Eur. Phys. J. E (2012) **35**: 92
DOI 10.1140/epje/i2012-12092-y

THE EUROPEAN PHYSICAL JOURNAL E

Regular Article

Linking the DNA strand asymmetry to the spatio-temporal replication program

I. About the role of the replication fork polarity in genome evolution

A. Baker[1,2], H. Julienne[1,2], C.L. Chen[3], B. Audit[1,2,a], Y. d'Aubenton-Carafa[3], C. Thermes[3], and A. Arneodo[1,2,b]

[1] Université de Lyon, F-69000 Lyon, France
[2] Laboratoire de Physique, ENS Lyon, CNRS, 69007 Lyon, France
[3] Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France

Received 6 June 2012 and Received in final form 8 August 2012
Published online: 26 September 2012

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). We elaborate on the simplest model of neutral substitution rates that takes into account the strand asymmetries generated by the transcription and replication processes. Using perturbation theory, we then solve the time evolution of the DNA composition under strand-asymmetric substitution rates. In our minimal model, the compositional and substitutional asymmetries are predicted to decompose into a transcription- and a replication-associated components. The transcription-associated asymmetry increases in magnitude with transcription rate and changes sign with gene orientation while the replication-associated asymmetry is proportional to the replication fork polarity. These results are confirmed experimentally in the human genome, using substitution rates obtained by aligning the human and chimpanzee genomes using macaca and orangutan as outgroups, and replication fork polarity determined in the HeLa cell line as estimated from the derivative of the mean replication timing. When further investigating the dynamics of compositional skew evolution, we show that it is not at equilibrium yet and that its evolution is an extremely slow process with characteristic time scales of several hundred Myrs.

1 Introduction

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 [1–10]. How is the spatio-temporal replication program regulated? 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? In this manuscipt we 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 (see sect. 2): 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. Only recently the existence of a replication-associated strand asymmetry, originally established in bacteria [11,12], has been confirmed in higher eukaryotes and in particular in the human genome [13].

A clear relationship between replication and compositional asymmetry was first established in prokaryotic genomes by Lobry [11]. In bacteria, the spatio-temporal replication program is particularly simple. Most prokaryotes follow the replicon model depicted in fig. 1A: the repli-

[a] e-mail: `benjamin.audit@ens-lyon.fr`
[b] e-mail: `alain.arneodo@ens-lyon.fr`

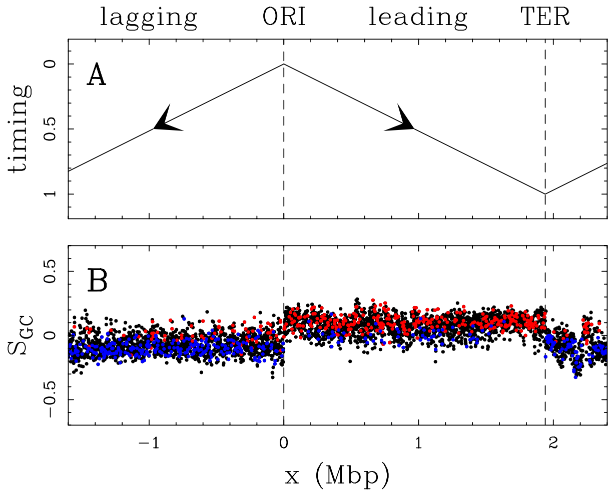

Fig. 1. (Colour on-line) Comparing GC skew $S_{GC} = \frac{G-C}{G+C}$ and replication timing in the *Bacillus subtilis* genome. (A) Spatio-temporal representation of the replicon model: two divergent replication forks porgress from the replication origin (ORI) to the replication terminus (TER). The replication 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 (respectively, blue) points correspond to (+) (respectively, (−)) genes, which coding sequences are on the Watson, respectively Crick, strand (the Watson strand corresponds to the DNA sequence as it is available in the database and the Crick strand to its reverse complementary sequence).

cation 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) (see sect. 2). As shown in fig. 1B for *Bacillus subtilis*, many prokaryotic genomes are circular and 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.

By contrast, the spatio-temporal replication program in eukaryotes is much more complex [14,15]. 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. Thus, as a consequence of 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 as quantified by the replication fork polarity, defined as the difference of proportions of right-moving and left-moving forks. Therefore in eukaryotes, in contrast to bacteria, leading and lagging strands cannot be unambiguously assigned to a genomic region. In that context, one may wonder whether the relationship observed between the replication program and the compositional asymmetry in bacteria (fig. 1) actually generalizes to eukaryotic genomes.

As shown in fig. 2 and originally discussed in [16,17] we still observe a clear relationship between the compositional asymmetry and the replication timing in the human genome: an 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 [18–20]. Those genomic domains, that were called N-domains, were shown to exhibit a very peculiar gene organization and chromatin state [21–24]. Based on the analogy with the bacterial case (the upward jump of the GC skew colocalizes with the ORI in fig. 1), the N-domains borders (upward jumps of the skew) were proposed to be replication origins, evolutionary conserved and active in the germline [18,19]. 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 ∼ 1 Mbp characteristic size, this model would imply that large ∼ 1 Mbp replicons are produced in the 30% of the human genome covered by N-domains, in conflict with the typical observed replicon size (∼ 100 kbp).

As suggested in [17] the key parameter that allows to link the compositional asymmetry to the replication timing is the replication fork polarity. Under the assumption that the replication fork velocity v is constant, and that origins are bidirectional, we have shown [17] that the replication fork polarity is proportional to the derivative of the mean replication timing (MRT):

$$p(x) \simeq v\, T_S\, \mathrm{d}\,\mathrm{MRT}/\mathrm{d}x, \tag{1}$$

where the MRT, expressed in unit of the S-phase fraction, multiplied by T_s the duration of the S phase, gives a reasonable proxy for the MRT expressed in unit of time. We have also argued why the skew S is also expected to be proportional to the replication fork polarity. The replication fork polarity being proportional to both $\mathrm{d}\,\mathrm{MRT}/\mathrm{d}x$ and S, it provides a unifying understanding of why the replicon model in bacteria (fig. 1A) results in the crenel-like skew profile (fig. 1B), and why the U-shaped replication timing profile observed in the human genome (fig. 2A) results in the N-shaped skew profile (fig. 2B).

In this paper we will detail the precise neutral molecular evolution scenario (from the strand asymmetries generated by the replication process, to their impact on substitutional rates and the resulting evolution of the DNA composition) that allows to link the compositional asymmetry to the replication fork polarity. We will further take into account the strand asymmetries generated by the transcription process. In particular, 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? As compositional asym-

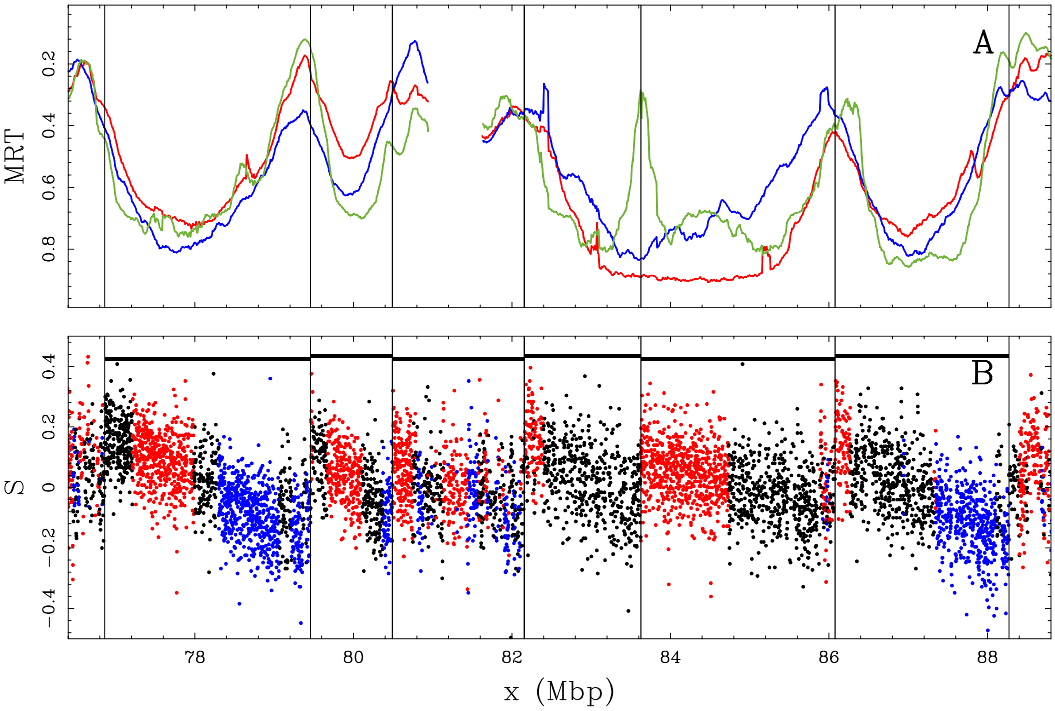

Fig. 2. (Colour on-line) Comparing compositional skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ and mean replication timing (MRT) in the human genome. (A) MRT profiles along a 11.4 Mbp long fragment of human chromosome 10, from early (0) to late (1) for BG02 embryonic stem cell (green), K562 erythroid (red) and GM06990 lymphoblastoid (blue) cell lines. Replication timing data was retrieved from [25] (appendix A.1). (B) S calculated in 1 kbp windows of repeat-masked sequences. The colors correspond to intergenic (black), (+) genes (red) and (−) genes (blue). Six skew N-domains (horizontal black bars) were detected in this genomic region [17].

metry is also associated to transcription in the human genome [26,27], is it possible to disentangle in the skew profile the contributions associated to transcription and replication?

The paper is organized as follows. In sect. 2 we review background knowledge on transcription and replication as processes that likely break DNA strand symmetry. Section 3 is devoted to the modelling of the impact of replication fork polarity, gene orientation and transcription rate on substitution rates. We elaborate on a simple model with a minimal number of parameters that takes into account the basic symmetries of the problem. We argue that most molecular mechanisms proposed so far to explain DNA strand asymmetry are particular cases of this minimal model. We justify the pertinence of this minimal model by combining the study of substitution rates in genic and intergenic regions of the human genome with the estimation of the replication fork polarity profile along human chromosomes, using experimental replication timing data determined in the HeLa cell line [28,29]. In sect. 4 we revisit the general formalism for DNA composition evolution when introducing strand-symmetric and strand-asymmetric variables. We use perturbation theory to demonstrate that in the framework of our minimal model of substitutional asymmetry, the compositional asymmetry linearly decomposes into a transcription-associated and a replication-associated components. The replication-associated compositional asymmetry is proportional to the replication fork polarity, whereas the transcription-associated one increases in magnitude with the transcription rate and changes sign with the gene orientation. In sect. 5 we confirm the results of our theoretical perturbative analysis in the human genome. We conclude in sect. 6 by discussing the robustness of our results when using replication timing data from different human cell types.

2 Transcription and replication as strand symmetry breaking processes

Due to base-pairing and anti-parallel orientation, the nucleotide sequences on the two DNA strands are related by *reverse complementarity*: a guanine G on one strand is always linked to a cytosine C on the other strand (three hydrogen bonds, strong coupling); a thymine T on one strand is always linked to an adenine A on the other strand (two hydrogen bonds, weak coupling) [14]. The nucleotide sequence on one DNA strand is conventionally read in the $5' \rightarrow 3'$ direction. This is typically the case for the published strand commonly called *reference strand*. The polarity of the DNA strands has great biological importance. For instance the DNA polymerase always synthesizes the newly replicated strand in the $5' \rightarrow 3'$ direction. Similarly, the RNA polymerase always synthesizes the messenger RNA in the $5' \rightarrow 3'$ direction.

2.1 DNA strand symmetry

2.1.1 Parity rule type 1

If the two DNA strands experience on average the same mutational and repair mechanisms, the substitution rates are expected to 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) [30]. PR1 is very well verified at the genome scale. For example, in the human genome, although substitution rates can vary over a large range of values (*e.g.*, the transition $C \to T$ is threefold higher than the transversion $C \to G$), reverse complementary substitution rates are nearly equal (data not shown).

2.1.2 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 computed on a given strand: $[G] \sim [C]$ and $[T] \sim [A]$. This second symmetry law is known as Parity rule type 2 (PR2) [30,31]. Like PR1, PR2 is very well verified at the chromosomal scale. For example, in mammalian genomes, the $G+C$ content ($\theta_{GC} = [G]+[C]$) can vary over a large range of values (*e.g.* from 36% to 50% for the 22 human autosomes), nevertheless the frequencies of complementary nucleotides are nearly equal [30, 32]. Actually PR2 formally derives from PR1: under symmetrical substitution rates (PR1), the DNA composition should verify PR2 [32,33].

