerful framework for the robust estimation of signal variations over any length scale (Mallat 1998; Arneodo et al. 2002b). The wavelet-transform (WT) is a space-

scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet, by means of dilations and translations. When using the derivatives of the Gaussian function, namely $g^{(n)}(x) = \mathrm{d}^n g^{(0)}(x)/\mathrm{d}x^n$, with $g^{(0)}(x) = e^{-x^2/2}$, then the WT of MRT profile takes the following expression:

$$\begin{aligned} T^{MRT}_{g^{(n)}}(x,a) &= \frac{1}{a}\int_{-\infty}^{+\infty} dy\, g^{(n)}\left(\frac{y-x}{a}\right) \mathrm{MRT}(y) \\ &= (-a)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(g^{(0)}_a * \mathrm{MRT}\right)(x), \end{aligned} \tag{3}$$

where x and a (> 0) are the space and scale parameters respectively. Eq. (3) shows that the WT computed with $g^{(n)}$ is proportional to the n^{th} derivative of the MRT profile smoothed by a dilated version $g^{(0)}_a(x) = \frac{1}{a}g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique (Mallat 1998; Arneodo et al. 2002b).

In the space-scale representation of replication timing second-order variations provided by $T^{MRT}_{g^{(2)}}$, we delineated loci that present a local maxima in the MRT curvature profile at scale 300 kbp (Fig. 5; $T^{MRT}_{g^{(2)}} \geq 0.02$) as loci of preferential replication initiation. In a second step, we characterized the regions encompassed between two preferential replication initiation loci using the MRT curvature at their mid-point (Fig. 5). We selected regions of length L with sufficiently negative values of $T^{MRT}_{g^{(2)}}$ (≤ -0.04; Eq. (3)) at some scale between $0.48L$ and $0.72L$ (for a parabolic shape profile of finite size L, the scale where extremal curvature is observed using the $T^{MRT}_{g^{(2)}}$ is proportional to L but also depends on the shape of the profile at the border of the region).

V.3 Chromatin state and long-range chromatin interactions in replication U-domains

V.3.1 Replication timing U-domains borders are enriched in open chromatin markers

Genome-wide investigation of chromatin architecture has revealed that, at large scales (from 100 kbp to 1 Mbp), regions enriched in open chromatin fibers correlate with regions of high gene density (Gilbert et al. 2004). Moreover there is a growing body of evidence that transcription factors are regulators of origin activation (reviewed in Kohzaki and Murakami 2005). We ask whether the remarkable genome organization observed around N-domain borders (Huvet et al. 2007) is maintained around replication timing U-domain borders and to what extent it is mediated by a particular chromatin structure favorable to early replication initiation (Audit et al. 2009).

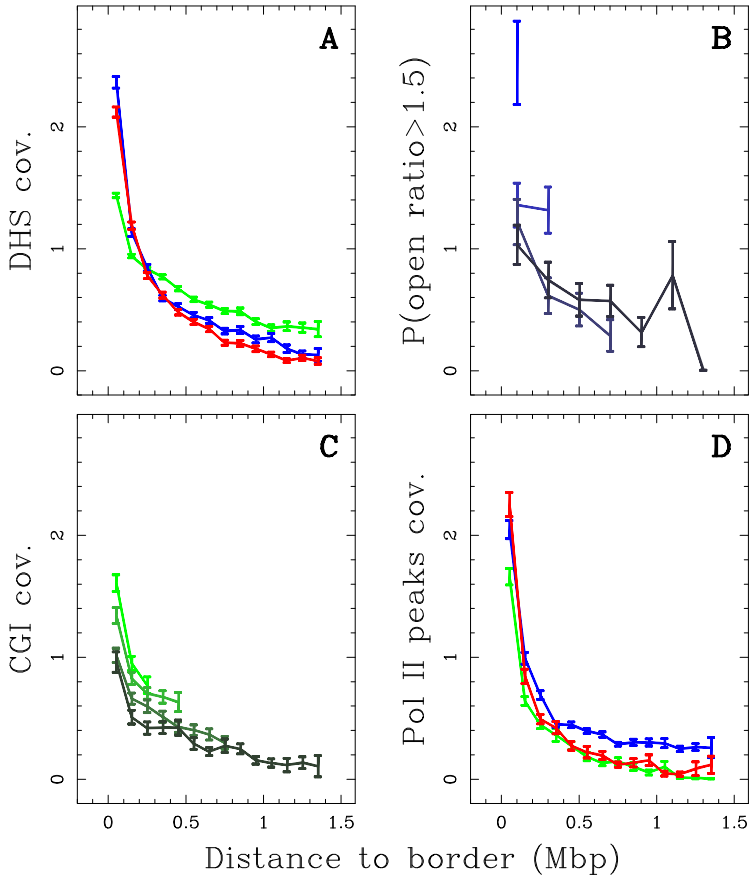

Figure 6: **Over representation of open chromatin markers at replication timing U-domain borders relative to the corresponding genome-wide value.** (A) Mean coverage by DNase I hypersensitive zones, as a function of the distance to the closest U-domain border in BG02 using DNase H1-hESC data (green, genome-wide mean value = 0.0073), K562 using DNase K562 data (red, genome-wide mean value = 0.0138), GM06990 using DNase GM06990 data (blue, genome-wide mean value = 0.0107). (B) Proportion of clones presenting a ratio of "open" over input chromatin greater than 1.5 versus the distance to the closest U-domain border in GM06990 for four U-domain size categories: L<0.8Mbp, 0.8Mb<L<1.2Mbp, 1.2Mb<L<1.8Mbp and 1.8Mb<L<3Mbp from light to dark blue curves (genome-wide mean value = 0.20). (C) Mean coverage by 1 kbp-enlarged CpG islands as a function of the distance to the closest U-domain border in BG02 for the four U-domain size categories defined in (B) from light to dark green curves (genome-wide mean value = 0.0254). (D) Mean coverage by Pol II peaks as a function of the distance to the closest U-domain border in BG02 (green: Pol II in H1 ESC, genome-wide mean value = 0.0026), K562 (red: Pol II in K562, genome-wide mean value = 0.0024), GM06990 (blue: Pol II in GM12878, genome-wide mean value = 0.0097).

When mapping DNase I sensitivity data (Section V.4) (Sabo et al. 2006) on the U-domains, we observed that the mean coverage is maximal at U-domain extremities and decreases significantly from the extremities to the center that is rather insensitive to DNase I cleavage (Fig. 6A). This decrease, from values significantly higher than the genome-wide average value, extends over ~ 150 kbp, whatever the size of the replication timing U-domain (Fig. 7A-C) suggesting that, for all examined cell lines, early replicating U-domains borders are at the center of ~ 300 kbp wide open chromatin regions. We observed a significant anti-correlation between DNase I cleavage sensitivity data and replication timing data in BG02 (DNase H1-hESC: $R = -0.55$, $P < 10^{-16}$), K562 ($R = -0.63$, $P < 10^{-16}$) and GM06990 ($R = -0.57$, $P < 10^{-16}$) cell lines as well as in the other four cell lines (data not shown; note that this was still observed when controlling for the GC content). This is further supported by open over input chromatin ratio data obtained from human lymphoblastoid cells (Gilbert et al. 2004). We observed that the regions presenting an open/input ratio > 1.5 also decreased significantly (3-fold) from U-domain borders to centers (Fig. 6B).

Cytosine DNA methylation is a mediator of gene silencing in repressed heterochromatic regions, while in potentially active open chromatin regions, DNA is essentially unmethylated (Suzuki and Bird 2008). DNA methylation is continuously distributed over mammalian chromosomes with the notable exception of CpG islands (CGIs) and in turn of certain CpG rich promoters and transcription start sites (TSSs). Along the observation that the hypomethylation level of CGIs extends to about 1 kbp in flanking regions, we used 1 kbp-enlarged CGI coverage as an hypomethylation marker (Section V.4) (Audit et al. 2009). When averaging over the U-domains detected in BG02, we robustly observed a maximum of CGI coverage at U-domain borders as the signature of hypomethylation and a decrease over a characteristic distance of ~ 150 kbp (Fig. 6C), similar to what found for DNase I sensitivity coverage (Fig. 6A). This contrasts with the GC-content profile that strongly depends on the U-domain size and decreases very slowly toward the U-domain center without exhibiting any characteristic scale (Fig. 7D-F). These observations are consistent with the hypothesis that early replication origins at U-domains borders are associated to CGIs possibly protected from methylation due to the colocalization with replication origins (Antequera and Bird 1999).

Open chromatin markers have been associated with genes. For example 16% of all DNase I hypersensitive sites (HS) are in the first exon or at the TSS of a gene and 42% are found inside a gene (Boyle et al. 2008a). Also, more than 90% of broadly expressed housekeeping genes have a CpG-rich promoter (Ponger et al. 2001). Remarkably, the mean profiles of Pol II binding Chip-Seq tag density (Section V.4) along U-domains detected in BG02, K562 and GM06990 cell lines strongly decay over ~ 150 kbp away from U-domain borders (Fig. 6D). This indicates that, whatever the cell line, the open chromatin regions around replication U-domains

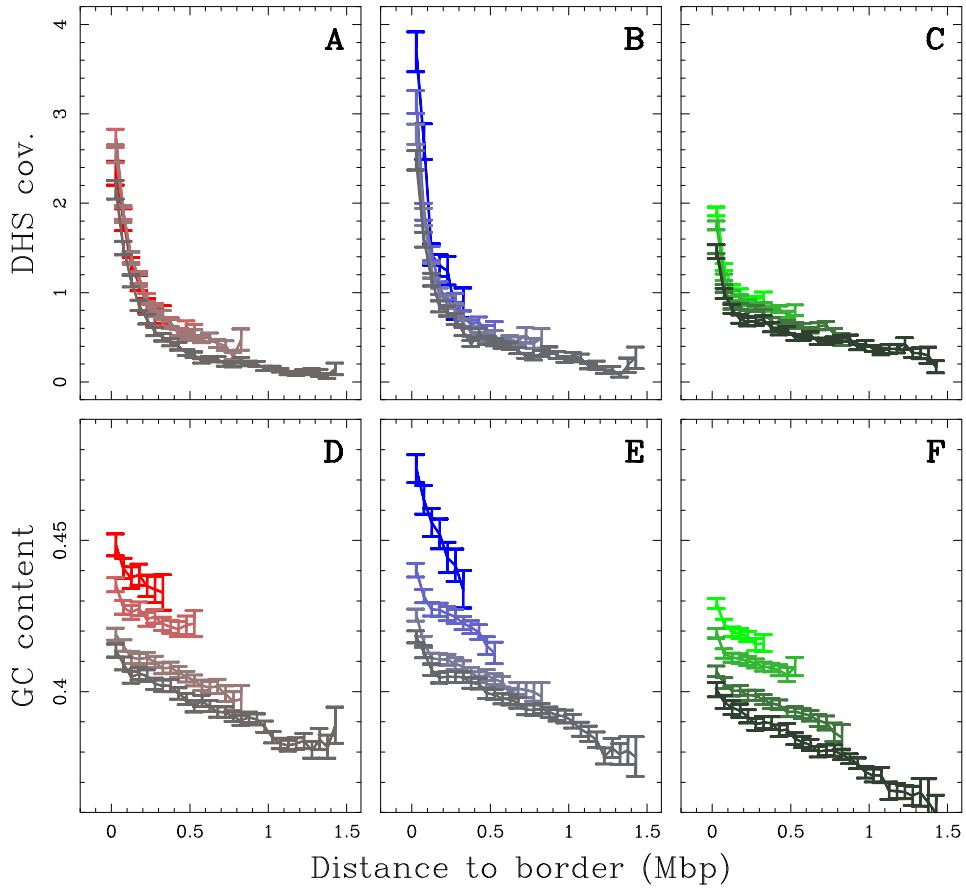

Figure 7: **Comparison of DHS coverage and GC content gradients along U-domains.** Mean coverage (relative to the genome average) of DNase I hypersensitive zones (A-C) and GC content (D-F) as a function of the distance to the closest U-domain border in K562 (A,D), GM06990 (B,E) and BG02 (C,F), for four U-domain size categories: L<0.8Mbp, 0.8Mb<L<1.2Mbp, 1.2Mb<L<1.8Mbp and 1.8Mb<L<3Mbp from light to dark curves.

are prone to transcription whereas U-domain central regions appear, on average, transcriptionally silent.

V.3.2 Replication timing U-domains are insulated compartments of genome-wide chromatin interactions (Hi-C)

It has early been recognized that the 3D chromatin tertiary structure provides some understanding to the experimental observation of the so-called replicon and replication foci (Buongiorno-Nardelli et al. 1982; Berezney et al. 2000). In particular, replicon size, which is dictated by the spacing of between active origins, has been shown to correlate with the length of chromatin loops (Buongiorno-Nardelli et al. 1982; Conti et al. 2007; Courbet et al. 2008). Very recently, the outstanding progress made in chromosome conformation capture technique (Lieberman-Aiden et al. 2009) has provided access to long-range chromatin interactions across the entire genome as a footprint of the different levels of chromatin folding in relation with gene activity and the functional state of the cell. From a comparative analysis of replication timing data and Hi-C data in the human genome, some dichotomic picture has been proposed where early and late replicating loci occur in separated compartments of open and closed chromatin respectively (Lieberman-Aiden et al. 2009; Ryba et al. 2010). More precisely, each chromosome has been consistently partitioned into two compartments, where the interaction profiles over the whole chromosome correlate for loci belonging to the same compartment but anticorrelates for loci belonging to separated compartments (Lieberman-Aiden et al. 2009). These two compartments very significantly overlap early and late replicating domains respectively (Ryba et al. 2010). Here, instead of considering this partitioning derived from the positive or negative correlations between interaction profiles over the whole chromosome, we focused on interactions between loci separated by short genomic distances ($\lesssim$ 10 Mbp) over which the contact probabilities are the highest (Lieberman-Aiden et al. 2009).

First, we performed this zoom in the Hi-C contact matrix in the K562 cell line at the 100 kbp resolution (Section V.4) for the 11.4 Mbp fragment of human chromosome 10 which contains four U-domains in K562 (Fig. 1; $[O_1, O_2]$, $[O_2, O_3]$, $[O_4, O_6]$ and $[O_6, O_7]$). We found that these four U-domains remarkably correspond to four matrix square-blocks of enriched interactions (Fig. 8A). We recovered that early replicating zones, when bordering a U-domain (*e.g.* O_4 and O_6 separated by 3.9 Mbp), have a high contact probability as the signature of 3D spatial proximity. However, we also observed a high contact probability of the two early replicating borders with the late replicating U-domain center and interactions appear sparse for loci in separate U-domains (*e.g.* O_1 and O_3 separated by 3.6 Mbp). Further examination of the average behavior of intrachromosomal contact probability as a function of genomic distance for the complete genome corroborated these observations. We found that the mean number of interactions between two 100 kbp loci of the same U-domain decays when increasing their distance as observed genome-wide (Fig. 8B). But importantly the mean number of pairwise interactions is significantly higher in-

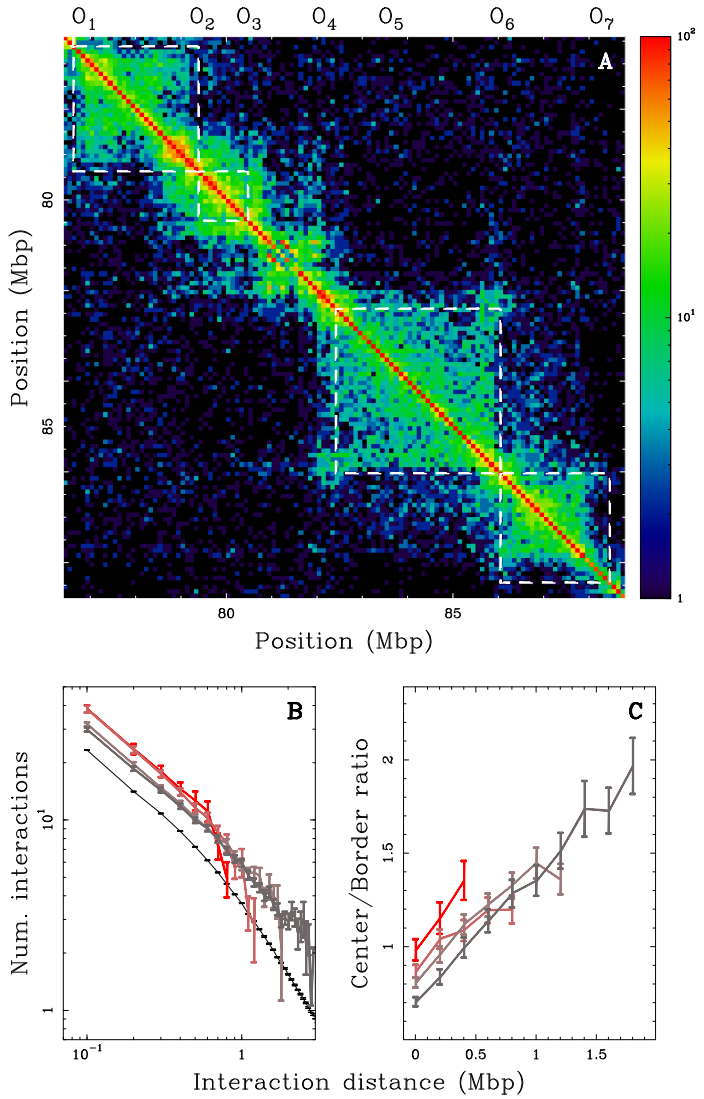

Figure 8: **Replication U-domains are self-interacting structural units in K562 erythroid cell line.** (A) Hi-C proximity matrix corresponding to intrachromosome interactions on the 11.4 Mbp long fragment of human chromosome 10 (Fig. 1), as measured in the K562 cell line. Each pixel represents all interactions between a 100 kbp locus and another 100 kbp locus; intensity corresponding to the total number of reads is color coded according to the colormap (right). The dashed squares correspond to replication timing U-domains detected in the K562 cell line. (B) Number of interactions between two 100 kbp loci versus the distance separating them (logarithmic scales) as computed genome wide (black) or in K562 replication U-domains only, for four U-domain size categories: L<0.8Mbp, 0.8Mb<L<1.2Mbp, 1.2Mb<L<1.8Mbp and 1.8Mb<L<3Mbp (from light to dark red). (C) Ratio of the number of interactions between two 100 kbp loci inside the same U-domain at equal distance from its center and the number of interactions between loci on opposite sides and equal distance from a U-domain border, versus the distance between them; colors as in (B).

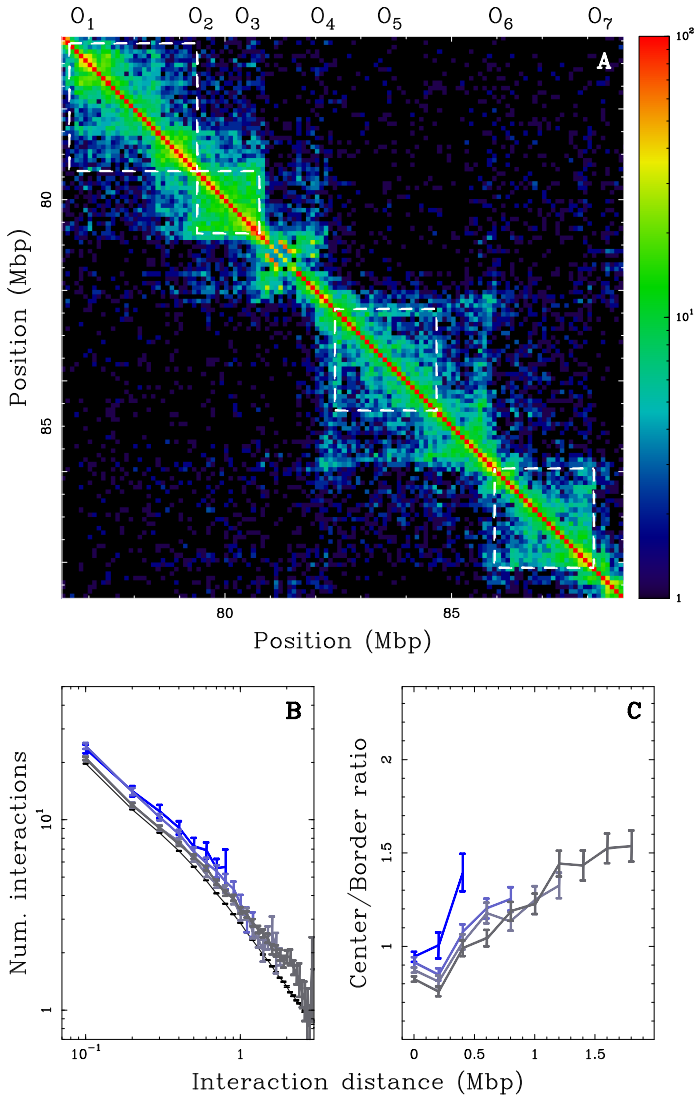

Figure 9: **Replication U-domains are self-interacting structural units in GM06990 lymphoblastoid cell line.** Same as in Fig. 8 but for the GM06990 lymphoblastoid cell line.