Let us point out that PR1 and PR2 are indeed approximate symmetries. If they are well verified at the chromosomal scale, some systematic deviations can be observed at finer scales [34]. 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.

2.2 Transcription

2.2.1 Transcription is a strand-asymmetric process

During the transcription of a gene (fig. 3A), 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 [14]. The coding strand and the transcribed strand could undergo different mutational and repair events that generate strand asymmetry [35–40]. 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 [35,37,39–42] than the transcribed strand. It has also been proposed that repair mechanisms [35,38,40,43] could generate strand asymmetries. A mechanism known as transcription-coupled repair (TCR) [14], associated with the passage of the RNA polymerase, preferentially repairs towards the coding strand [44]. Strand asymmetry associated to transcription has been observed across the whole life tree: in several bacterial strains [37,43], in human [26, 38–40,45,46] and in many other eukaryotes [27].

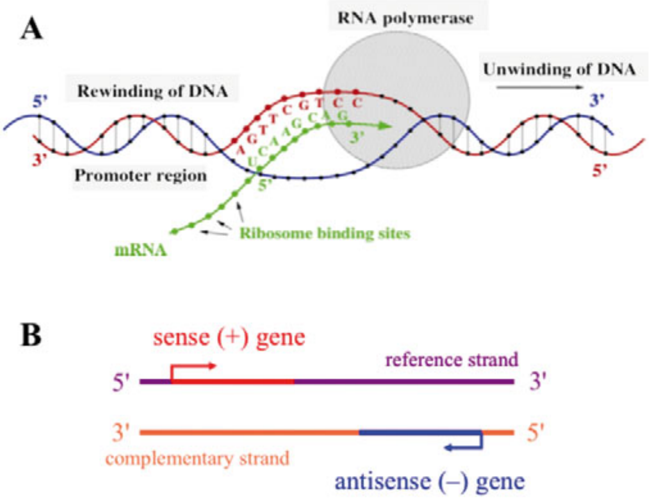

Fig. 3. (Colour on-line) Schematic view of transcription. (A) Schematic representation of transcription elongation. During the transcription elongation, *i.e.* the progression of the RNA polymerase protein complex, this enzyme catalyzes the unwinding of the DNA double helix on a little more than one turn of the helix (one turn corresponds to 10-11 base pairs). The template strand is exposed to the synthesis of a complementary RNA chain, in the $5' \to 3'$ direction. This nascent RNA is first hybridized to the template DNA and further separated from its template, while the two DNA strands reform the double helix during the progression of the "transcription bubble". (B) Definition of sense (+) and antisense (−) genes.

2.2.2 Gene orientation and transcription rate as natural parameters to describe transcription-associated strand asymmetry

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. 3B). For a sense (+) gene the reference strand is the coding

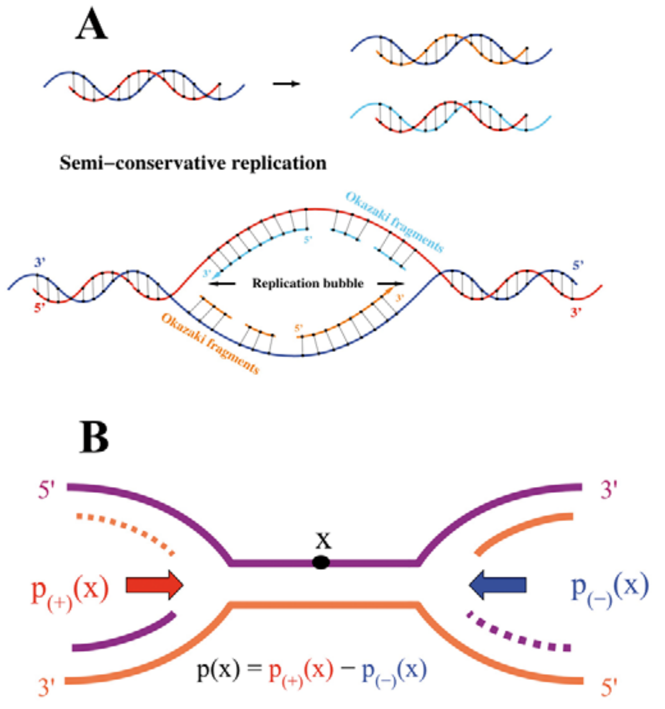

Fig. 4. (Colour on-line) Schematic representation of replication initiation. (A) Replication initiates at local regions on the genomes called replication origins (there is a unique origin in most eubacterial chromosomes and many origins in eukaryotic genomes). At each replication origin position, the two strands of the parental DNA double helix are separated from each other to serve as templates for the synthesis of the two daughter strands. On each extremity of the resulting "replication bubble", two replication forks are formed: one moves rightward with the leading (bottom) and lagging (top) strands; the other moves leftward with the leading (top) and lagging (bottom) strands. All strands are synthesized in the $5' \to 3'$ direction. It results that the DNA synthesized on the leading strand is made continuously while the lagging strand is made discontinuously, *i.e.* as successive short fragments called Okazaki fragments. (B) Definition of replication fork polarity $p(x)$. The purple strand is the reference strand. For newly synthesized strands, continuous (respectively, dashed) lines correspond to leading (respectively, lagging) strands.

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.

2.3 Replication

2.3.1 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. 4A): 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 [14]. 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. 4) 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 main DNA polymerases that replicate separately the two parental strands. 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*).

Replication could induce strand asymmetries by several means [35,36,47–49]. 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 [35,36]. In eukaryotes, the leading and lagging strands are presumably synthesized by two distinct DNA polymerases [50]. Strand asymmetries could result from the different error spectra of the two DNA polymerases [49]. Strand asymmetry associated to replication was observed across the whole life tree: in several bacterial strains [11,12,47,48,51,52], in viruses [12,53], in yeast [54], in mitochondria [55], and in human [13,18,19,49]. Very convincing demonstrations were reported recently in human [13] and *S. cerevisiae* [56].

2.3.2 Replication fork polarity as the natural parameter to describe replication-associated strand asymmetry

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 corresponding to the DNA sequence used. 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. 4B). 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 replication origin that fired downstream ($3'$ direction of the reference strand). For a sense $(+)$ fork (fig. 4B), the reference strand is the leading strand whereas the complementary strand is the lagging strand. For an antisense $(-)$ fork (fig. 4B) 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. Due to the intrinsic stochasticity of the replication program, the locus x will be replicated by a proportion $p_{(\pm)}(x)$ of $(\pm)$ forks over cell cycles. 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*:

$$p(x) = p_{(+)}(x) - p_{(-)}(x). \tag{2}$$

We define the replication fork polarity for a locus x, but it can be equally defined for a genomic region by averaging $p(x)$ over space. When, as in the bacterial replicon model (fig. 1) [57] the replication fork polarity $p = +1$ (respectively, $p = -1$), the reference strand only undergoes leading (respectively, 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.

2.4 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 forces that determine fixation or disappearance of a mutation [58]. Random genetic drift corresponds to the stochastic variation of a mutation frequency, due to the random sampling of alleles [58]. 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 [58,59] (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) [58]. The fixation probability can also be affected by neutral processes such as biased gene conversion [60–62]: gene conversion, a common event during meiosis recombination, is biased towards the fixation of $G + C$ rich alleles. When selection plays no role at a locus, the site is said to evolve neutrally.

Remark. In the following study of substitutional asymmetry, we will only consider sites that can be considered as neutral.

3 Modelling substitutional asymmetry

3.1 Minimal model

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 dependence of substitution rates upon the replication fork polarity p (sect. 2.3.2), gene orientation and transcription rate α (sect. 2.2.2). First of all, the model has to respect *strand-exchange symmetry*. For example, 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 $-p$:

$$\tau[\xi, p, \alpha, (+)] = \tau^c[\xi, -p, \alpha, (-)], \tag{3}$$

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). In fact, 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:

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

$$\tau^a[\xi, p, \alpha, (+)] = -\tau^a[\xi, -p, \alpha, (-)]. \tag{5}$$

Hereafter, we will forget about the "ξ dependence" to focus only on the effect of gene orientation (±), 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.

3.1.1 Substitution rates in intergenic regions

In our minimal model, substitution rates in intergenic regions are given by

$$\tau_{\text{intergenic}}[p] = \tau_0^s + p_{(+)}\tau_R + p_{(-)}\tau_R^c. \tag{6}$$

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. (3)), the passage of an antisense (−) fork contributes by the reverse complementary substitution rate τ_R^c. We assume that mutational and fixation 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.

3.1.2 Substitution rates in genic regions

In genic regions, we propose to model the net effect of transcription by

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

$$\tau_{\text{genic }(-)}[p, \alpha] = \tau_{\text{intergenic}}[p] + \tau_T^c[\alpha]. \tag{8}$$

The reverse complementary coefficient τ_T^c appears in antisense gene due to strand exchange symmetry (eq. (3)). 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 [63], producing non-coding transcripts [64–68]. Non-coding transcripts could generate strand asymmetries in intergenic regions not associated to replication [69]. In our model non-coding transcripts are not taken into account, and we will always assume that substitutional asymmetry in intergenic regions is mainly due to replication [13].

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

For the model defined by eqs. (6) to (8), 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:

$$\tau^s[p, \alpha, (\pm)] = \tau_0^s + \tau_R^s + \tau_T^s[\alpha], \tag{9}$$

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:

$$\tau^a[p, \alpha, (\pm)] = p\tau_R^a \pm \tau_T^a[\alpha], \tag{10}$$

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

3.2 Molecular mechanisms

Various molecular mechanisms were proposed to explain strand-asymmetry in the human genome As briefly reviewed in this section, most of them actually reduce to the minimal model presented in sect. 3.1.

Remark. In the following we will confound substitutions with mutations to make the discussion easier to follow. The relationship between substitution and mutation rate is direct [59] if there are no (neutral or selective) fixation bias. To our knowledge, no concrete neutral fixation bias were proposed to generate strand asymmetry, but a fixation bias can modulate the strength of the substitutional asymmetry.

3.2.1 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 [50]. In the human genome, the substitutional asymmetries associated to replication were proposed to result from the different error spectra of the two DNA polymerases [13, 49]. 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 [49]. For clarity let us call pol ϵ (respectively, pol δ) the leading (respectively, 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} (respectively, 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 (respectively, δ) polymerase. For simplicity we assume that the mispaired base $j : i^c$ will persist until the next replication round, where the mispaired base results in the $i \to j$ substitution in 50% of the cases. 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 an 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:

$$(+) \quad \text{fork} \quad \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{\Sigma_{ji}} \begin{matrix} j \\ i^c \end{matrix} \xrightarrow{\frac{1}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1}{2}} \begin{matrix} i \\ j^c \end{matrix} \xleftarrow{\Delta^c_{ji}} \begin{matrix} i \\ i^c \end{matrix}, \tag{11}$$

$$(-) \quad \text{fork} \quad \begin{matrix} i \\ i^c \end{matrix} \xrightarrow{\Delta_{ji}} \begin{matrix} j \\ i^c \end{matrix} \xrightarrow{\frac{1}{2}} \begin{matrix} j \\ j^c \end{matrix} \xleftarrow{\frac{1}{2}} \begin{matrix} i \\ j^c \end{matrix} \xleftarrow{\Sigma^c_{ji}} \begin{matrix} i \\ i^c \end{matrix}, \tag{12}$$

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

$$p_{(+)} \frac{(\Sigma + \Delta^c)_{ji}}{2} + p_{(-)} \frac{(\Sigma^c + \Delta)_{ji}}{2}. \tag{13}$$

We recover our minimal model eq. (6) with

$$\tau_R = \frac{\Sigma + \Delta^c}{2}. \tag{14}$$

The substitutional asymmetry associated to replication is then given by

$$p\tau_R^a, \qquad \text{with} \qquad \tau_R^a = \frac{\Sigma^a - \Delta^a}{2}, \tag{15}$$

in agreement with eq. (10).

3.2.2 Other examples of molecular mechanisms

In this subsection, we just list other molecular mechanisms that reduce to our minimal model (eqs. (9) and (10)). We refer the reader to A. Baker's thesis [70] for more detailed discussion.

Cytosine deamination in single-stranded DNA. Cytosine can spontaneously deaminates into uracil. After two replication rounds the uracil can become a thymine ($U:G \rightarrow U:A \rightarrow T:A$) and the cytosine deamination results in a proper $C \rightarrow T$ substitution. The cytosine deamination is much more frequent (140 fold) in single-stranded DNA (ssDNA) than in double-stranded DNA [71]. 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 [36], in mitochondria [55], and recently in human [13,49]. Similarly the coding strand is transiently in ssDNA during transcription and could undergo an excess of cytosine deamination [41,42]. This process was proposed to explain the strand asymmetry observed in *E. coli* genes [37], and in human genes [40, 39]. The cytosine deamination theory predicts positive $(C \rightarrow T)^a$ asymmetries for both replication [36,49] and transcription [37,40].

Transcription-induced mutations. Along with the cytosine deamination, other types of mutagenic reactions can be considered. Mugal *et al.* [40] 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 $C \rightarrow T$, $A \rightarrow G$, $G \rightarrow T$ and $Y \rightarrow T$ substitutions. During transcription, the coding strand is possibly more exposed to those mutagenic reactions [41,42]. These mutagenic reactions lead to an increase of $C \rightarrow T$, $A \rightarrow G$, $G \rightarrow T$ and $A \rightarrow T$ substitution rates on the coding strand compared to the flanking intergenic regions, as observed in the human genome by Mugal *et al.* [40]. Under transcription-induced mutations alone, those substitution rates should be the same in the transcribed strand and in the flanking intergenic region.

Transcription-coupled repair. Transcription-coupled repair (TCR), see [44] for review, has also been proposed to generate strand asymmetries [38,40,43]. 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 [44]. TCR can therefore reduce rates of mutagenic reactions on the transcribed strands. Mugal *et al.* [40] proposed that the $C \rightarrow T$, $A \rightarrow G$, $G \rightarrow T$ and $A \rightarrow T$ substitution rates were lower in the transcribed strand than in the flanking intergenic region due to TCR, or other repair mechanisms.

TCR acting on misinserted bases. Green *et al.* [38] 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 [44]. 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. This molecular mechanism leads to an asymmetrical rate τ^a that still satisfies our minimal model eq. (10). However an additional contribution, proportional to the replication fork polarity times the gene orientation, is found in the symmetrical rate τ^s eq. (9).

Remark. 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.

3.3 Analysis of substitution rates in the human genome

To demonstrate the validity and pertinence of the minimal model described in sect. 3.1, we determined the substitution rates in the human genome as explained in appendix A.3. The substitution rates were tabulated in the human lineage since the divergence with chimpanzee [13]. Substitution rates were computed separately in genic (+), intergenic and genic (−) regions of given replication fork polarity values estimated from the mean replication timing (MRT) according to eq. (1). As a substitute to germline replication fork polarity, we used the replication fork polarity determined in HeLa cell line, where the replication fork velocity $v = 0.64\,\mathrm{kbp/min}$ has been measured by DNA combing and where the S phase duration was estimated to be $T_S \sim 7\,\mathrm{h}$ [29]. The conservation of the replication fork polarity profile across differentiation will be addressed in sect. 6.

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

As shown in fig. 5, 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 (+) (respectively, (−)) regions, we recover the same linear behaviour adding up (respectively, subtracting down) a constant corresponding to the transcription-associated asymmetry. The substitutional asymmetry τ^a is therefore consistent with the following model:

$$\tau^a = \begin{cases} p\tau_R^a + \tau_T^a & \text{genic } (+), \\ p\tau_R^a & \text{intergenic}, \\ p\tau_R^a - \tau_T^a & \text{genic } (-), \end{cases} \qquad (16)$$

in agreement with the minimal model for substitutional asymmetry proposed in sect. 3.1 (eq. (10)).

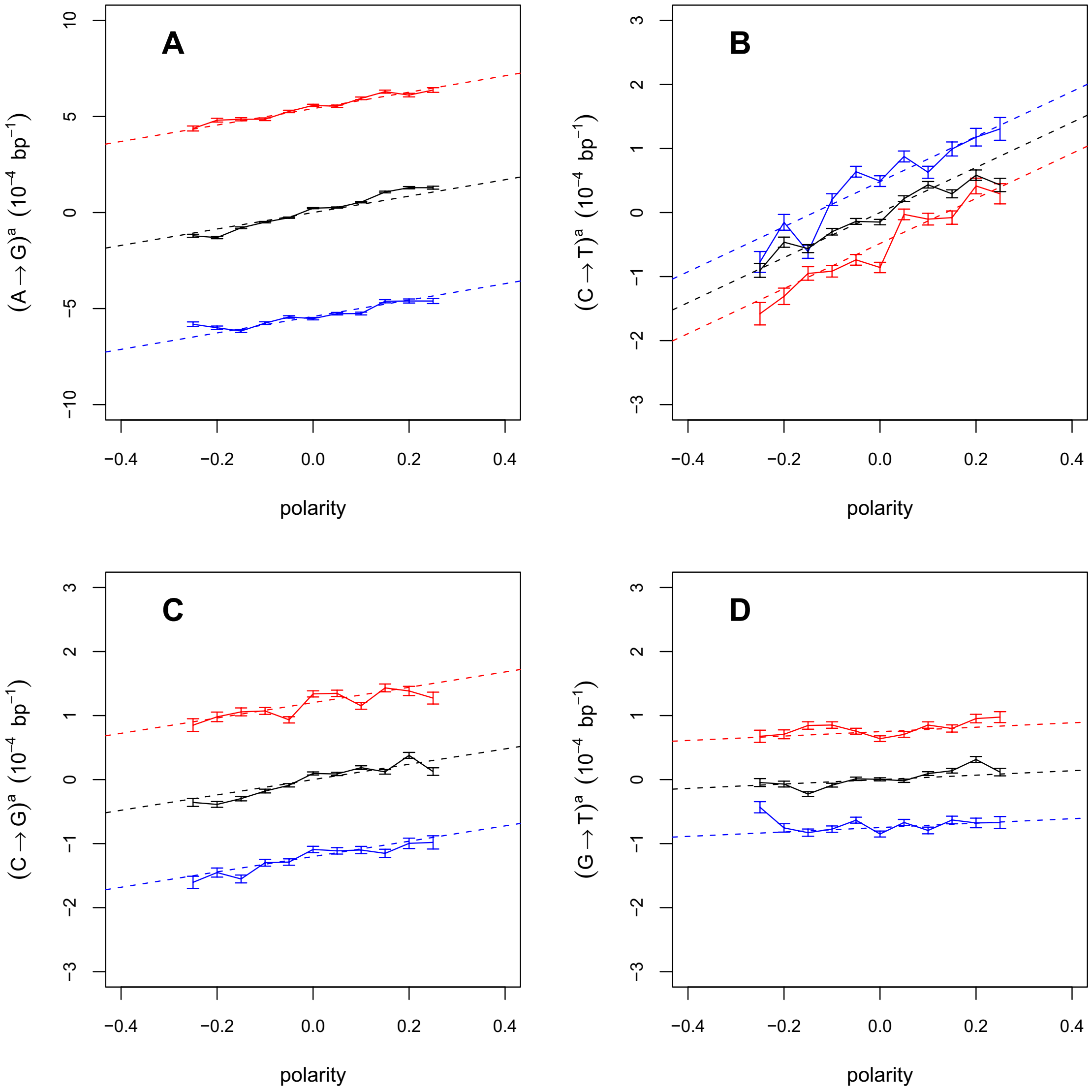

Fig. 5. (Colour on-line) Substitutional asymmetry *versus* replication fork polarity (determined in HeLa cell line using eq. (1) and replication timing data from [29]) in genic sense (red), intergenic (black), and genic antisense (blue) regions, for (A) the $A \rightarrow G$ substitution, (B) the $C \rightarrow T$ substitution, (C) the $C \rightarrow G$ substitution, and (D) the $G \rightarrow T$ 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. (16). The linear regression coefficients are reported in table 1.

Table 1. Transcription- and replication-associated substitutional asymmetries. Coefficients τ_R^a, τ_T^a of the linear model eq. (16), obtained by least-squares fits to a line in fig. 5.

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

The coefficients τ_T^a and τ_R^a, estimated by least-squares fits to a line (dashed lines in fig. 5), 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 $\mathrm{d\,MRT}/\mathrm{d}x$) 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 $\mathrm{d\,MRT}/\mathrm{d}x$ which is a strand-

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 1 Mbp windows genome wide. For substitution rates we only retained intergenic nucleotides. Only 1 Mbp 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 $= 3 \cdot 10^{-5}$).

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

asymmetric variable. On the opposite the symmetrical substitution rates highly correlate with the MRT [72,28], but not with d MRT/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.* [40,73] reported that the substitutional asymmetry correlates strongly with the relative distance to skew N-domains borders (fig. 2B). In our current perspective, the relative distance to N-domains borders is directly related to the replication fork polarity in the germline [74]. So far the substitutional asymmetry follows closely the model proposed in sect. 3.1 (eq. (10)): 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 [13,40,49,73]. 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).

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

When computed genome wide, the average symmetrical substitution rates in genic regions were found to be lower than the corresponding rates in intergenic regions (data not shown). 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 [72, 28], 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. 6, 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. For the $C \to T$ (fig. 6B) and $G \to T$ (fig. 6D) 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 $(G \to T)^s_T < 0$ observed genome wide. The variation of the $A \to G$ rate (fig. 6A) and the $C \to G$ rate (fig. 6C) are compliant with, but not demonstrative of, the weak $(A \to G)^s_T < 0$ and $(C \to G)^s_T < 0$ observed genome wide [70]. 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 [38].

Remark. We first note on fig. 6A that there is a strong $(A \to G)^a_T > 0$ asymmetry (the purple curve is above the orange one), which extends on the whole transcript, as previously observed [38,39]. Similarly significant $(C \to G)^a_T > 0$ and $(G \to T)^a_T > 0$ asymmetries are observed along the whole transcript (figs. 6C and D). In contrast, no significant asymmetry is observed for the $C \to T$ substitution rate (fig. 6B). 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 [39]. The regional variation of substitution rates observed in fig. 6 thus confirms the transcription-associated substitutional asymmetries observed genome wide [70]. Let us note a residual $(A \to G)^a_T > 0$ asymmetry in the flanking intergenic region in fig. 6A. 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 tenfold lower than the asymmetry observed inside the gene, which suggests that unannotated transcripts are not numerous enough to affect our previous observations.

4 DNA composition evolution

4.1 General formalism

In the case of neighbor-independent (for neighbor-dependent see paper II) and time homogeneous substitutions, the time evolution of the DNA composition is given by [58]

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t) = M X(t), \tag{17}$$

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. (17) is called

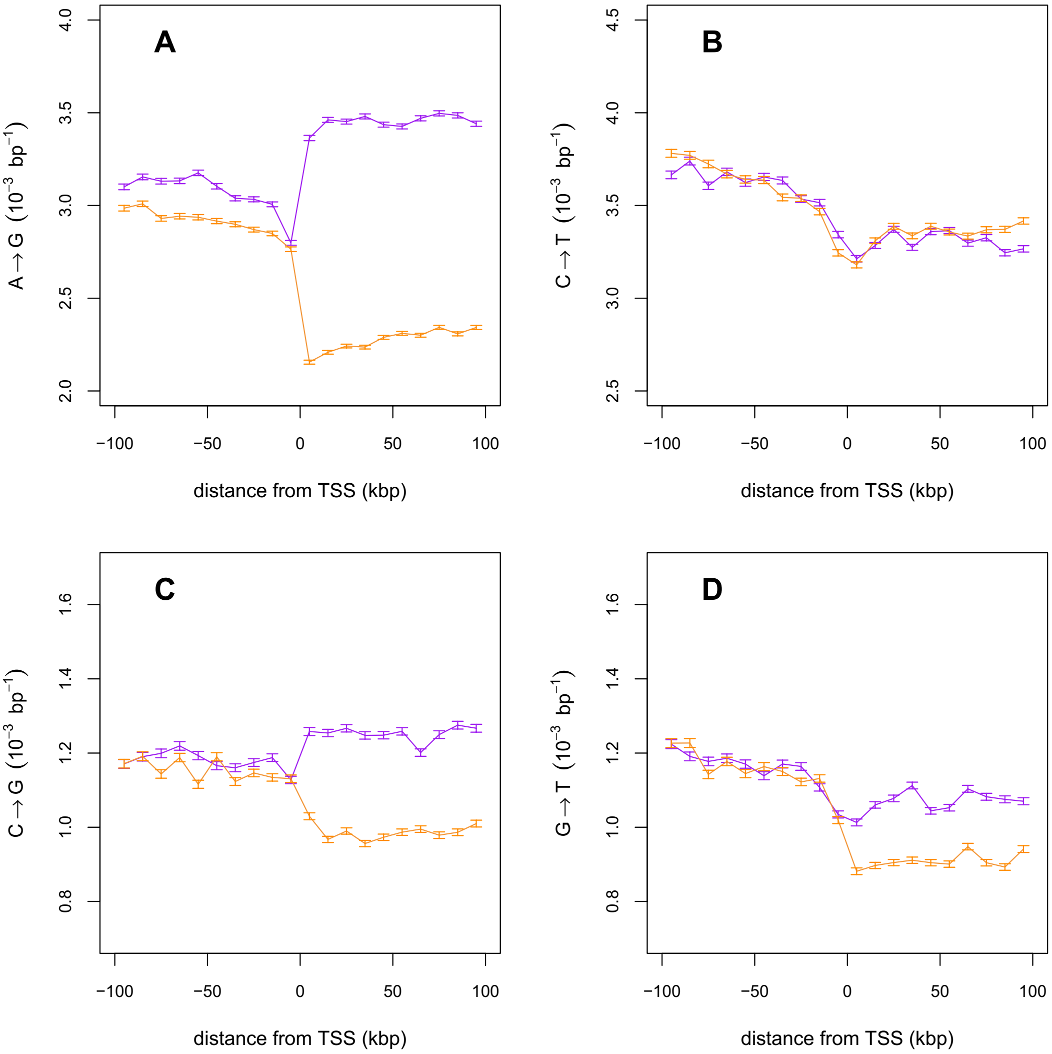

Fig. 6. (Colour on-line) Substitution rates along large (> 100 kbp) human genes. Average substitution rates in large human genes were computed every 10 kbp from 100 kbp upstream to 100 kbp downstream of the TSS. As genes are larger than 100 kbp, data points at 0 kbp < distance to TSS < 100 kbp correspond 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 sect. 2.2). (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 sect. 2.2). (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 master equation; it is the time continuous formulation of a Markov chain. The general properties of a Markov chain are well known [75], including the evolution towards equilibrium.