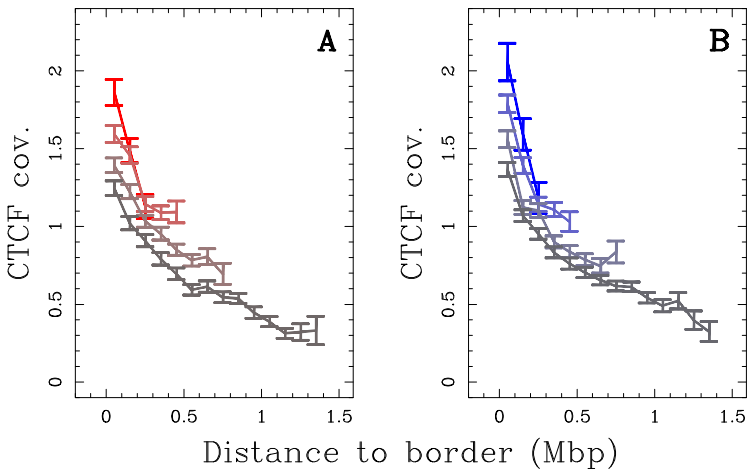

Figure 10: **Enrichment in insulator-binding protein CTCF at replication U-domains borders.** (A) Mean coverage by CTCF enriched signals versus the distance to the closest U-domain border in K562 cell line for four U-domain size categories: L<0.8Mbp, 0.8Mb<L<1.2Mbp, 1.2Mb<L<1.8Mbp and 1.8Mb<L<3Mbp, from light to dark red curves (genome-wide mean value = 0.0051). (B) Same as in (A) but for the GM06990 cell line (blue code shades) (genome-wide mean value = 0.0046).

side the U-domains than genome-wide and this seems to depend on the U-domain length, smaller the domain, higher the mean number of interactions probably as the signature of a more open chromatin structure. When comparing the contact probability between two loci inside a U-domain or lying in neighboring U-domains (Fig. 8C), we observed that the latter is higher than the former for distances smaller than the characteristic size ($\sim$ 300 kbp) of the open chromatin structure at U-domain borders (Fig. 6). Above this characteristic distance, the tendency is reversed and the ratio increases up to 2 for distances $\sim$ 1.8 Mbp (Fig. 8C). These data suggest that the segmentation of the genome into replication timing U-domains corresponds to some spatial compartmentalization into self-interacting structural chromatin units insulated by two boundaries of open, accessible, actively transcribed chromatin. This conclusion is strengthened by the observation that U-domain borders are significantly enriched in the insulator binding protein CTCF (Fig. 10), that is known to be involved in chromatin loop formation conditioning communication between transcriptional regulatory elements (Phillips and Corces 2009; Hou et al. 2010; Ohlsson et al. 2010; Handoko et al. 2011). Quantitatively similar results were obtained for the lymphoblastoid GM06990 cell line for which both replication timing and Hi-C data were available (Fig. 9).

V.4 Material and Methods

Determining mean replication timing profiles.
We determined the mean replication timing profiles along the complete human genome using Repli-Seq data (Hansen et al., 2010; Chen et al., 2010). This method consists in labeling newly synthesized DNA using a pulse of BrdU, sorting cells into several S-phase fractions using FACS and to reveal the locus of DNA synthesis in each fraction using anti-BrdU antibody combined to next-generation sequencing. For embryonic stem cells (BG02), three lymphoblastoid cell lines (GM06990, H0287, TL010) a fibroblast cell line (BJ, replicate R1) and erythroid K562 cell line, Repli-Seq tags for 6 FACS fractions were downloaded from the NCBI SRA website (Studies accession: SPR0013933) (Hansen et al., 2010). For a given cell line and for each S-phase fraction, we computed the tag densities in 100 kb windows, and following the authors (Hansen et al., 2010) the tag densities were normalized to the same genome-wide sequence tag counts for each fraction, and a second normalization was performed so that at each genomic position, the sum over S-phase fractions be one. To filter out noise which could critically bias mean timing profile estimate (Fig. 11A), we proceeded as follows. We noticed that the genome-wide distribution of the normalized tag density (Fig. 11D) presents a mode at $0.01 < m < 0.08$ (mainly noise) and a long tail up to 1 (mainly corresponding to the replication signal). For each S-phase fraction we set to 0 the normalized tag density $< 4m$, and re-normalized at each genomic position by the sum over S-phase fractions. The mean replication timing profile computed on these denoised tag densities superimposes on the original one, but is much less noisy (Fig. 11B,C).

For the HeLa cell line, the denoised tag densities were obtained from (Chen et al., 2010). Instead of computing the S50 (median replication timing) as the authors in (Chen et al., 2010), we computed the mean replication timing (MRT).

Sequence and annotation data. Sequence and annotation data were retrieved from the Genome Browsers of the University of California Santa Cruz (UCSC) (Karolchik et al. 2003). Analyses were performed using the human genome assembly of March 2006 (NCBI36 or hg18). As human gene coordinates, we used the UCSC Known Genes table. When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates, resulting in 23818 distinct genes. We used CpG islands (CGIs) annotation provided in UCSC table "cpgIslandExt".

Replication N-domains. The coordinates of the 678 human replication N-domains for assembly NCBI35/hg17 were obtained in (Huvet et al. 2007) and mapped using LiftOver to hg18 coordinates; we kept only the 663 N-domains that had the same size after conversion.

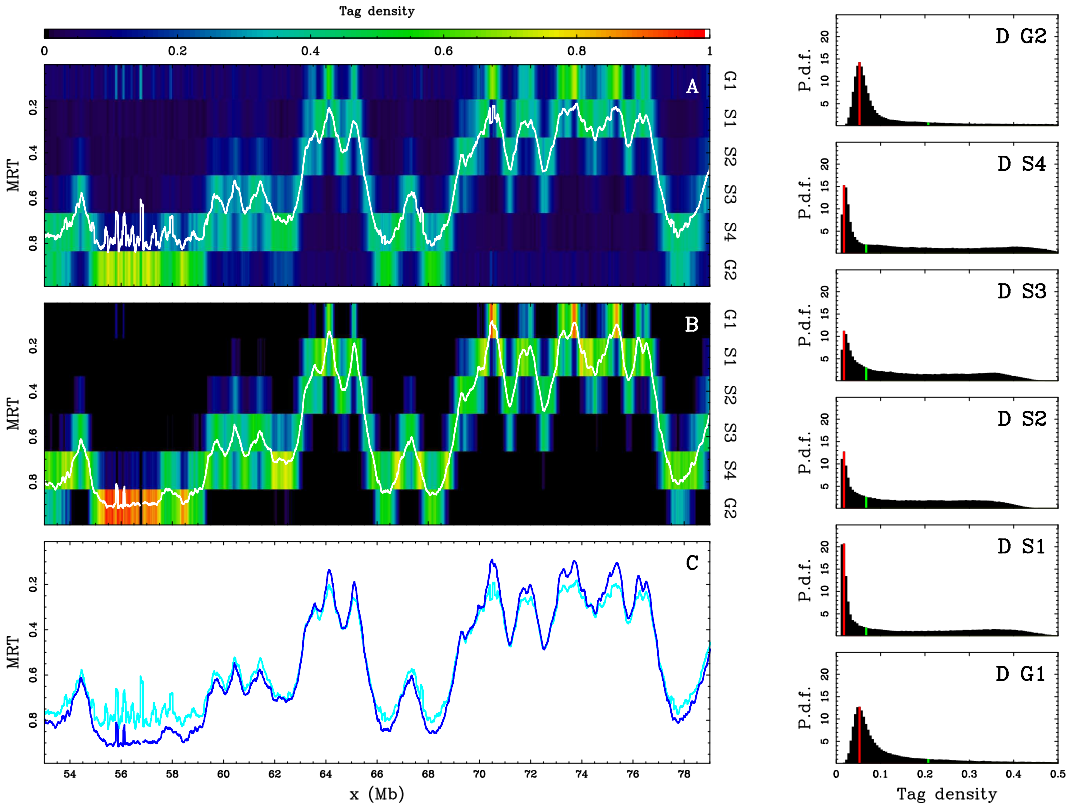

Figure 11: **Determining mean replication timing profiles from Repli-Seq data.** (A) Normalized tag densities on a 25 Mb long fragment of chromosome 10, for the GM06990 cell line, and the corresponding computed MRT (white line). (B) "Denoised" normalized tag densities on the same genomic fragment and the corresponding MRT (white line). In (A) and (B) the tag densities for each S-phase fraction (G1-G2) are color coded using the color map situed at the top. (C) Comparison on the same genomic fragment of the MRT computed on the normalized tag densities (cyan line) and the MRT computed on the "denoised" normalized tag densities (blue line). (D) Probability density function (P.d.f.) of the genome-wide distribution of the normalized tag densities for each S-phase fraction from G1 to G2 from bottom to top (black histogram). The mode m of the distribution is given by the red bar, the threshold $4m$ used for denoising is given by the green bar.

Matchings of replication timing U-domains expected by chance. We randomly re-positioned replication timing U-domains in all cell lines including skew N-domains in the germline, conserving the statistics of domains size and inter-domain distance. We then computed for each cell line pair the number of matchings (1000 simulations were used to obtain the mean values), and the percentage of matchings as in Table 4. None of the 1000 simulations gave number of matchings as important as the one observed, so we concluded that the matchings observed had a p-value $P < 10^{-3}$.

DNase I hypersensitive site data. We used the DNaseI sensitivity measured genome-wide (Sabo et al. 2006). Data corresponding to Release 3 (Jan 2010) of the ENCODE UW DNaseI HS track, were downloaded from the UCSC FTP site:
`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/`.
We plotted the coverage by DNase Hypersentive Sites (DHSs) identified as signal peaks at a false discovery rate threshold of 0.5% within hypersensitive zones delineated using the HotSpot algorithm ("wgEncodeUwDnaseSeqPeaks" tables). When several replicates were available, data were merged.

Genome-wide maps of Pol II and CTCF binding. We used ChIP-seq data using antibody for Pol II and CTCF from Release 3 (Mar 2010) of the ENCODE Open Chromatin track (Bhinge et al. 2007; The ENCODE Project Consortium 2007). Data were downloaded from the UCSC FTP site:
`ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap`.
We plotted coverage by regions of enriched signal in ChIP experiments, called based on signals created using F-Seq (Boyle et al. 2008b) ("wgEncodeUtaChIPseqPeaks" tables). Significant regions were determined at an approximately 95% sensitivity level. We always used the most recent version of data.

Whole genome chromatin conformation data. We used the spatial proximity maps of the human genome generated using Hi-C method (Lieberman-Aiden et al. 2009). We downloaded 100 kbp resolution maps for GM06990 and K562 cell lines from the GEO web site (GSE18199_binned_heatmaps):
`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18199`.

Chromatin fiber density data. Open over input chromatin ratio data from human lymphoblastoid cells were obtained from the authors (Gilbert et al. 2004).

Summary of Chapter V

In several human cell lines, about half of the human genome is covered by large-scale ($\sim$ 1 Mbp) replication domains characterized by a U-shaped replication timing profile (U-domains). Interestingly, a majority of replication U-domains are cell line specific, and thus belong to genomic regions of high replication timing plasticity. The mapping of open chromatin marks along U-domains revealed that they are bordered by early replication initiation zones likely specified by a $\sim$ 300 kbp wide region of accessible, open chromatin permissive to transcription (Fig. 6). Long-range chromatin interaction data (Hi-C) suggest that U-domains correspond to chromatin self-interacting structural units. Replication U-domains indeed remarkably correspond to square block of enriched interaction in the Hi-C contact matrix (Figs. 8A and 9A). Loci interact more inside a U-domains than loci separated by a U-domain border (Figs. 8C and 9C), as if the border was insulating the domain, consistently with the strong enrichment of insulator binding protein CTCF at U-domain borders (Fig. 10).

Chapter VI

Conclusion and perspectives

VI.1 Establishment of the compositional asymmetry

Compositional asymmetry has long been proposed to result from the asymmetric mutational pressure generated by the transcription and replication processes. The originality of this work was to formalize this neutral molecular evolution scenario, and to propose a theoretical framework to discuss the compositional asymmetry dynamic and its relation to the spatio-temporal program of DNA replication.

A minimal model for the compositional asymmetry evolution

In the minimal model proposed in Chapter I, both the substitutional and compositional asymmetries decompose into transcription and replication associated components, the latter being proportional to the replication fork polarity. Most studies of replication-associated asymmetry focused on the relationship between strand asymmetry and replication origins, in direct analogy with the bacterial replicon case, sometimes leading to inconclusive and even contradictory results. In Chapter II, we proved under quite reasonable assumptions (constant replication fork velocity and bidirectional replication origins) that the replication fork polarity is proportional to the derivative of the mean replication timing, whatever the complexity of the spatio-temporal replication program. Thanks to this relationship, we succeeded in estimating experimentally the replication fork polarity in the human genome, using recently available replication timing data. In Chapter III, we were then able to show that the substitutional and compositional asymmetries observed in the human genome were consistent with the minimal model proposed theoretically in Chapter I.

Study of replication-associted strand asymmetry in eukaryotes

As experimental replication timing data are now becoming available in an increasing number of organisms, the relationship between replication fork polarity and replication timing opens new perspectives regarding the study of replication-associated strand asymmetry in eukaryotic genomes. It would be for instance possible to extend, on a genome-wide scale, previous analyses of strand asymmetry made in the

sub-telomeric regions of yeast (Gierlik et al. 2000). As the mutational spectra of the DNA polymerases has been investigated experimentally in yeast (Pursell et al. 2007; Larrea et al. 2010), it would be in principle possible to test if the substitutional asymmetries associated to replication can be explained by the different error spectra of the leading and lagging DNA polymerases.

Conservation of the replication timing profile

The comparative analysis of replication timing and compositional asymmetry in the eukaryotic kingdom will probably shed a new light on the conservation of the replication program across evolution. Indeed the establishment of a replication-associated compositional asymmetry not only requires a molecular mechanism that generates a different mutational pattern on the leading and lagging strands, but further requires the stability of the replication fork polarity profile on evolutionary time-scales. The current values of the compositional asymmetries and substitution rates observed in the human genome suggest that the compositional skew has been generated over 400 Myrs, a time-scale comparable to the last common ancestor of amniotes ($\sim$ 350 Myrs). Interestingly, skew N-domains are observed in all amniotic genomes (Claude Thermes, personal communication). These observations suggest that the replication timing profile has been well conserved in amniotes, despite considerable chromosomal rearrangements. Consistently, evolutionary breakpoints in N-domains preferentially occur at the borders (Lemaitre et al. 2009; Zaghloul 2009), thus likely preserving the overall N-shaped skew profile.

Do all skew N-domains correspond to germline replication U-domains?

In this thesis, N-domains were proposed to result from a U-shaped mean replication timing profile in the germline. We could not directly test this hypothesis, as no germline replication timing data is available today. It would be very interesting to know how many skew N-domains are actually germline replication U-domains. What are the characteristics of skew N-domains that are not U-domains? Were they generated by another mechanism, such as non-coding transcription as proposed in (Necsulea et al. 2009)? Were they U-domains in an ancestral replication timing profile that have persisted as skew N-domains (as it takes several hundred Myrs to erase an initially non null skew)?

VI.2 Which spatio-temporal replication program for the replication U-domains?

As proposed in this thesis, the N-shaped skew profile in N-domains may result from a U-shaped mean replication timing profile. A U-shaped replication timing profile is not specific to the germline but is robustly observed in several human cell lines (Chapter V). However it doesn't tell where and when replication initiates in the N-domains and U-domains, which would completely specify the spatio-temporal

replication program. Actually several replication programs, with drastically different distributions of initiation sites, can yield the same mean replication timing.

Determining the firing times and initiation sites from the replication kinetics

In Chapter II, we derived the "analytical inversion of the KJMA model" which could have very promising applications. This result allows, at least theoretically, to infer the distribution of initiation sites and firing times directly from the replicated fractions obtained by time-course micro-array or Repli-Seq experiments. Unfortunately, the current time resolution ($\sim$ 2h) of replicated fractions in the human genome is too coarse to allow the application of this result. Hopefully, it would be possible in the next future to apply this procedure to better-resolved experimental data. Of course, it would be then very interesting to compare the distribution of initiation sites predicted from the replicated fractions and the distributions obtained experimentally by other approaches, for instance DNA combing and nascent-strand studies. Maybe the combination of all these methods will permit to decipher the spatio-temporal replication program in human cell lines, in particular within U-domains.

Previous model: replication origins located at the borders

We recall the first model proposed to explain the N-domains (Brodie of Brodie et al. 2005; Touchon et al. 2005): the skew upward jumps at the N-domains borders correspond to replication origins as in bacteria, but the absence of downward jumps (associated to the replication terminus in bacteria) likely reflects the randomness of the termination site. We note that the variable firing times of the two bordering replication origins lead inevitably to greatly dispersed termination sites over cell cycles, which provides a very simple mechanism accounting for the random termination site postulated in (Brodie of Brodie et al. 2005; Touchon et al. 2005). More generally in each model of DNA replication program with well-positioned replication origins (with variable efficiencies and firing times), as it seems to be the case in yeast, we expect a U-shaped replication timing profile, where timing peaks correspond to replication origins but where the converse is not necessarily true (Chapter II). However we already know that this model is certainly wrong for the N-domains and U-domains in the human genome, especially for the larger ones. Indeed this model would imply too large inter-origin distances (several Mbp), as compared to the typical inter-origin distances (40 kbp) observed in DNA combing experiments (Rappailles et al. 2011). We foresee that other initiation events could be observed in the central regions of N-domains and U-domains.

The domino model

The "analytical inversion of the KJMA model" relies on an important assumption: the independent firing of replication origins. Our collaborators, Olivier Hyrien and Arach Goldar, have proposed a "domino model" for the replication program in U-

domains which explicitly breaks this assumption. In this model, a moving fork is supposed to stimulate the initiation in nearby unreplicated DNA. As each initiation event is associated to the propagation of two diverging replication forks, this results in the sequential activation of replication origins in a domino-like fashion (Rappailles et al. 2011). In U-domains, replication first initiates at the borders as they are local minima of the replication timing. From the borders, two replication waves of secondary initiations would then propagate until complete replication of the U-domain. If the "master" replication origins at the border (that fire first) have variable firing times, or if the stimulated initiations are more and more synchronous as S phase progresses, or if some rare additional initiation events (not associated to the replication wave of secondary initiations) take place within the domain, then it can easily lead to a U-shaped mean replication timing profile.

VI.3 Genome 3D structure and replication timing

How the chromatin fiber folds into higher-order structures within the cell nucleus is arguably one of the most important open problem in cell biology. The genome 3D structure is presumably tightly related to the chromatin state and to the organization of many cellular processes, including transcription and replication. The recent development of the Hi-C technique offers the unprecedented possibility to measure long-range chromatin interactions on a genome-wide scale. The comparative analysis of the replication timing and the Hi-C contact matrix revealed qualitatively two different regimes. Replication U-domains, especially when they are large, often correspond to a square like block of enriched interactions, as if they were self-interacting. This correspondence is particularly impressive on U-domains larger than 3 Mbp[1], which contain in their central regions heterochromatic and late replicating gene deserts. Besides replication U-domains, the human genome is also covered by large (several Mbp) early replicating domains, which overlap in part the GC-rich isochores, are characterized by an open chromatin state and a high transcriptional activity. In the early replicating domains, a locus interacts a lot with neighboring loci on a Mbp characteristic distance.