4.1.1 Time evolution of the composition

First, we can easily integrate eq. (17) to get the composition $X(t)$ at any time t, knowing the initial composition $X(t_0)$ at a time t_0

$$X(t) = W(t,t_0)X(t_0), \qquad \text{where} \qquad W(t,t_0) = e^{(t-t_0)M}. \tag{18}$$

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

$$MX^* = 0 \qquad \text{and} \qquad \sum_i X_i^* = 1. \tag{19}$$

So X^* is an eigenvector of M with eigenvalue 0. The three other eigenvalues ($k = 1$ to 3) have all a strictly negative real part

$$MX^{(k)} = \left[-\frac{1}{\lambda^{(k)}} + i\omega^{(k)}\right] X^{(a)}, \qquad \text{with} \qquad \lambda^{(k)} > 0. \tag{20}$$

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 [75]. There are some exceptional cases where eqs. (19) and (20) are not verified, but such cases are never encountered in DNA composition evolution and therefore not relevant for our purpose. It follows from eqs. (19) and (20) that the composition $X(t)$ converges asymptotically towards the equilibrium value X^*, whatever the initial composition $X(t_0)$:

$$X(t) = e^{M(t-t_0)}X(t_0) \to X^*, \qquad \text{when} \qquad t \to \infty. \tag{21}$$

4.1.2 Exploiting strand exchange symmetry

The time evolution on the complementary strand is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}X^c(t) = M^c X^c(t), \tag{22}$$

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 (sect. 2). 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 $X^c_i(t) = X_{i^c}(t)$ and $M^c_{ij} = M_{i^c j^c}$. We can decompose M into a symmetrical and an asymmetrical part under strand exchange symmetry, $M = M^s + M^a$ with

$$M^s = \frac{M + M^c}{2}, \qquad M^a = \frac{M - M^c}{2}. \tag{23}$$

It is more convenient to consider the evolution of DNA composition through the following variables [33]:

$$Y = \begin{pmatrix} \theta \\ S \end{pmatrix} = \begin{pmatrix} \theta_{TA} \\ \theta_{GC} \\ S_{TA} \\ S_{GC} \end{pmatrix} = UX = \begin{pmatrix} X_T + X_A \\ X_G + X_C \\ X_T - X_A \\ X_G - X_C \end{pmatrix}, \tag{24}$$

where S_{TA} and S_{GC} are the compositional skews and θ_{TA} and θ_{GC} are the $T + A$ and $G + C$ contents. The $T + A$ and $G + C$ 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

$$U = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \qquad U^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix}. \tag{25}$$

It is easy to get the evolution of Y through a linear transformation of eq. (17)

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = NY(t), \qquad \text{with} \qquad N = UMU^{-1}. \tag{26}$$

Similarly it is straightforward to get the equilibrium composition Y^* through a linear transformation of eq. (19)

$$NY^* = 0 \qquad \text{and} \qquad \theta^*_{TA} + \theta^*_{GC} = 1. \tag{27}$$

The symmetry properties of M^s and M^a imply the following block forms for [33]:

$$N^s = UM^sU^{-1} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \tag{28}$$

$$N^a = UM^aU^{-1} = \begin{pmatrix} 0 & B \\ C & 0 \end{pmatrix}. \tag{29}$$

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:

$$M^s = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \mu & \nu & \kappa & \epsilon \\ \nu & \mu & \epsilon & \kappa \end{pmatrix}, \qquad M^a = \begin{pmatrix} a & -b & c & -d \\ b & -a & d & -c \\ m & -n & k & -e \\ n & -m & e & -k \end{pmatrix}. \tag{30}$$

The matrices A and D introduced in eq. (28) are equal to

$$A = \begin{pmatrix} \alpha + \beta & \gamma + \delta \\ \mu + \nu & \kappa + \epsilon \end{pmatrix}, \qquad D = \begin{pmatrix} \alpha - \beta & \gamma - \delta \\ \mu - \nu & \kappa - \epsilon \end{pmatrix}, \tag{31}$$

and the matrices B and C introduced in eq. (29) are equal to

$$B = \begin{pmatrix} a + b & c + d \\ m + n & k + e \end{pmatrix}, \qquad C = \begin{pmatrix} a - b & c - d \\ m - n & k - e \end{pmatrix}. \tag{32}$$

Following [61], the coefficients of the matrix A can also be expressed as substitution rates between weak ($\mathbb{W} = A, T$) and strong ($\mathbb{S} = G, C$) nucleotides:

$$\mu + \nu = (T \to G)^s + (T \to C)^s = (\mathbb{W} \to \mathbb{S}), \tag{33}$$

$$\gamma + \delta = (G \to T)^s + (G \to A)^s = (\mathbb{S} \to \mathbb{W}). \tag{34}$$

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

$$A\theta_A = 0, \qquad \text{with} \qquad \theta_A = \frac{1}{\mu + \nu + \gamma + \delta}\begin{pmatrix} \gamma + \delta \\ \mu + \nu \end{pmatrix}, \tag{35}$$

and

$$A\begin{pmatrix} 1 \\ -1 \end{pmatrix} = -\frac{1}{\lambda_A}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \text{with} \quad \lambda_A = \frac{1}{\mu + \nu + \gamma + \delta}. \tag{36}$$

Expressed in terms of weak to strong and strong to weak substitution rates, the $G+C$ component of θ_A and λ_A are equal to [61]

$$\theta_{A,GC} = \frac{(\mathbb{W} \to \mathbb{S})}{(\mathbb{W} \to \mathbb{S}) + (\mathbb{S} \to \mathbb{W})}\,, \tag{37}$$

$$\lambda_A = \frac{1}{(\mathbb{W} \to \mathbb{S}) + (\mathbb{S} \to \mathbb{W})}\,. \tag{38}$$

The two eigenvalues of D ($k = 1, 2$) have a strictly negative real part

$$DS^{(k)} = \left[-\frac{1}{\lambda_D^{(k)}} + i\omega^{(k)}\right] S^{(k)}, \qquad \text{with} \qquad \lambda_D^{(k)} > 0. \tag{39}$$

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 λ_A and $\lambda_D^{(1,2)}$ are characteristic time scales of the DNA composition evolution, when the substitution rate matrix satisfies PR1.

Proof. As $\sum_i M^s_{ij} = 0$ and for $i \neq j, M^s_{ij} > 0$, we know according to eqs. (19) and (20) 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 $-\frac{1}{\lambda_A}$ are the two eigenvalues of A, the two remaining eigenvalues of N^s, those of D, have a strictly negative real part.

4.2 Strand symmetry

4.2.1 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 [30]. It implies in turn that the matrices B and C are null (eq. (29)). The equilibrium $T+A$ and $G+C$ contents and the equilibrium skews satisfy

$$A\theta^* = 0 \qquad \text{and} \qquad DS^* = 0, \tag{40}$$

with the constraint $\theta^*_{TA} + \theta^*_{GC} = 1$ (eq. (27)). According to the spectral properties of A and D derived just above (eqs. (35) and (39)), the solutions of eq. (40) are

$$\theta^* = \theta_A \qquad \text{and} \qquad S^* = 0. \tag{41}$$

The equations of evolution for the $T+A$ and $G+C$ contents and the compositional skews are respectively given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = A\theta(t), \tag{42}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = DS(t), \tag{43}$$

whose solutions are

$$\theta(t) = e^{A(t-t_0)}\theta(t_0) \to \theta_A, \quad \text{when} \quad t - t_0 \gg \lambda_A, \tag{44}$$

$$S(t) = e^{D(t-t_0)}S(t_0) \to 0, \quad \text{when} \quad t - t_0 \gg \lambda_D^{(1)}, \lambda_D^{(2)}. \tag{45}$$

We recover the result of Lobry [32]: if the substitution rate matrix is symmetrical (PR1), then the compositional skews are null at equilibrium (PR2):

$$\text{if} \quad M = M^s \text{ (PR1)} \quad \text{then} \quad S^*_{TA} = S^*_{GC} = 0 \text{ (PR2)}. \tag{46}$$

The $T+A$ and $G+C$ contents converge exponentially toward their equilibrium values θ_A with the characteristic time scale λ_A. More explicitly the evolution of the $G+C$ content[1] is given by

$$\theta_{GC}(t) = e^{-\frac{(t-t_0)}{\lambda_A}}\theta_{GC}(t_0) + \left(1 - e^{-\frac{(t-t_0)}{\lambda_A}}\right)\theta^*_{GC}. \tag{47}$$

Hence the half-time $t_{1/2}$ of the $G+C$ content evolution, defined as the time necessary to divide by two the difference between the $G+C$ content and the equilibrium $G+C$ content, is given by [61]

$$\frac{\theta^*_{GC} - \theta_{GC}(t)}{\theta^*_{GC} - \theta_{GC}(t_0)} = e^{-\frac{(t-t_0)}{\lambda_A}} = \frac{1}{2}, \tag{48}$$

$$\text{for} \quad t - t_0 = t_{1/2} = \ln 2\, \lambda_A. \tag{49}$$

The compositional skews decay towards zero with two time scales $\lambda_D^{(1)}$ and $\lambda_D^{(2)}$. More precisely, the projections of the compositional skew $S(t)$ onto the eigenvectors $S^{(1)}$ and $S^{(2)}$ of D (eq. (39)) decay exponentially with respective characteristic time scales $\lambda_D^{(1)}$ and $\lambda_D^{(2)}$, and the corresponding half times are given by $\ln 2\, \lambda_D^{(1)}$ and $\ln 2\, \lambda_D^{(2)}$.

4.2.2 Numerical simulations

In fig. 7, we illustrate the time evolution under PR1, using the symmetrical substitution rate matrix $M = M^s_0 + M^s_R$ estimated in the human genome as explained in appendix B (eq. (B.1)). To express substitution rates in per bp per Myrs units, we used 5 Myrs as an estimation of the human-chimpanzee divergence. As predicted by eq. (44), the $T+A$ and $G+C$ contents converge towards their equilibrium values whatever their initial values (figs. 7A and B). The equilibrium $G+C$ content is equal to $\theta^*_{GC} = 44\%$ and the characteristic time scales are equal to $\lambda_A = 558$ Myrs (corresponding half-time $t_{1/2} = 387$ Myrs), $\lambda_D^{(1)} = 555$ Myrs and $\lambda_D^{(2)} = 1360$ Myrs (corresponding half-times 385 Myrs and 943 Myrs). The dynamics of the $G+C$ content and the skews are therefore extremely slow. As predicted by eq. (45), the TA and GC skews decay towards 0 whatever their initial values (fig. 7C and D).

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

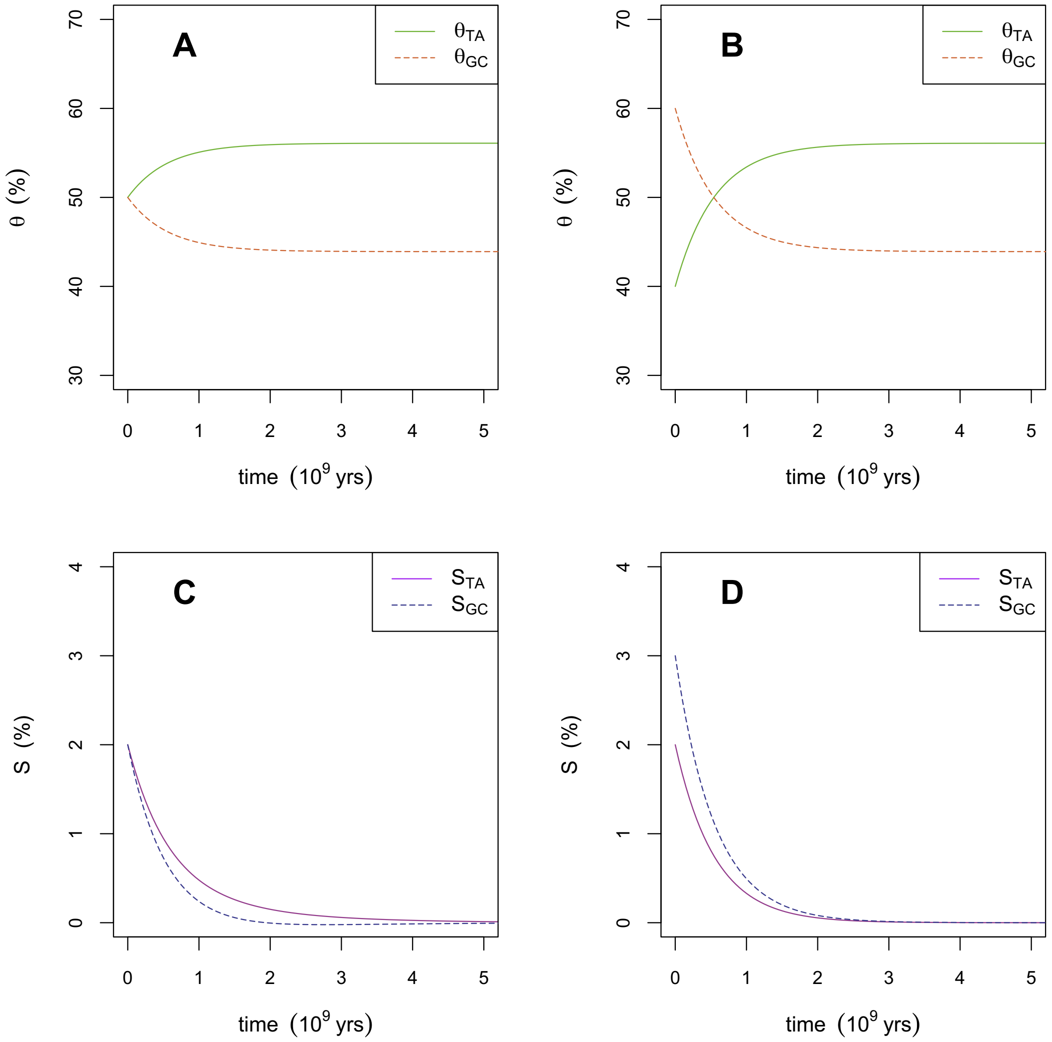

Fig. 7. (Colour on-line) DNA composition evolution under PR1 satisfies PR2 asymptotically. The substitution rate matrix is symmetric $M = M_0^s + M_R^s$ (eq. (B.1)). Time evolution of the $T+A$ and $G+C$ contents with the initial conditions (A) $\theta_{TA}(0) = \theta_{GC}(0) = 50\%$, and (B) $\theta_{TA}(0) = 40\%$ 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 $S_{GC}(0) = 3\%$.

Remark on the $G+C$ content evolution. The symmetrical substitution rates (and therefore the $G+C$ 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 [61], or replication timing [28,72]. The value found for $G+C^*$ (44%) corresponds to the highest values found in [61], for reasons currently unclear. In the BGC model, the half-time $t_{1/2}$ strongly depends on the recombination rate and on the effective population size. In the absence of recombination, the BGC model predicts $t_{1/2} \sim 470\,$Myrs, and the $G+C$ content evolution is extremely slow [61]. But in genomic region of high recombination rate, for species with large effective population size, the $G+C$ content evolution is predicted to be much faster $t_{1/2} \sim 62\,$Myrs [61]. Our value for $t_{1/2}$ (387 Myrs) therefore corresponds to an intermediate value between the two extremes proposed by the BGC model.

4.3 Perturbative analysis of the compositional asymmetry

4.3.1 Strand asymmetry establishment

When the PR1 symmetry is broken $M^a \neq 0$, the matrices B and C are no longer null (eq. (29)). The equilibrium $T+A$ and $G+C$ contents and the equilibrium skews are now solutions of the equations:

$$A\theta^* + BS^* = 0 \qquad \text{and} \qquad C\theta^* + DS^* = 0, \tag{50}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = A\theta(t) + BS(t), \tag{51}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = C\theta(t) + DS(t). \tag{52}$$

How are the time evolutions of the skews and the $G+C$ content affected by the substitutional asymmetry? How are their equilibrium values modified? Is PR2 still verified? As reported in appendix B, in the human genome the substitutional asymmetry is actually very small as compared to its substitutional symmetric counterpart. More precisely for the values reported in appendix B, we have

$$|\tau_R^a| = \epsilon|\tau_0^s + \tau_R^s|, \qquad \text{with} \qquad \epsilon \leq 0.14, \tag{53}$$

$$|\tau_T^a| = \epsilon|\tau_0^s + \tau_R^s|, \qquad \text{with} \qquad \epsilon \leq 0.17. \tag{54}$$

This justifies the use of perturbation theory to describe the skew evolution.

4.3.2 Perturbative analysis

Let us illustrate the principles of perturbation theory on the time evolution of the composition $X(t)$ governed by eq. (17). 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. (53) and (54)):

$$X(t) = X^{(0)}(t) + \epsilon X^{(1)}(t) + \epsilon^2 X^{(2)}(t) + \dots. \tag{55}$$

We have then to solve the time evolution eq. (17) order by order in ϵ, considering M^a of order ϵ. Explicitly we have to solve the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}X^{(0)}(t) = M^s X^{(0)}(t), \tag{56}$$

at order zero and the differential equation

$$\epsilon^n \frac{\mathrm{d}}{\mathrm{d}t}X^{(n)}(t) = \epsilon^n M^s X^{(n)}(t) + \epsilon^{n-1} M^a X^{(n-1)}(t), \tag{57}$$

at order $n \geq 1$. The solutions of these differential equations are:

$$X^{(0)}(t) = e^{M^s(t-t_0)}X(t_0), \tag{58}$$

$$\epsilon X^{(1)}(t) = \int_{t_0}^{t} \mathrm{d}t_1 e^{M^s(t-t_1)} M^a e^{M^s(t_1-t_0)} X(t_0), \tag{59}$$

$$\epsilon^2 X^{(2)}(t) = \int_{t_0}^{t} \mathrm{d}t_2 e^{M^s(t-t_2)} M^a \times \int_{t_0}^{t_2} \mathrm{d}t_1 e^{M^s(t_2-t_1)} M^a e^{M^s(t_1-t_0)} X(t_0), \tag{60}$$

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

$$X(t) = e^{M^s(t-t_0)}X(t_0) + \int_{t_0}^{t} \mathrm{d}u\, e^{M^s(t-u)} M^a e^{M^s(u-t_0)} X(t_0) + O(\epsilon^2). \tag{61}$$

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 $M^a \neq 0$.

In our minimal model, the asymmetrical part M^a follows the decomposition eq. (10) yielding the same decomposition for the matrices B and C, and the symmetrical part M^s follows the decomposition eq. (9) yielding the same decomposition for the matrices A and D. Below, we give the results of the pertubative analysis considering the matrices $M_T^s[\alpha]$, $M_T^a[\alpha]$ and M_R^a as small perturbations of order $O(\epsilon)$ to the symmetrical matrix $M_0^s + M_R^s$.

The skew can be decomposed into transcription- and replication-associated components. Equation (26) gives the following time evolution of the TA and GC skews:

$$S[p,\alpha,(\pm)](t) = pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2), \tag{62}$$

where

$$S_R(t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, C_R\, \tilde{\theta}_0(u), \tag{63}$$

$$S_T[\alpha](t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[D_0+D_R](t-u)}\, C_T[\alpha]\, \tilde{\theta}_0(u). \tag{64}$$

The pertubative resolution of eq. (27) gives the following equilibrium TA and GC skews:

$$S^*[p,\alpha,(\pm)] = pS_R^* \pm S_T^*[\alpha] + O(\epsilon^2), \tag{65}$$

where

$$S_R^* = -[D_0 + D_R]^{-1}\, C_R\, \tilde{\theta}_0^*, \tag{66}$$

$$S_T^*[\alpha] = -[D_0 + D_R]^{-1}\, C_T[\alpha]\tilde{\theta}_0^*. \tag{67}$$

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. (10)). The former is proportional to the replication fork polarity whereas the latter increases in magnitude with the transcriptional rate and changes sign with gene orientation.

Weak impact on the $T+A$ and $G+C$ contents. The perturbative resolution of eq. (26) gives the following time evolution of the $T+A$ and $G+C$ contents:

$$\theta[p,\alpha,(\pm)](t) = \tilde{\theta}_0(t) + \theta_T[\alpha](t) + O(\epsilon^2), \tag{68}$$

where

$$\tilde{\theta}_0(t) = e^{[A_0+A_R](t-t_0)}\, \theta(t_0), \tag{69}$$

$$\theta_T[\alpha](t) = \int_{t_0}^{t} \mathrm{d}u\, e^{[A_0+A_R](t-u)}\, A_T[\alpha]\, \tilde{\theta}_0(u). \tag{70}$$

The pertubative resolution of eq. (27) yields the following equilibrium $T+A$ and $G+C$ contents:

$$\theta^*[p,\alpha,(\pm)] = \tilde{\theta}_0^* + \theta_T^*[\alpha] + O(\epsilon^2), \tag{71}$$

where

$$\tilde{\theta}_0^* = \theta_{[A_0+A_R]}, \tag{72}$$

$$\theta_T^*[\alpha] = \lambda_{[A_0+A_R]}\, A_T[\alpha]\tilde{\theta}_0^*. \tag{73}$$

As expected the $G+C$ content does not depend on replication fork polarity and gene orientation. Hence our minimal model (eqs. (9) and (10)) does not provide a satisfactory treatment of the $G+C$ content evolution. The $G+C$ 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 [61], should be considered to account for the $G+C$ content evolution. Our model only predicts a slight dependence of the $G+C$ content upon transcription rate through the $\theta_T[\alpha]$ coefficient. This change is however presumably small as compared to the variation of the $G+C$ content with recombination rate.

Long-term memory of the initial skews. If the skews are initially null, they increase according to eqs. (62)–(64) to ultimately reach their equilibrium values. Depending linearly on the substitutional asymmetry (eqs. (65)–(67)), 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

$$S[p,\alpha,(\pm)](t) = S_{\text{ini}}(t) + pS_R(t) \pm S_T[\alpha](t) + O(\epsilon^2), \tag{74}$$

where

$$S_{\text{ini}}(t) = e^{[D_0+D_R](t-t_0)} S(t_0). \tag{75}$$

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. (45)) 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 $\tau^{(2)}_{[D_0+D_R]}$.

4.3.3 Numerical tests

In fig. 8, we compare the exact and pertubative solutions for the toy model substitution rate matrix

$$M[p] = M_0^s + M_R^s + pM_R^a, \tag{76}$$

where $M_0^s + M_R^s$ and M_R^a are the substitution rate matrices estimated in the human genome as explained in appendix B (eqs. (B.1) and (B.3)). To express the substitution rates in per bp per Myrs units, we used 5 Myrs as an estimation of the human-chimp divergence. According to our minimal model eqs. (9) and (10), $M[p]$ is equal to the substitution rate matrix obtained in intergenic regions of replication fork polarity p. As shown in figs. 8A and 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 (figs. 8C and 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. (65), the skews at equilibrium are proportional to p (fig. 8D). Similarly, as predicted by eq. (71), the $T+A$ and $G+C$ contents at equilibrium do not depend upon p (fig. 8C). As governed by eq. (68), the time evolution of the $G+C$ content (fig. 8A) is not affected by the substitutional asymmetry pM_R^a, which explains that we recover the time evolution under the symmetrical matrix $M_0^s + M_R^s$ previously shown in fig. 7A. According to eq. (63), the skews converge towards their equilibrium values with time scales $\lambda_{A_0+A_R}$, $\lambda^{(1)}_{D_0+D_R}$ and $\lambda^{(2)}_{D_0+D_R}$. For the $M_0^s + M_R^s$ matrix given in eq. (B.1), these time scales are equal to 558 Myrs, 555 Myrs and 1360 Myrs. Hence the convergence of the skews towards their equilibrium values is a very long process.

5 From substitutional to compositional asymmetries

In the biological literature the skews are often normalized by the $G+C$ and $T+A$ contents [18,19]

$$S_{TA} = \frac{T-A}{T+A} \qquad \text{and} \qquad S_{GC} = \frac{G-C}{G+C}. \tag{77}$$

In the human genome, as the TA and GC skews correlate [26], 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.