The relationship between replication timing and long-range chromatin interactions is under current investigation at Laboratoire Joliot-Curie, both experimentally and bioinformatically. On the bioinformatical side, in order to study quantitatively this relationship on a genome-wide scale, it will be necessary to objectively delineate square-like blocks in the Hi-C contact matrix and to measure the overlap with replication U-domains. The square-like Hi-C blocks have different sizes, and are sometimes organized hierarchically (a square-like block can be made of smaller

[1]Actually, U-domains larger than 3 Mbp were not retained in the analysis of Chapter V. These U-domains significantly overlap the so-called split-N domains, skew domains systematically detected and studied in (Zaghloul 2009; Arneodo et al. 2011). Split-N domains contain a central region of null skew, which corresponds to a heterochromatic gene desert.

square-like blocks). The wavelet transform, which can perform a pattern recognition at multiple scales, and previously used as such for the detection of skew N-domains and replication U-domains, seems equally well adapted for this new task.

Bibliography

Aladjem MI. 2007. Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nat Rev Genet* **8**: 588–600.

Alberts B, Jonhson A, Lewis J, Raff M, Roberts K, Walter P. 2008. *Molecular Biology of the Cell, 5th ed.* Garland Publishing, New York.

Anglana M, Apiou F, Bensimon A, Debatisse M. 2003. Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114**: 385–394.

Antequera F, Bird A. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr Biol* **9**: R661–R667.

Arndt PF, Burge CB, Hwa T. 2003. DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol* **10**: 313–322.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**: 2322–2328.

Arneodo A, Argoul F, Bacry E, Elezgaray J, Freysz E, Grasseau G, Muzy JF, Pouligny B. 1992a. *Wavelets and Applications*, chapter Wavelet transform of fractals, pages 286 – 352. Springer, Berlin.

Arneodo A, Argoul F, Bacry E, Elezgaray J, Muzy JF. 1995a. *Ondelettes Multifractales et Turbulences : de l'ADN aux croissances cristallines.* Diderot Editeur, Arts et Sciences, Paris.

Arneodo A, Argoul F, Bacry E, Muzy JF, Tabard M. 1992b. Golden mean arithmetic in the fractal branching of diffusion-limited aggregates. *Phys Rev Lett* **68**: 3456–3459.

Arneodo A, Argoul F, Elezgaray J, Grasseau G. 1989. Wavelet transform analysis of fractals: application to nonequilibrium phase transitions. In G Turchetti, ed., *Nonlinear Dynamics*, pages 130–180. World Scientific, Singapore.

Arneodo A, Argoul F, Muzy JF, Tabard M. 1992c. Structural 5-fold symmetry in the fractal morphology of diffusion-limited aggregates. *Physica A* **188**: 217–242.

Arneodo A, Argoul F, Muzy JF, Tabard M. 1992d. Uncovering Fibonacci sequences in the fractal morphology of diffusion-limited aggregates. *Phys Lett A* **171**: 31–36.

Arneodo A, Audit B, Brodie of Brodie EB, Nicolay S, Touchon M, d'Aubenton-Carafa Y, Huvet M, Thermes C. 2009. Fractals and wavelets : what can we learn on transcription and replication from wavelet-based multifractal analysis of DNA sequences ? *Encyclopedia of Complexity and System Science* **3893**.

Arneodo A, Audit B, Decoster N, Muzy JF, Vaillant C. 2002a. *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*, chapter Wavelet based multifractal formalism: Application to DNA sequences, satellite images of the cloud structure and stock market data, pages 26–102. Springer Verlag, Berlin.

Arneodo A, Audit B, Decoster N, Muzy JF, Vaillant C. 2002b. Wavelet based multifractal formalism: Application to DNA sequences, satellite images of the cloud structure and stock market data. In A Bunde, J Kropp, HJ Schellnhuber, eds., *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes*, pages 26–102. Springer Verlag, Berlin.

Arneodo A, Audit B, Faivre-Moskalenko C, Moukhtar J, Vaillant C, Argoul F, d'Aubenton-Carafa Y, Thermes C. 2008. *From DNA sequence to chromatin organization : the fundamental role of genomic longe-range correlations.* Bulletin de l'Académie Royale de Belgique, Mémoire de la Classe des Sciences, Collection 8, 3 série, Tome XXVIII, n 2049.

Arneodo A, Bacry E, Graves PV, Muzy JF. 1995b. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys Rev Lett* **74**: 3293–3296.

Arneodo A, Bacry E, Muzy JF. 1995c. The thermodynamics of fractals revisited with wavelets. *Physica A* **213**: 232–275.

Arneodo A, d'Aubenton-Carafa Y, Audit B, Bacry E, Muzy JF, Thermes C. 1998a. Nucleotide composition effects on the long-range correlations in human genes. *Eur Phys J B* **1**: 259–263.

Arneodo A, d'Aubenton-Carafa Y, Audit B, Brodie of Brodie EB, Nicolay S, St-Jean P, Thermes C, Touchon M, Vaillant C. 2007. DNA in chromatin: from genome-wide sequence analysis to the modeling of replication in mammals. *Advances in Chemical Physics* **135**: 203–252.

Arneodo A, d'Aubenton-Carafa Y, Bacry E, Graves PV, Muzy JF, Thermes C. 1996. Wavelet based fractal analysis of DNA sequences. *Physica D* **96**: 291–320.

Arneodo A, Decoster N, Kestener P, Roux SG. 2003. A wavelet-based method for multifractal image analysis: From theoretical concepts to experimental applications. *Adv Imaging Electr Phys* **126**: 1–92.

Arneodo A, Decoster N, Roux SG. 1999a. Intermittency, log-normal statistics, and multifractal cascade process in high-resolution satellite images of cloud structure. *Phys Rev Lett* **83**: 1255–1258.

Arneodo A, Decoster N, Roux SG. 2000. A wavelet-based method for multifractal image analysis. I. Methodology and test applications on isotropic and anisotropic random rough surfaces. *Eur Phys J B* **15**: 567–600.

Arneodo A, Grasseau G, Holschneider M. 1988. Wavelet transform of multifractals. *Phys Rev Lett* **61**: 2281–2284.

Arneodo A, Manneville S, Muzy JF. 1998b. Towards log-normal statistics in high reynolds number turbulence. *Eur Phys J B* **1**: 129–140.

Arneodo A, Manneville S, Muzy JF, Roux SG. 1999b. Revealing a lognormal cascading process in turbulent velocity statistics with wavelet analysis. *Phil Trans R Soc Lond A* **357**: 2415–2438.

Arneodo A, Muzy JF, Sornette D. 1998c. "Direct" causal cascade in the stock market. *Eur Phys J B* **2**: 277–282.

Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton Carafa Y, Thermes C. 2011. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys Rep* **498**: 45–188.

Arrault J, Arneodo A, Davis A, Marshak A. 1997. Wavelet based multifractal analysis of rough surfaces: Application to cloud models and satellite data. *Phys Rev Lett* **79**: 75–78.

Audit B, Nicolay S, Huvet M, Touchon M, d'Aubenton Carafa Y, Thermes C, Arneodo A. 2007. DNA replication timing data corroborate in silico human replication origin predictions. *Phys Rev Lett* **99**: 248102.

Audit B, Thermes C, Vaillant C, d'Aubenton Carafa Y, Muzy JF, Arneodo A. 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys Rev Lett* **86**: 2471–2474.

Audit B, Vaillant C, Arneodo A, d'Aubenton Carafa Y, Thermes C. 2002. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J Mol Biol* **316**: 903–918.

Audit B, Zaghloul L, Vaillant C, Chevereau G, d'Aubenton Carafa Y, Thermes C, Arneodo A. 2009. Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res* **37**: 6064–6075.

Avrami M. 1939. Kinectics of phase change I. General theory. *J Chem Phys* **7**: 1103–1112.

Avrami M. 1940. Kinectics of phase change II. Transformation-time relations for random distribution of nuclei. *J Chem Phys* **8**: 212–224.

Avrami M. 1941. Kinetics of phase change III. Granulation, phase change, and microstucture. *J Chem Phys* **9**: 177–184.

Azbel' MY. 1995. Universality in a DNA statistical structure. *Phys Rev Lett* **75**: 168–171.

Bacry E, Muzy JF, Arneodo A. 1993. Singularity spectrum of fractal signals from wavelet analysis: exact results. *J Stat Phys* **70**: 635–674.

Baker A, Audit B, Chen C, Moindrot B, Leleu A, Guilbaud G, Rappailles A, Vaillant C, Goldar A, Mongelard F, et al. 2011. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *Nucleic Acids Res* **submitted**.

Baker A, Nicolay S, Zaghloul L, d'Aubenton-Carafa Y, Thermes C, Audit B, Arneodo A. 2010. Wavelet-based method to disentangle transcription- and replication- associated strand asymmetries in mammalian genomes. *Appl Comput Harmon Annal* **28**: 150–170.

Bechhoefer J, Marshall B. 2007. How Xenopus laevis replicates DNA reliably even though its origins of replication are located and initiated stochastically. *Phys Rev Lett* **98**: 098105.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proc Natl Acad Sci USA* **93**: 13919–13924.

Beletskii A, Bhagwat AS. 1998. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. *Biol Chem* **379**: 549–551.

Bell SP, Dutta A. 2002. DNA replication in eukaryotic cells. *Annu Rev Biochem* **71**: 333–374.

Belmont AS. 2001. Visualizing chromosome dynamics with GFP. *Trends Cell Biol* **11**: 250–257.

Bensimon A, Simon A, Chiffaudel A, Croquette V, Heslot F, Bensimon D. 1994. Alignment and sensitive detection of DNA by a moving interface. *Science* **265**: 2096–2098.

Bérard J, Gouéré JB, Piau D. 2008. Solvable models of neighbor-dependent substitution processes. *Math Biosci* **211**: 56–88.

Berezney R, Dubey DD, Huberman JA. 2000. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* **108**: 471–484.

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

Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR. 2007. Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome Res* **17**: 910–916.

Blow JJ, Ge XQ. 2009. A model for DNA replication showing how dormant origins safeguard against replication fork failure. *EMBO Rep* **10**: 406–412.

Blow JJ, Gillespie PJ, Francis D, Jackson DA. 2001. Replication origins in xenopus egg extract are 5-15 kilobases apart and are activated in clusters that fire at different times. *J Cell Biol* **152**: 15–25.