5.1 The compositional skew decomposes into transcription- and replication-associated components

If the subtitutional asymmetry follows the decomposition observed in fig. 5 and formalized in eq. (16), the mathematical demonstration in sect. 4.3 shows that the same decomposition is expected for the compositional asymmetry, as measured by the TA and GC skews (eq. (77)). 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. 9A and B, the equilibrium skews S_{GC}^* and S_{TA}^* indeed decompose into transcription- and replication-associated components, consistent with the formal derivations made in sect. 4.3 (eq. (65)). If the current substitutional pattern is representative of the substitutional patterns that have

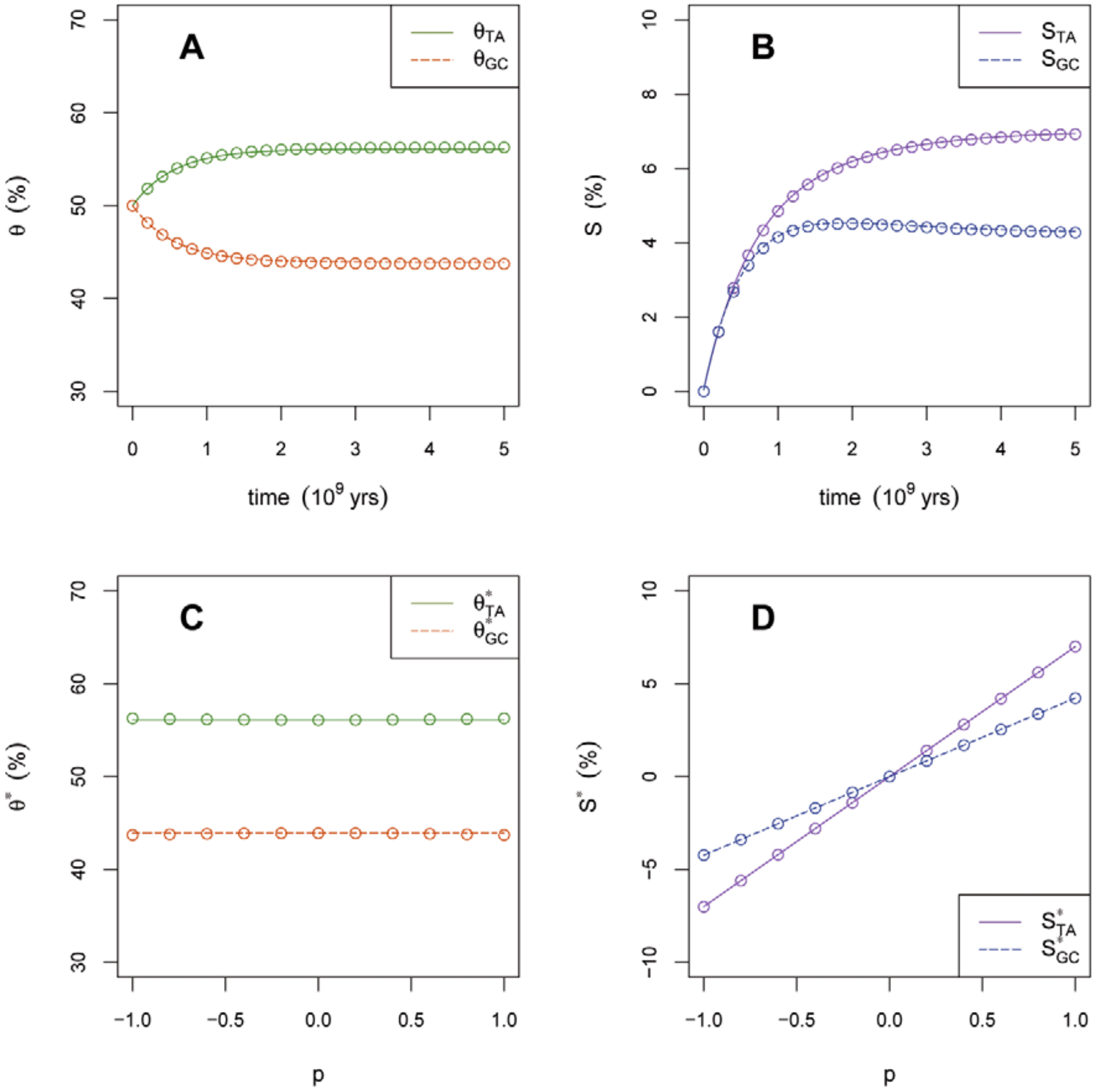

Fig. 8. (Colour on-line) DNA composition evolution in the presence of strand asymmetry: comparison of exact and pertubative solutions. The minimal model substitution rate matrix $M[p] = M_0^s + M_R^s + pM_R^a$ depends on the replication fork polarity p (eq. (76)). Exact solution is represented as circles, pertubative solution as solid line. Time evolution of the $T + A$ and $G + C$ 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 $p = 1$. Equilibrium $T + A$ and $G + C$ contents (C) and TA and GC skews (D) *versus* p.

shaped our genome, we expect the GC and TA compositional skews observed presently to follow the same decomposition. This is verified in figs. 9C and D, 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:

$$S = \begin{cases} pS_R + S_T & \text{genic } (+), \\ pS_R & \text{intergenic}, \\ pS_R - S_T & \text{genic } (-), \end{cases} \qquad (78)$$

in agreement with the minimal model for the compositional asymmetry proposed in sect. 4.3 (eqs. (62) and (65)). The coefficients S_T and S_R, estimated by least-squares fits to a line (dashed lines in fig. 9), 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 [18,19,26,27]. 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 $p > 0.45$),

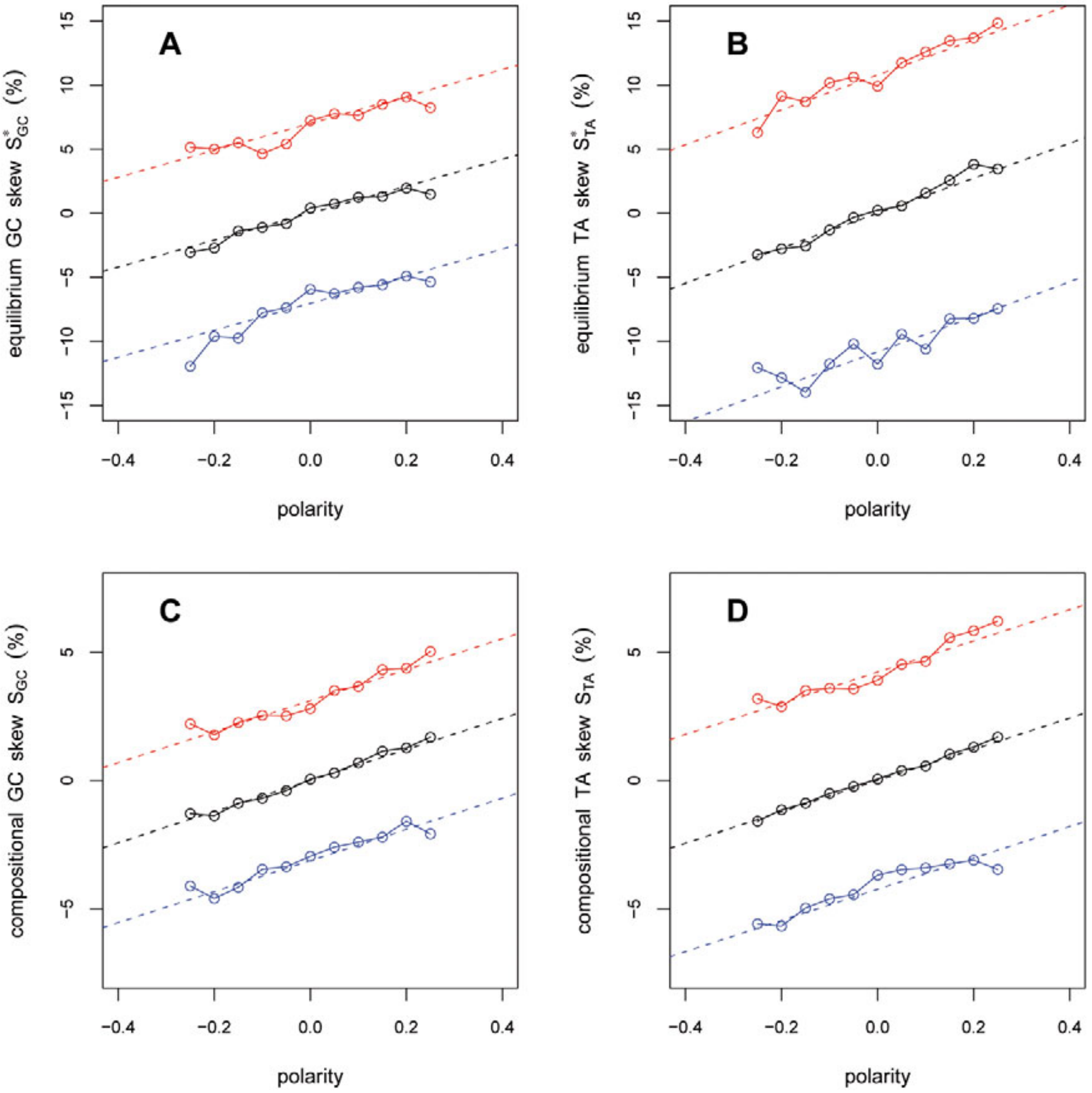

Fig. 9. (Colour on-line) Compositional asymmetry *versus* the replication fork polarity (determined in HeLa cell line) in genic sense (red), intergenic (black), and genic antisense (blue) regions, for (A) the equilibrium GC skew, (B) the equilibrium TA skew, (C) the compositional GC skew, and (D) the compositional TA skew. The equilibrium composition was directly computed from the substitution rate matrix (eq. (19) in sect. 4.1.1). 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. (78). The linear regression coefficients are reported in table 3.

Table 3. Transcription- and replication-associated compositional asymmetries. Coefficients S_R, S_T of the linear model eq. (78), obtained by least-squares fits to a line in fig. 9.

	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

which is strand-symmetric. As shown in fig. 10 along large (> 100 kbp) human genes, the GC and TA skews extend on the whole transcript [26,27].

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 1 Mbp 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 $< 10^{-16}$.

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

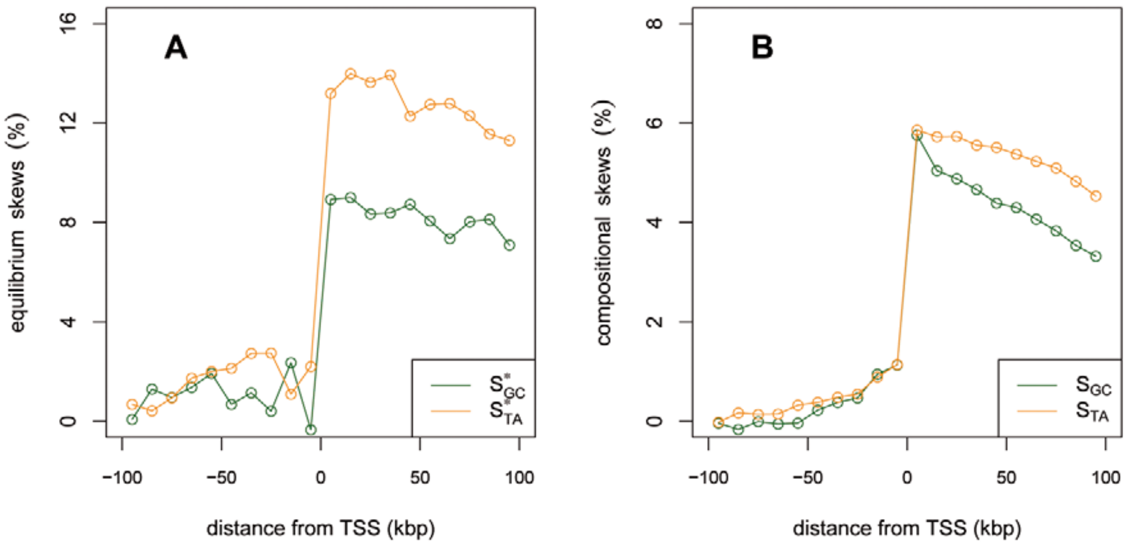

Fig. 10. (Colour on-line) Equilibrium and compositional skews along large (> 100 kbp) human genes. Equilibrium skews (A) and compositional skews (B) *versus* the distance to the TSS. Average substitution rates and nucleotide composition in large human genes were computed every 10 kbp from 100 kbp upstream (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.

5.2 The observed compositional skews were generated over several hundreds Myrs

If we compare the numerical values in figs. 9C and D, the observed compositional skews are twofold lower than the equilibrium skews shown in figs. 9A and B. The compositional skews have clearly not reached equilibrium yet. In this section we investigate the dynamics of compositional skew evolution. The convergence of the compositional skews towards equilibrium is governed by the time scales λ_A, $\lambda_D^{(1)}$, and $\lambda_D^{(2)}$ introduced in sect. 4.1 (eqs. (36) and (39)). 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. 11, if we start from initial null skew, and if the sequence is submitted to the substitution rates found in intergenic region at given replication fork polarity (fig. 5), the compositional skew increases over time. It is equal to the observed compositional skew (black circles in fig. 11) after 400 Myrs. It almost reaches equilibrium after three billion years (black crosses in fig. 11). Interestingly, the estimated time to reach the observed skew (400 Myrs) is much larger than the age of the mammalian radiation. However, we note that the 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

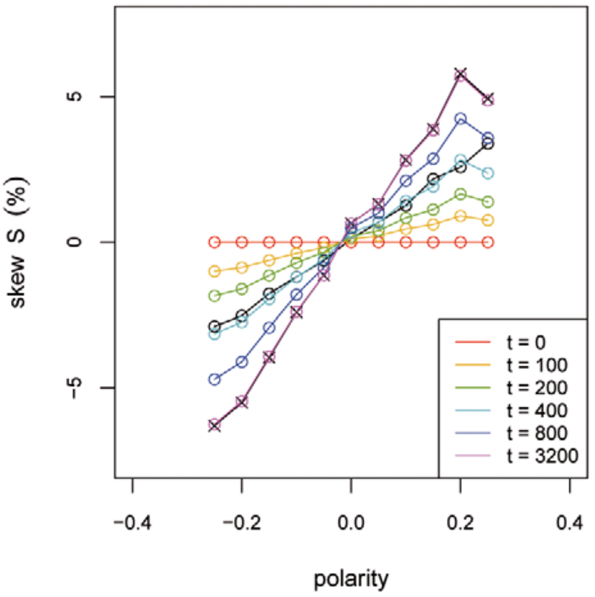

Fig. 11. (Colour on-line) Establishment of the compositional skew is a very slow process. Time evolution of the total skew $S = S_{TA} + S_{GC}$, from initially null skews, under the current substitutional pattern obtained in intergenic regions for different replication fork polarity values in HeLa (fig. 5). 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 sect. 4.1. 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).

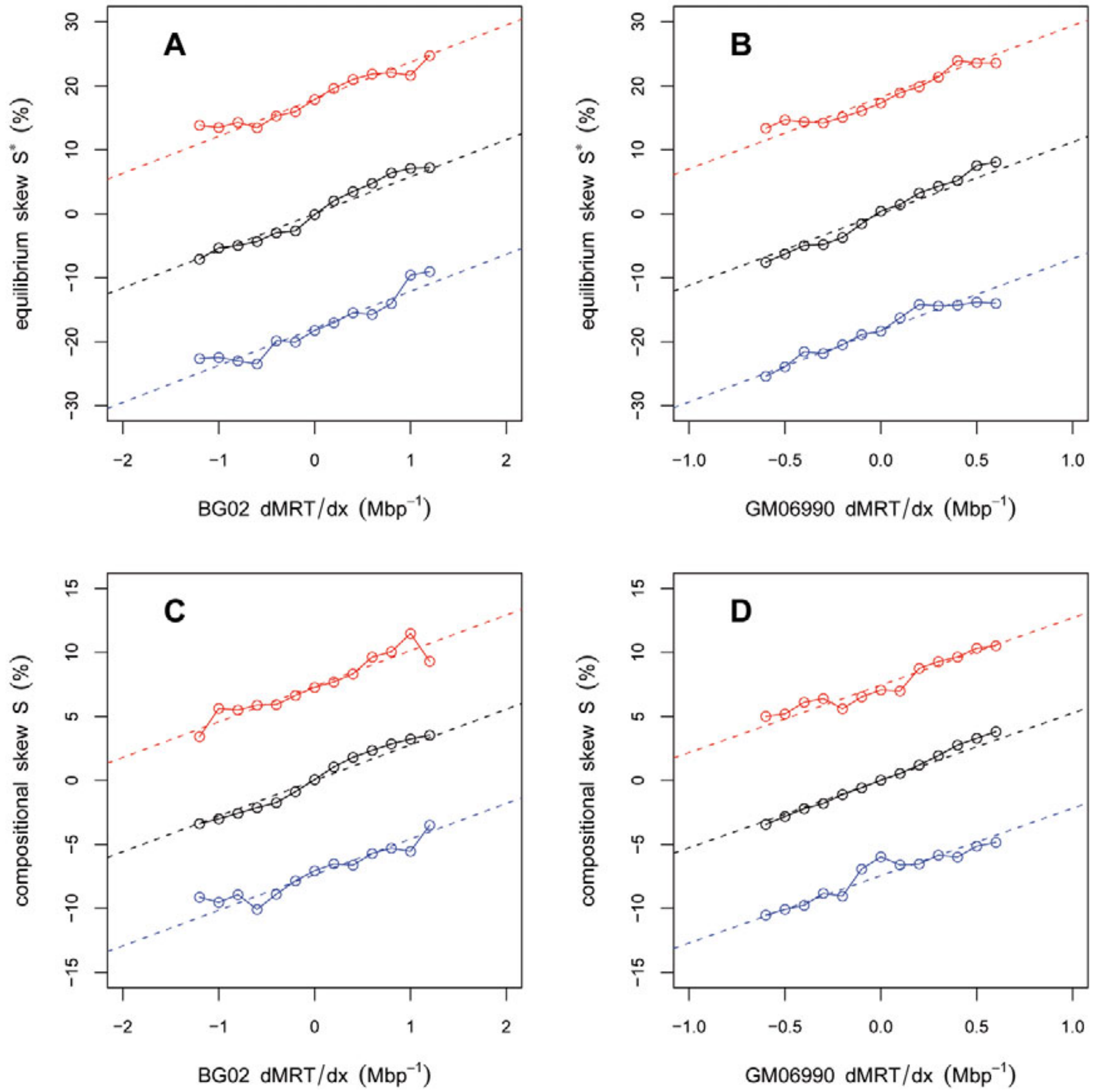

Fig. 12. (Colour on-line) The decomposition of S and S^* into transcription- and replication-associated components is observed for all examined cell types. (A) Equilibrium skew S^* *versus* $\mathrm{d\,MRT}/\mathrm{d}x$ in BG02 embryonic stem cell line for genic sense (red), intergenic (black), and genic antisense (blue) regions. (B) Same as in (A) but using $\mathrm{d\,MRT}/\mathrm{d}x$ determined in the GM06990 lymphoblastoid cell line. (C) Compositional skew S *versus* $\mathrm{d\,MRT}/\mathrm{d}x$ in BG02 embryonic stem cell line for genic sense (red), intergenic (black), and genic antisense (blue) regions. (D) Same as in (C) but using $\mathrm{d\,MRT}/\mathrm{d}x$ determined in the GM06990 lymphoblastoid cell line. Equilibrium and observed compositional skews, $\mathrm{d\,MRT}/\mathrm{d}x$, 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. (78). The replication timing data were retrieved from [25] (see appendix A.1).

Mugal *et al.* [40] 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 paper, 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 [73], 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 [76, 77].

6 Discussion

6.1 The good conservation of d MRT/dx across differentiation ensures the robustness of our analysis

In this study, 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 types (data from [25] see appendix A.1), in contrasts to HeLa (data from [29]), we had not access to both the replication fork velocity v and duration of the S phase T_S, and we were therefore not able to convert the $\mathrm{d\,MRT}/\mathrm{d}x$ 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 $\mathrm{d\,MRT}/\mathrm{d}x$, as exemplified in fig. 12 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 $\mathrm{d\,MRT}/\mathrm{d}x$ profile in any examined cell line (table 5). We infer from the good correlation obtained between the $\mathrm{d\,MRT}/\mathrm{d}x$ profiles of the different cell types that they all correlate with the $\mathrm{d\,MRT}/\mathrm{d}x$ 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 $\mathrm{d\,MRT}/\mathrm{d}x$ profiles is as high as between the $\mathrm{d\,MRT}/\mathrm{d}x$ profiles themselves (table 5).

Table 5. Conservation of $\mathrm{d\,MRT}/\mathrm{d}x$ across differentiation. Pearson correlation (R values) between $\mathrm{d\,MRT}/\mathrm{d}x$ profiles from various cell types: 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 $\mathrm{d\,MRT}/\mathrm{d}x$. S, S^*, and $\mathrm{d\,MRT}/\mathrm{d}x$ were calculated genome wide in non-overlapping 1 Mbp windows using replication timing data from [25,29]. 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 $< 10^{-16}$.

	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

6.2 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 [78], the replication-associated asymmetries observed at compositional skew upward jumps (S-jumps) in the human genome [18,19] bordering replication N-domains (see fig. 2B) are about threefold 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 $S = 3.72\%$ [13], 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 polarity, 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 the 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 were equal to twice its value measured by DNA combing in HeLa cells [29], 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 sects. 3 and 4, 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}

$$p_{\mathrm{germline}} = K p_{\mathrm{HeLa}}. \tag{79}$$

According to our minimal model (sect. 3), we expect to observe the following substitutional asymmetries:

$$\tau^a = \begin{cases} p_{\text{germline}}\tau_R^a + \tau_T^a = p_{\text{HeLa}}(K\tau_R^a) + \tau_T^a & \text{genic } (+), \\ p_{\text{germline}}\tau_R^a = p_{\text{HeLa}}(K\tau_R^a) & \text{intergenic}, \\ p_{\text{germline}}\tau_R^a - \tau_T^a = p_{\text{HeLa}}(K\tau_R^a) - \tau_T^a & \text{genic } (-). \end{cases} \tag{80}$$

Hence the coefficient $\tau^a_{R,\text{HeLa}} = K\tau^a_R$, as estimated by the linear regression *versus* p_{HeLa}, is expected to be proportional to τ^a_R. If $K > 1$ (respectively, $K < 1$), the coefficients reported in table 1 would actually overestimate (respectively, underestimate) the replication-associated asymmetries.

6.3 Effect of gene expression on substitution rates

In the minimal model proposed in sect. 3, 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 [79], the $(A \to G)^a$ 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 [80], expression in testis [81], and breath of expression [82], 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 $(A \to G)^a$ asymmetry and the $G + T$ content most strongly correlate with the expression in spermatogonia [79]. On the opposite, the $(C \to T)^a$ asymmetry does not correlate significantly with the expression in germline cells [79]. Our results suggest that the $(C \to T)^a$ asymmetry, in transcribed and non-transcribed regions, is mainly driven by the replication fork polarity (fig. 5B), which could explain the poor correlation observed in [79]. The correlation could also be affected by the variation of the $(C \to T)^a$ asymmetry along transcripts, as observed in [39]. Note that the poor correlation between the $(C \to T)^a$ 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.* [73] reported a strong correlation between the $(C \to T)^a$ 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 [74]) [40,73]. According to [73], a linear model based on these two predictors has the best explanatory power which strongly supports the minimal model proposed in sect. 3 for substitutional asymmetry.

We note the lower symmetrical substitution rates in genes (see fig. 6) among which the strong to weak $C \to T$ and $G \to T$ substitutions were the most affected (figs. 6B and D). 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 $G + C$ rich alleles, can neither be invoked, as it impacts on weak to strong substitution rates, but not on strong to weak substitution rates [61]. However BGC, along with reduced recombination rates observed in genes [79], could explain the weakly reduced weak to strong $(A \to G)^s$ symmetrical substitution rate (fig. 6A). 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.

This work was supported by the Agence Nationale de la Recherche (projet "REFOPOL": Spatio-temporal program of replication of the human genome, ANR 10 BLAN 1615).

Appendix A. Material and methods

Appendix A.1. Determining the replication fork polarity from replication timing profiles

The mean replication timing profile was determined in [17] using replication timing data available in seven human cell types [25,28]. In the Hela cell line, we used the values of the replication fork velocity v and duration of the S phase T_s from [29], and we converted the $\mathrm{d\,MRT}/\mathrm{d}x$ profile to a replication fork polarity profile according to eq. (1).