Blumenthal AB, Kriegstein HJ, Hogness DS. 1974. The units of DNA replication in drosophila melanogaster chromosomes. *Cold Spring Harb Symp Quant Biol* **38**: 205–223.

Bogan JA, Natale DA, Depamphilis ML. 2000. Initiation of eukaryotic DNA replication: conservative or liberal? *J Cell Physiol* **184**: 139–150.

Borštnik B, Pumpernik D, Lukman D. 1993. Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences. *Europhys Lett* **23**: 389–394.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008a. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311–322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008b. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537–2538.

Branco MR, Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**: e138.

Brodie of Brodie EB, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2005. From DNA sequence analysis to modeling replication in the human genome. *Phys Rev Lett* **94**: 248103.

Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE. 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys Rev E* **51**: 5084–5091.

Bulmer M. 1991. Strand symmetry of mutation rates in the beta-globin region. *J Mol Evol* **33**: 305–310.

Buongiorno-Nardelli M, Micheli G, Carri MT, Marilley M. 1982. A relationship between replicon size and supercoiled loop domains in the eukaryotic genome. *Nature* **298**: 100–102.

Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* **89**: 1358–1362.

Caddle LB, Grant JL, Szatkiewicz J, van Hase J, Shirley BJ, Bewersdorf J, Cremer C, Arneodo A, Khalil A, Mills KD. 2007. Chromosome neighborhood composition determines translocation outcomes after exposure to high-dose radiation in primary cells. *Chrom Res* **15**: 1061–1073.

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau MN. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci USA* **105**: 15837–15842.

Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SSW, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jégou B, et al. 2007. The conserved transcriptome in human and rodent male gametogenesis. *Proc Natl Acad Sci U S A* **104**: 8346–8351.

Chambeyron S, Bickmore WA. 2004. Does looping and clustering in the nucleus regulate gene expression? *Curr Opin Cell Biol* **16**: 256–262.

Chatzidimitriou-Dreismann CA, Larhammar D. 1993. Long-range correlations in DNA. *Nature* **361**: 212–213.

Chen CL, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, Huvet M, d'Aubenton Carafa Y, Hyrien O, Arneodo A, et al. 2011. Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol* **28**: 2327–2337.

Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**: 447–457.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**: 1149–1154.

Christian J. 2002. *The Theory of Phase Transformations in Metals and Alloys, Part I: Equilibrium and General Kinectics Theory.* Pergamon Press, Oxford.

Combes J, Grossmann A, Tchamitchian P, eds. 1989. *Wavelets.* Springer, Berlin.

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**: 1293–1304.

Conti C, Sacca B, Herrick J, Lalou C, Pommier Y, Bensimon A. 2007. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol Biol Cell* **18**: 3059–3067.

Cook PR. 1995. A chromomeric model for nuclear and chromosome structure. *J Cell Sci* **108**: 2927–2935.

Cook PR. 1999. The organization of replication and transcription. *Science* **284**: 1790–1795.

Cook PR. 2001. *Principles of Nuclear Structure and Functions*. Wiley, New York.

Cook PR. 2002. Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet* **32**: 347–352.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845–1848.

Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, Debatisse M. 2008. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature* **455**: 557–560.

Coverley D, Laskey RA. 1994. Regulation of eukaryotic DNA replication. *Annu Rev Biochem* **63**: 745–776.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**: 292–301.

Czajkowsky DM, Liu J, Hamlin JL, Shao Z. 2008. DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI. *J Mol Biol* **375**: 12–19.

Daubechies I. 1992. *Ten Lectures on Wavelets*. SIAM, Philadelphia.

de Moura APS, Retkute R, Hawkins M, Nieduszynski CA. 2010. Mathematical modelling of whole chromosome replication. *Nucleic Acids Res* **38**: 5623–5633.

Decoster N, Roux SG, Arneodo A. 2000. A wavelet-based method for multifractal image analysis. II. Applications to synthetic multifractal rough surfaces. *Eur Phys J B* **15**: 739–764.

Dekker J. 2003. A closer look at long-range chromosomal interactions. *Trends Biochem Sci* **28**: 277–280.

Delour J, Muzy JF, Arneodo A. 2001. Intermittency of 1D velocity spatial profiles in turbulence: a magnitude cumulant analysis. *Eur Phys J B* **23**: 243–248.

Demeret C, Vassetzky Y, Méchali M. 2001. Chromatin remodelling and DNA replication: from nucleosomes to loop domains. *Oncogene* **20**: 3086–3093.

DePamphilis ML, ed. 2006. *DNA Replication and Human Disease*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New-York.

Desprat R, Thierry-Mieg D, Lailler N, Lajugie J, Schildkraut C, Thierry-Mieg J, Bouhassira EE. 2009. Predictable dynamic program of timing of DNA replication in human cells. *Genome Res* **19**: 2288–2299.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**: 640–649.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4**: e1000071.

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

Erlebacher G, Hussaini M, Jameson L, eds. 1996. *Wavelets : Theory and Applications*. Oxford University Press, Oxford.

Evans J. 1993. Random and cooperative sequential adsorption. *Rev Mod Phys* **65**: 1281–1329.

Fanfoni M, Tomellini M. 1998. The Johnson-Mehl-Avrami-Kolmogorov model: a brief review. *Il Nuevo Cimento* **20D**: 1171–1181.

Farge M, Hunt J, Vassilicos J, eds. 1993. *Wavelets, Fractals and Fourier*. Clarendon Press, Oxford.

Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, Yakhini Z, Simon I. 2008. Global organization of replication time zones of the mouse genome. *Genome Res* **18**: 1562–1570.

Fisher D, Méchali M. 2003. Vertebrate HoxB gene expression requires DNA replication. *EMBO J* **22**: 3737–3748.

Francino MP, Chao L, Riley MA, Ochman H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**: 107–109.

Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. *Trends Genet* **13**: 240–245.

Francino MP, Ochman H. 2000. Strand symmetry around the beta-globin origin of replication in primates. *Mol Biol Evol* **17**: 416–422.

Francino MP, Ochman H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. *Mol Biol Evol* **18**: 1147–1150.

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**: 65–77.

Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**: 2532–2537.

Friedman KL, Brewer BJ, Fangman WL. 1997. Replication profile of Saccharomyces cerevisiae chromosome VI. *Genes Cells* **2**: 667–678.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**: 907–911.

Gardiner K. 1996. Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet* **12**: 519–524.

Gasser SM. 2002. Visualizing chromatin dynamics in interphase nuclei. *Science* **296**: 1412–1416.

Gauthier MG, Bechhoefer J. 2009. Control of DNA replication by anomalous reaction-diffusion kinetics. *Phys Rev Lett* **102**: 158104.

Gerbi SA, Bielinsky AK. 2002. DNA replication and chromatin. *Curr Opin Genet Dev* **12**: 243–248.

Gierlik A, Kowalczuk M, Mackiewicz P, Dudek MR, Cebrat S. 2000. Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? *J Theor Biol* **202**: 305–314.

Gilbert DM. 2001. Making sense of eukaryotic DNA replication origins. *Science* **294**: 96–100.

Gilbert DM. 2004. In search of the holy replicator. *Nat Rev Mol Cell Biol* **5**: 848–855.

Gilbert DM. 2010. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat Rev Genet* **11**: 673–684.

Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. 2004. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**: 555–566.

Gilbert N, Gilchrist S, Bickmore WA. 2005. Chromatin organization in the mammalian nucleus. *Int Rev Cytol* **242**: 283–336.

Goupillaud P, Grossmann A, Morlet J. 1984. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration* **23**: 85–102.

Graur D, Li WH. 1999. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**: 514–517.

Grigoriev A. 1999. Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res* **60**: 1–19.

Grossmann A, Morlet J. 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J of Math Anal* **15**: 723–736.

Grossmann A, Morlet J. 1985. Decomposition of functions into wavelets of constant shape and related transforms. In L Streit, ed., *Mathematics and Physics, Lectures on Recent Results*, pages 135–165. World Scientific, Singapore.

Hamlin JL, Mesner LD, Dijkwel PA. 2010. A winding road to origin discovery. *Chromosome Res* **18**: 45–61.

Hamlin JL, Mesner LD, Lar O, Torres R, Chodaparambil SV, Wang L. 2008. A revisionist replicon model for higher eukaryotic genomes. *J Cell Biochem* **105**: 321–329.

Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F, et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**: 630–638.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci USA* **107**: 139–144.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. *Science* **322**: 1855–1857.

Hentschel HGE. 1994. Stochastic multifractality and universal scaling distributions. *Phys Rev E* **50**: 243–261.

Herrick J, Jun S, Bechhoefer J, Bensimon A. 2002. Kinetic model of DNA replication in eukaryotic organisms. *J Mol Biol* **320**: 741–750.

Herrick J, Stanislawski P, Hyrien O, Bensimon A. 2000. Replication fork density increases during DNA synthesis in X. laevis egg extracts. *J Mol Biol* **300**: 1133–1142.

Herzel H, Große I. 1995. Measuring correlations in symbol sequences. *Physica A* **216**: 518–542.

Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. *J Mol Biol* **236**: 1022–1033.

Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, Fussner E, Bazett-Jones DP, Plath K, Dalton S, et al. 2010. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**: 155–169.

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* **6**: e245.

Holschneider M. 1988. On the wavelet transform of fractal objects. *J Stat Phys* **50**: 963–993.

Holschneider M, Tchamitchian P. 1990. Régularité locale de la fonction non-différentiable de Riemann. In PG Lemarié, ed., *Les Ondelettes en 1989*, pages 102–124. Springer, Berlin.

Horn PJ, Peterson CL. 2002. Molecular biology. Chromatin higher order folding–wrapping up transcription. *Science* **297**: 1824–1827.

Hou C, Dale R, Dean A. 2010. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* **107**: 3651–3656.

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome Res* **17**: 1278–1285.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* **101**: 13994–14001.

Hyrien O, Goldar A. 2010. Mathematical modelling of eukaryotic DNA replication. *Chromosome Res* **18**: 147–161.

Hyrien O, Marheineke K, Goldar A. 2003. Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *Bioessays* **25**: 116–125.

Hyrien O, Méchali M. 1993. Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. *EMBO J* **12**: 4511–4520.