Appendix A.2. Sequence and annotation data

Sequence and annotation data were retrieved from the Genome Browsers of the University of California Santa Cruz (UCSC) [83]. 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".

$$M_0^s + M_R^s = \begin{array}{|c|c|c|c|c|}\hline \swarrow & T & A & G & C \\ \hline T & & 0.638 \pm 0.005 & 1.248 \pm 0.012 & 3.783 \pm 0.017 \\ \hline A & 0.638 \pm 0.005 & & 3.783 \pm 0.017 & 1.248 \pm 0.012 \\ \hline G & 0.816 \pm 0.007 & 3.123 \pm 0.023 & & 1.224 \pm 0.009 \\ \hline C & 3.123 \pm 0.023 & 0.816 \pm 0.007 & 1.224 \pm 0.009 & \\ \hline \end{array} \tag{B.1}$$

$$M_T^s = \begin{array}{|c|c|c|c|c|}\hline \swarrow & T & A & G & C \\ \hline T & & -0.085 \pm 0.006 & -0.245 \pm 0.015 & -0.419 \pm 0.021 \\ \hline A & -0.085 \pm 0.006 & & -0.419 \pm 0.021 & -0.245 \pm 0.015 \\ \hline G & -0.045 \pm 0.009 & -0.160 \pm 0.028 & & -0.070 \pm 0.011 \\ \hline C & -0.160 \pm 0.028 & -0.045 \pm 0.009 & -0.070 \pm 0.011 & \\ \hline \end{array} \tag{B.2}$$

$$M_R^a = \begin{array}{|c|c|c|c|c|}\hline \swarrow & T & A & G & C \\ \hline T & & 0.050 \pm 0.008 & 0.035 \pm 0.013 & 0.352 \pm 0.023 \\ \hline A & -0.050 \pm 0.008 & & -0.352 \pm 0.023 & -0.035 \pm 0.013 \\ \hline G & -0.012 \pm 0.009 & 0.428 \pm 0.026 & & 0.120 \pm 0.012 \\ \hline C & -0.428 \pm 0.026 & 0.012 \pm 0.009 & -0.120 \pm 0.012 & \\ \hline \end{array} \tag{B.3}$$

$$M_T^a = \begin{array}{|c|c|c|c|c|}\hline \swarrow & T & A & G & C \\ \hline T & & 0.043 \pm 0.002 & 0.075 \pm 0.002 & -0.048 \pm 0.005 \\ \hline A & -0.043 \pm 0.002 & & 0.048 \pm 0.005 & -0.075 \pm 0.002 \\ \hline G & -0.011 \pm 0.002 & 0.541 \pm 0.005 & & 0.120 \pm 0.002 \\ \hline C & -0.541 \pm 0.005 & 0.011 \pm 0.002 & -0.120 \pm 0.002 & \\ \hline \end{array} \tag{B.4}$$

Appendix A.3. Determination of substitution rates in the human genome

Substitutions were tabulated in the human lineage since its divergence with chimpanzee using macaca and orangutan as outgroups [28]. 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 [84] for review. As a result the CpG dinucleotide is hypermutable, and the CpG → TpG and its reverse complementary CpG → CpA are by far the principal neighbor-dependent substitutions rates [85, 86]. The twelve neighbor-independent substitution rates were determined on non-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 [27].

Appendix B. Estimating the coefficients of our minimal model

We computed substitution rates separately in genic (+), intergenic, and genic (−) genomic regions of given replication fork polarity values (determined in the HeLa cell line (appendix A.1)). Genomic regions were classed as genic (+), genic (−), and intergenic using RefGene transcripts (appendix A.2). The substitution rates correspond to averaged values, usually on several Mbp of aligned sequences.

As in fig. 5, the asymmetrical coefficients τ_R^a and τ_T^a of our minimal model (eq. (10)) were estimated by least-squares fits to a line following the linear model eq. (16) for the substitutional asymmetry. The symmetrical coefficients $\tau_0^s + \tau_R^s$ and τ_T^s of our minimal model (eq. (9)) were estimated by least-squares fits to a line following the linear model

$$\tau^s = \begin{cases} \tau_0^s + \tau_R^s + \tau_T^s & \text{genic } (+), \\ \tau_0^s + \tau_R^s & \text{intergenic}, \\ \tau_0^s + \tau_R^s + \tau_T^s & \text{genic } (-). \end{cases} \tag{B.5}$$

The coefficients obtained for the twelve neighbor independent substitution rates are tabulated in form of a substitution rate matrix (see sect. 2) and are reported as

see eqs. (B.1)–(B.4) *above*

We clearly see from these estimates that the asymmetrical parts M_R^a and M_T^a are small compared to the symmetrical part $M_R^s + M_0^s$, condition required for the perturbative analysis made in sect. 4.

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.

References

1. R. Berezney, D.D. Dubey, J.A. Huberman, Chromosoma **108**, 471 (2000).
2. D.M. Gilbert, Science **294**, 96 (2001).
3. M. Méchali, Nat. Rev. Genet. **2**, 640 (2001).
4. S.P. Bell, A. Dutta, Annu. Rev. Biochem. **71**, 333 (2002).
5. A.J. McNairn, D.M. Gilbert, Bioessays **25**, 647 (2003).
6. M.I. Aladjem, Nat. Rev. Genet. **8**, 588 (2007).
7. S. Courbet, S. Gay, N. Arnoult, G. Wronka, M. Anglana, O. Brison, M. Debatisse, Nature **455**, 557 (2008).
8. J.L. Hamlin, L.D. Mesner, O. Lar, R. Torres, S.V. Chodaparambil, L. Wang, J. Cell. Biochem. **105**, 321 (2008).
9. M. Méchali, Nat. Rev. Mol. Cell Biol. **11**, 728 (2010).
10. D.M. Gilbert, Nat. Rev. Genet. **11**, 673 (2010).
11. J.R. Lobry, Mol. Biol. Evol. **13**, 660 (1996).
12. J. Mrázek, S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **95**, 3720 (1998).
13. C.L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton Carafa, O. Hyrien, A. Arneodo *et al.*, Mol. Biol. Evol. **28**, 2327 (2011).
14. B. Alberts, A. Jonhson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, 5th edition (Garland Publishing, New York, 2008).
15. M.L. DePamphilis (Editor), *DNA Replication and Human Disease* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2006).
16. B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **99**, 248102 (2007).
17. A. Baker, B. Audit, C. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard *et al.*, PLoS Comput. Biol. **8**, e1002443 (2012).
18. M. Touchon, S. Nicolay, B. Audit, E.B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005).
19. E.B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **94**, 248103 (2005).
20. S. Nicolay, PhD Thesis, University of Liège, Belgium (2006).
21. M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Genome Res. **17**, 1278 (2007).
22. B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton Carafa, C. Thermes, A. Arneodo, Nucl. Acids Res. **37**, 6064 (2009).
23. C. Lemaitre, L. Zaghloul, M.F. Sagot, C. Gautier, A. Arneodo, E. Tannier, B. Audit, BMC Genomics **10**, 335 (2009).
24. L. Zaghloul, PhD Thesis, ENS de Lyon, France (2009).
25. R.S. Hansen, S. Thomas, R. Sandstrom, T.K. Canfield, R.E. Thurman, M. Weaver, M.O. Dorschner, S.M. Gartler, J.A. Stamatoyannopoulos, Proc. Natl. Acad. Sci. U.S.A. **107**, 139 (2010).
26. M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, FEBS Lett. **555**, 579 (2003).
27. M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Nucl. Acids Res. **32**, 4969 (2004).
28. C.L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, O. Hyrien *et al.*, Genome Res. **20**, 447 (2010).
29. G. Guilbaud, A. Rappailles, A. Baker, C.L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, O. Hyrien, PLoS Comput. Biol. **7**, e1002322 (2011).
30. N. Sueoka, J. Mol. Evol. **40**, 318 (1995).
31. R. Rudner, J.D. Karkas, E. Chargaff, Proc. Natl. Acad. Sci. U.S.A. **60**, 921 (1968).
32. J.R. Lobry, J. Mol. Evol. **40**, 326 (1995).
33. J.R. Lobry, C. Lobry, Mol. Biol. Evol. **16**, 719 (1999).
34. A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa, C. Thermes, Phys. Rep. **498**, 45 (2011).
35. M.P. Francino, H. Ochman, Trends Genet **13**, 240 (1997).
36. A.C. Frank, J.R. Lobry, Gene **238**, 65 (1999).
37. M.P. Francino, H. Ochman, Mol. Biol. Evol. **18**, 1147 (2001).
38. P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Nat. Genet. **33**, 514 (2003).
39. P. Polak, P.F. Arndt, Genome Res. **18**, 1216 (2008).
40. C.F. Mugal, H.H. von Grunberg, M. Peifer, Mol. Biol. Evol. **26**, 131 (2009).
41. A. Beletskii, A.S. Bhagwat, Proc. Natl. Acad. Sci. U.S.A. **93**, 13919 (1996).
42. A. Beletskii, A.S. Bhagwat, Biol. Chem. **379**, 549 (1998).
43. M.P. Francino, L. Chao, M.A. Riley, H. Ochman, Science **272**, 107 (1996).
44. J.Q. Svejstrup, Nat. Rev. Mol. Cell Biol. **3**, 21 (2002).
45. S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. Lett. **93**, 108101 (2004).
46. S. Nicolay, E.B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Phys. Rev. E **75**, 032902 (2007).
47. E.P. Rocha, A. Danchin, A. Viari, Mol. Microbiol. **32**, 11 (1999).
48. E.P.C. Rocha, M. Touchon, E.J. Feil, Genome Res. **16**, 1537 (2006).
49. P. Polak, P.F. Arndt, Genome Biol. Evol. **1**, 189 (2009).
50. T.A. Kunkel, P.M. Burgers, Trends Cell Biol. **18**, 521 (2008).
51. J.R. Lobry, Science **272**, 745 (1996).
52. E.R. Tillier, R.A. Collins, J. Mol. Evol. **50**, 249 (2000).
53. A. Grigoriev, Virus Res. **60**, 1 (1999).
54. A. Gierlik, M. Kowalczuk, P. Mackiewicz, M.R. Dudek, S. Cebrat, J. Theor. Biol. **202**, 305 (2000).
55. A. Reyes, C. Gissi, G. Pesole, C. Saccone, Mol. Biol. Evol. **15**, 957 (1998).
56. M.C. Marsolier-Kergoat, A. Goldar, Mol. Biol. Evol. **29**, 893 (2012).
57. F. Jacob, S. Brenner, F. Cuzin, Cold Spring Harb. Symp. Quant. Biol. **28**, 329 (1963).
58. D. Graur, W.H. Li, *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1999).
59. M. Kimura, Nature **217**, 624 (1968).
60. N. Galtier, G. Piganeau, D. Mouchiroud, L. Duret, Genetics **159**, 907 (2001).
61. L. Duret, P.F. Arndt, PLoS Genet. **4**, e1000071 (2008).
62. L. Duret, N. Galtier, Annu. Rev. Genomics Hum. Genet. **10**, 285 (2009).
63. The ENCODE Project Consortium, Nature **447**, 799 (2007).
64. J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt *et al.*, Science **308**, 1149 (2005).

65. L.J. Core, J.J. Waterfall, J.T. Lis, Science **322**, 1845 (2008).
66. Y. He, B. Vogelstein, V.E. Velculescu, N. Papadopoulos, K.W. Kinzler, Science **322**, 1855 (2008).
67. P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M.S. Christensen, C.K. Mapendano, M.H. Schierup, T.H. Jensen, Science **322**, 1851 (2008).
68. A.C. Seila, J.M. Calabrese, S.S. Levine, G.W. Yeo, P.B. Rahl, R.A. Flynn, R.A. Young, P.A. Sharp, Science **322**, 1849 (2008).
69. A. Necsulea, C. Guillet, J.C. Cadoret, M.N. Prioleau, L. Duret, Mol. Biol. Evol. **26**, 729 (2009).
70. A. Baker, PhD Thesis, University of Lyon, France (2011).
71. L.A. Frederico, T.A. Kunkel, B.R. Shaw, Biochemistry **29**, 2532 (1990).
72. J.A. Stamatoyannopoulos, I. Adzhubei, R.E. Thurman, G.V. Kryukov, S.M. Mirkin, S.R. Sunyaev, Nat. Genet. **41**, 393 (2009).
73. C.F. Mugal, J.B.W. Wolf, H.H. von Grunberg, H. Ellegren, Genome Biol. Evol. **2**, 19 (2010).
74. C.L. Chen, A. Baker, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, in preparation (2012).
75. N. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd edition (North-Holland, Amsterdam, 2007).
76. T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T.C. Schulz, A.J. Robins, S. Dalton, D.M. Gilbert, Genome Res. **20**, 761 (2010).
77. E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, I. Simon, PLoS Genet. **6**, e1001011 (2010).
78. A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton-Carafa, C. Thermes, B. Audit, A. Arneodo, Appl. Comput. Harmon. Annal. **28**, 150 (2010).
79. G. McVicker, P. Green, Genome Res. **20**, 1503 (2010).
80. J. Majewski, Am. J. Hum. Genet. **73**, 688 (2003).
81. J.M. Comeron, Genetics **167**, 1293 (2004).
82. L. Duret, Curr. Opin. Genet. Dev. **12**, 640 (2002).
83. D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas *et al.*, Nucl. Acids Res. **31**, 51 (2003).
84. M.M. Suzuki, A. Bird, Nat. Rev. Genet. **9**, 465 (2008).
85. S.T. Hess, J.D. Blake, R.D. Blake, J. Mol. Biol. **236**, 1022 (1994).
86. P.F. Arndt, T. Hwa, Bioinformatics **21**, 2322 (2005).