Ivanov PC, Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, Stanley HE. 1999. Multifractality in human heartbeat dynamics. *Nature* **399**: 461–465.

Ivanov PC, Rosenblum MG, Peng CK, Mietus J, Havlin S, Stanley HE, Goldberger AL. 1996. Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature* **383**: 323–327.

Jacob F, Brenner S, Cuzin F. 1963. On the regulation of DNA replication in bacteria. *Cold Spring Harb Symp Quant Biol* **28**: 329–342.

Jaffard S. 1989. Hölder exponents at given points and wavelet coefficients. *C R Acad Sci Paris Sér I* **308**: 79–81.

Jaffard S. 1991. Pointwise smoothness, two-microlocalization and wavelet coefficients. *Publ Mat* **35**: 155–168.

Jaffard S. 1997a. Multifractal formalism for functions part I: results valid for all functions. *SIAM J Math Anal* **28**: 944–970.

Jaffard S. 1997b. Multifractal formalism for functions part II: self-similar functions. *SIAM J Math Anal* **28**: 971–998.

Jaffard S, Lashermes B, Abry P. 2006. Wavelet leaders in multifractal analysis. In T Qian, MI Vai, Y Xu, eds., *Wavelet Analysis and Applications*, pages 219–264. Birkhäuser Verlag, Basel, Switzerland.

Jaffard S, Meyer Y, Ryan R, eds. 2001. *Wavelets : Tools for Science and Technology.* SIAM, Philadelphia.

Johnson WA, Mehl P. 1939. Reaction kinetics in processes of nucleation and growth. *Trans AIME* **135**: 416–442.

Jun S, Bechhoefer J. 2005. Nucleation and growth in one dimension. II. Application to DNA replication kinetics. *Phys Rev E* **71**: 011909.

Jun S, Herrick J, Bensimon A, Bechhoefer J. 2004. Persistence length of chromatin determines origin spacing in Xenopus early-embryo DNA replication: quantitative comparisons between theory and experiment. *Cell Cycle* **3**: 223–229.

Jun S, Zhang H, Bechhoefer J. 2005. Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model. *Phys Rev E* **71**: 011908.

Karlin S, Brendel V. 1993. Patchiness and correlations in DNA sequences. *Science* **259**: 677–680.

Karnani N, Taylor CM, Dutta A. 2009. Microarray analysis of DNA replication timing. *Methods Mol Biol* **556**: 191–203.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. 2003. The UCSC genome browser database. *Nucleic Acids Res* **31**: 51–54.

Kestener P, Arneodo A. 2003. Three-dimensional wavelet-based multifractal method: The need for revisiting the multifractal description of turbulence dissipation data. *Phys Rev Lett* **91**: 194501.

Kestener P, Arneodo A. 2004. Generalizing the wavelet-based multifractal formalism to random vector fields: Application to three-dimensional turbulence velocity and vorticity data. *Phys Rev Lett* **93**: 044501.

Kestener P, Arneodo A. 2007. A multifractal formalism for vector-valued random fields based on wavelet analysis: application to turbulent velocity and vorticity 3D numerical data. *Stoch Environ Res Risk Assess* **22**: 421–435.

Kestener P, Colon P, Khalil A, Fennel L, McAteer R, Gallagher P, Arneodo A. 2010. Characterizing complexity in solar magnetization data using a wavelet-based segmentation method. *Astrophysical Journal* **717**: 995–1005.

Kestener P, Lina JM, Saint-Jean P, Arneodo A. 2001. Wavelet-based multifractal formalism to assist in diagnosis in digitized mammograms. *Image Anal Stereol* **20**: 169–174.

Khalil A, Grant J, Caddle L, Atzema E, Mills K, Arneodo A. 2007. Chromosome territories have a highly non-spherical morphology and non-random positioning. *Chrom Res* **15**: 889–916.

Khalil A, Joncas G, Nekka F, Kestener P, Arneodo A. 2006. Morphological analysis of H_I features. II. Wavelet-based multifractal formalism. *Astrophysical Journal Supplement Series* **165**: 512–550.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.

Kolmogorov AN. 1937. On the static theory of crystallization in metals. *Bull Acad Sci URSS* **3**: 335.

Kuhn A, Argoul F, Muzy JF, Arneodo A. 1994. Structural-analysis of electroless deposits in the diffusion-limited regime. *Phys Rev Lett* **73**: 2998–3001.

Kunkel TA, Burgers PM. 2008. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol* **18**: 521–527.

Larhammar D, Chatzidimitriou-Dreismann CA. 1993. Biological origins of long-range correlations and compositional variations in DNA. *Nucleic Acids Res* **21**: 5167–5170.

Larrea AA, Lujan SA, McElhinny SAN, Mieczkowski PA, Resnick MA, Gordenin DA, Kunkel TA. 2010. Genome-wide model for the normal eukaryotic DNA replication fork. *Proc Natl Acad Sci U S A* **107**: 17674–17679.

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, ichi Isono K, et al. 2006. Control of developmental regulators by polycomb in human embryonic stem cells. *Cell* **125**: 301–313.

Lemaitre C, Zaghloul L, Sagot MF, Gautier C, Arneodo A, Tannier E, Audit B. 2009. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics* **10**: 335.

Li W. 1992. Generating non trivial long-range correlations and $1/f$ spectra by replication and mutation. *Int J Bifurc Chaos* **2**: 137–154.

Li W. 1997. The study of correlation structures of DNA sequences: a critical review. *Comput Chem* **21**: 257–271.

Li W, Stolovitzky G, Bernaola-Galván P, Oliver JL. 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Res* **8**: 916–928.

Li WT, Marr TG, Kaneko K. 1994. Understanding long-range correlations in DNA-sequences. *Physica D* **75**: 392–416.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lobry JR. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* **40**: 326–330.

Lobry JR. 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**: 660–665.

Lobry JR. 1996b. Origin of replication of mycoplasma genitalium. *Science* **272**: 745–746.

Lobry JR, Lobry C. 1999. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol Biol Evol* **16**: 719–723.

Lucas I, Palakodeti A, Jiang Y, Young DJ, Jiang N, Fernald AA, Le Beau MM. 2007. High-throughput mapping of origins of replication in human cells. *EMBO Rep* **8**: 770–777.

Luo H, Li J, Eshaghi M, Liu J, Karuturi RKM. 2010. Genome-wide estimation of firing efficiencies of origins of DNA replication from time-course copy number variation data. *BMC Bioinformatics* **11**: 247.

Lygeros J, Koutroumpas K, Dimopoulos S, Legouras I, Kouretas P, Heichinger C, Nurse P, Lygerou Z. 2008. Stochastic hybrid modeling of DNA replication across a complete genome. *Proc Natl Acad Sci USA* **105**: 12295–12300.

Majewski J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet* **73**: 688–692.

Mallat S. 1998. *A Wavelet Tour in Signal Processing.* Academic Press, New York.

Mallat S, Hwang W. 1992. Singularity detection and processing with wavelets. *IEEE Trans Info Theory* **38**: 617–643.

Mallat S, Zhong S. 1992. Characterization of signals from multiscale edges. *IEEE Trans Patt Recog Mach Intell* **14**: 710–732.

Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, Stanley HE. 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys Rev E* **52**: 2939–2950.

Marheineke K, Goldar A, Krude T, Hyrien O. 2009. *DNA Replication Methods and Protocols*, chapter Use of DNA combing to study DNA replication in Xenopus and human cell-free systems, pages 575–603. S. Vengrova and J.Z. Dalgraad (Eds) Springer, New York.

McAteer R, Kestener P, Arneodo A, Khalil A. 2010. Automated detection of coronal loop using a wavelet transform modulus maxima method. *Solar Physics* **262**: 387–397.

McCune HJ, Danielson LS, Alvino GM, Collingwood D, Delrow JJ, Fangman WL, Brewer BJ, Raghuraman MK. 2008. The temporal program of chromosome replication: genomewide replication in clb5Δ Saccharomyces cerevisiae. *Genetics* **180**: 1833–1847.

McNairn AJ, Gilbert DM. 2003. Epigenomic replication: linking epigenetics to DNA replication. *Bioessays* **25**: 647–656.

McVicker G, Green P. 2010. Genomic signatures of germline gene expression. *Genome Res* **20**: 1503–1511.

Méchali M. 2001. DNA replication origins: from sequence specificity to epigenetics. *Nat Rev Genet* **2**: 640–645.

Méchali M. 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat Rev Mol Cell Biol* **11**: 728–738.

Mesner LD, Crawford EL, Hamlin JL. 2006. Isolating apparently pure libraries of replication origins from complex genomes. *Mol Cell* **21**: 719–726.

Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, Bekiranov S. 2011. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. *Genome Res* **21**: 377–389.

Meyer Y, ed. 1992. *Wavelets and their Applications.* Springer, Berlin.

Meyer Y, Roques S, eds. 1993. *Progress in Wavelets Analysis and Applications.* Éditions Frontières, Gif-sur-Yvette.

Mordant N, Delour J, Léveque E, Arneodo A, Pinton JF. 2002. Long-time correlations in Lagrangian dynamics: a key to intermittency in turbulence. *Phys Rev Lett* **89**: 254502.

Mordant N, Delour J, Léveque E, Michel O, Arneodo A, Pinton JF. 2003. Lagrangian velocity fluctuations in fully developed turbulence : scaling, intermittency, and dynamics. *J Stat Phys* **113**: 701–717.

Moukhtar J, Faivre-Moskalenko C, Milani P, Audit B, Vaillant C, Fontaine E, Mongelard F, Lavorel G, St-Jean P, Bouvet P, et al. 2010. Effect of genomic long-range correlations on DNA persistence length: from theory to single molecule experiments. *J Phys Chem B* **114**: 5125–5143.

Moukhtar J, Fontaine E, Faivre-Moskalenko C, Arneodo A. 2007. Probing persistence in DNA curvature properties with atomic force microscopy. *Phys Rev Lett* **98**: 178101.

Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci USA* **95**: 3720–3725.

Mugal CF, von Grunberg HH, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol* **26**: 131–142.

Mugal CF, Wolf JBW, von Grunberg HH, Ellegren H. 2010. Conservation of neutral substitution rate and substitutional asymmetries in mammalian genes. *Genome Biol Evol* **2**: 19–28.

Müller WG, Rieder D, Kreth G, Cremer C, Trajanoski Z, McNally JG. 2004. Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol Cell Biol* **24**: 9359–9370.

Münkel C, Eils R, Dietzel S, Zink D, Mehring C, Wedemann G, Cremer T, Langowski J. 1999. Compartmentalization of interphase chromosomes observed in simulation and experiment. *J Mol Biol* **285**: 1053–1065.

Muzy JF, Bacry E, Arneodo A. 1991. Wavelets and multifractal formalism for singular signals: Application to turbulence data. *Phys Rev Lett* **67**: 3515–3518.

Muzy JF, Bacry E, Arneodo A. 1993. Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. *Phys Rev E* **47**: 875–884.

Muzy JF, Bacry E, Arneodo A. 1994. The multifractal formalism revisited with wavelets. *Int J Bifurc Chaos* **4**: 245–302.

Muzy JF, Sornette D, Delour J, Arneodo A. 2001. Multifractal returns and hierarchical portfolio theory. *Quant Finance* **1**: 131–148.

Necsulea A, Guillet C, Cadoret JC, Prioleau MN, Duret L. 2009. The relationship between DNA replication and human genome organization. *Mol Biol Evol* **26**: 729–741.

Nee S. 1992. Uncorrelated DNA walks. *Nature* **357**: 450.

Nicolay S. 2006. *Analyse des séquences d'ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles.* Ph.D. thesis, University of Liège, Belgium.

Nicolay S, Argoul F, Touchon M, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2004. Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? *Phys Rev Lett* **93**: 108101.

Nicolay S, Brodie of Brodie EB, Touchon M, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2007. Bifractality of human DNA strand-asymmetry profiles results from transcription. *Phys Rev E* **75**: 032902.

Nieduszynski CA, Hiraga SI, Ak P, Benham CJ, Donaldson AD. 2007. OriDB: a DNA replication origin database. *Nucleic Acids Res* **35**: D40–D46.

Ohlsson R, Lobanenkov V, Klenova E. 2010. Does CTCF mediate between nuclear organization and gene expression? *Bioessays* **32**: 37–50.

Ostashevsky J. 1998. A polymer model for the structural organization of chromatin loops and minibands in interphase chromosomes. *Mol Biol Cell* **9**: 3031–3040.

Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N. 2006. DNA replication origins fire stochastically in fission yeast. *Mol Biol Cell* **17**: 308–316.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE. 1992. Long-range correlations in nucleotide sequences. *Nature* **356**: 168–170.

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Simons M, Stanley HE. 1993. Finite-size effects on long-range correlations: Implications for analysing DNA sequences. *Phys Rev E* **47**: 3730–3733.

Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. 1994. Mosaic organization of DNA nucleotides. *Phys Rev E* **49**: 1685–1689.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**: 1194–1211.

Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res* **18**: 1216–1223.

Polak P, Arndt PF. 2009. Long-range bidirectional strand asymmetries originate at CpG islands in the human genome. *Genome Biol Evol* **1**: 189–197.

Ponger L, Duret L, Mouchiroud D. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res* **11**: 1854–1860.

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**: 1851–1854.

Pursell ZF, Isoz I, Lundström EB, Johansson E, Kunkel TA. 2007. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. *Science* **317**: 127–130.

Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ, Fangman WL. 2001. Replication dynamics of the yeast genome. *Science* **294**: 115–121.

Rappailles A, Guilbaud G, Baker A, Chen C, Arneodo A, Goldar A, d'Aubenton-Carafa Y, Thermes C, Audit B, Hyrien O. 2011. Sequential activation of replication origins along replication timing gradients in the human genome. *PLoS Comput Biol* **in press**.

Retkute R, Nieduszynski CA, de Moura APS. 2011. Dynamics of DNA replication in yeast. *Phys Rev Lett* **107**: 068103.

Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* **15**: 957–966.

Rhind N. 2006. DNA replication timing: random thoughts about origin firing. *Nat Cell Biol* **8**: 1313–1316.

Rhind N, Yang SCH, Bechhoefer J. 2010. Reconciling stochastic origin firing with defined replication timing. *Chromosome Res* **18**: 35–43.

Rocha EP, Danchin A, Viari A. 1999. Universal replication biases in bacteria. *Mol Microbiol* **32**: 11–16.

Rocha EPC, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. *Genome Res* **16**: 1537–1547.

Roux S, Muzy JF, Arneodo A. 1999. Detecting vorticity filaments using wavelet analysis: About the statistical contribution of vorticity filaments to intermittency in swirling turbulent flows. *Eur Phys J B* **8**: 301–322.

Roux S, Venugopal V, Fineberg K, Arneodo A, Foufoula-Georgiou E. 2009. Evidence for inherent nonlinearity in temporal rainfall. *Adv in Water Resources* **32**: 41–48.

Roux SG, Arneodo A, Decoster N. 2000. A wavelet-based method for multifractal image analysis. III. Applications to high-resolution satellite images of cloud structure. *Eur Phys J B* **15**: 765–786.

Rudner R, Karkas JD, Chargaff E. 1968. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci USA* **60**: 921–922.

Ruskai M, Beylkin G, Coifman R, Daubechies I, Mallat S, Meyer Y, Raphael L, eds. 1992. *Wavelets and their Applications*. Jones and Barlett, Boston.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**: 761–770.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3**: 511–518.

Sachs RK, van den Engh G, Trask B, Yokota H, Hearst JE. 1995. A random-walk/giant-loop model for interphase chromosomes. *Proc Natl Acad Sci USA* **92**: 2710–2714.

Sasaki T, Sawado T, Yamaguchi M, Shinomiya T. 1999. Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalphadE2F locus of Drosophila melanogaster. *Mol Cell Biol* **19**: 547–555.

Schübeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M. 2002. Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat Genet* **32**: 438–442.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322**: 1849–1851.

Sekimoto K. 1984a. Kinetics of magnetization switching in 1D system (1): size distribution of unswitched domains. *Physica A* **125**: 261–269.

Sekimoto K. 1984b. Kinetics of magnetization switching in a 1D system (2): long time behaviour of switched domains. *Physica A* **128**: 132–149.

Sekimoto K. 1986. Evolution of domain structure during the nucleation-and-growth process with non-conserved order paramter. *Physica A* **135**: 328–346.

Shopland LS, Lynch CR, Peterson KA, Thornton K, Kepper N, von Hase J, Stein S, Vincent S, Molloy KR, Kreth G, et al. 2006. Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J Cell Biol* **174**: 27–38.

Silverman B, Vassilicos J, eds. 2000. *Wavelets : The Key to Intermittent Information?* Oxford University Press, Oxford.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393–395.

Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Ossadnik SM, Peng CK, Simons M. 1993. Fractal landscapes in biological systems. *Fractals* **1**: 283–301.

Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* **40**: 318–325 (erratum in idib **42**:373).

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465–476.

Svejstrup JQ. 2002. Mechanisms of transcription-coupled DNA repair. *Nat Rev Mol Cell Biol* **3**: 21–29.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Tillier ER, Collins RA. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* **50**: 249–257.

Tomellini M, Fanfoni M. 1997. Why phantom nuclei must be considered in the Johnson-Mehl-Avrami-Kolmogorov kinetics. *Phys Rev E* **55**: 14071–14073.

Torresani B. 1998. *Analyse Continue par Ondelettes*. Éditions de Physique, Les Ulis.

Touchon M. 2005. *Biais de composition chez les mammifères : rôle de la transcription et de la réplication*. Ph.D. thesis, University Denis Diderot, Paris VII, France.

Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res* **32**: 4969–4978.

Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett* **555**: 579–582.

Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci USA* **102**: 9836–9841.

Vaillant C, Audit B, Arneodo A. 2005. Thermodynamics of DNA loops with long-range correlated structural disorder. *Phys Rev Lett* **95**: 068101.

Vaillant C, Audit B, Arneodo A. 2007. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys Rev Lett* **99**: 218103.

Vaillant C, Audit B, Thermes C, Arneodo A. 2006. Formation and positioning of nucleosomes: effect of sequence-dependent long-range correlated structural disorder. *Eur Phys J E* **19**: 263–277.

Van Kampen N. 2007. *Stochastic Processes in Physics and Chemistry, 3rd ed.* North-Holland, Amsterdam.

Venugopal V, Roux SG, Foufoula-Georgiou E, Arneodo A. 2006a. Revisiting multifractality of high-resolution temporal rainfall using a wavelet-based formalism. *Water Resour Res* **42**: W06D14.

Venugopal V, Roux SG, Foufoula-Georgiou E, Arneodo A. 2006b. Scaling behavior of high resolution temporal rainfall: New insights from a wavelet-based cumulant analysis. *Phys Lett A* **348**: 335–345.

Viswanathan GM, Buldyrev SV, Havlin S, Stanley HE. 1998. Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA. *Physica A* **249**: 581–586.

Voss RF. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys Rev Lett* **68**: 3805–3808.

Voss RF. 1994. Long-range fractal correlations in DNA introns and exons. *Fractals* **2**: 1–6.

Wald R. 1984. *General Relativity.* University of Chicago Press, Chicago.

Wolffe AP. 1998. *Chromatin Structure and Function, 3rd ed.* Academic Press, London.

Woodfine K, Beare DM, Ichimura K, Debernardi S, Mungall AJ, Fiegler H, Collins VP, Carter NP, Dunham I. 2005. Replication timing of human chromosome 6. *Cell Cycle* **4**: 172–176.

Wu CI. 1991. DNA strand asymmetry. *Nature* **352**: 114.

Wu CI, Maeda N. 1987. Inequality in mutation rates of the two strands of DNA. *Nature* **327**: 169–170.

Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, Simon I. 2010. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet* **6**: e1001011.

Yang SCH, Bechhoefer J. 2008. How Xenopus laevis embryos replicate reliably: investigating the random-completion problem. *Phys Rev E* **78**: 041917.

Yang SCH, Rhind N, Bechhoefer J. 2010. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol Syst Biol* **6**: 404.

Zaghloul L. 2009. *Transcriptional activity, chromatin state and replication timing in domains of compositional skew in the huamn genome.* Ph.D. thesis, ENS de Lyon, France.

Zhang H, Bechhoefer J. 2006. Reconstructing DNA replication kinetics from small DNA fragments. *Phys Rev E* **73**: 051903.