Replication-Associated Mutational Asymmetry in the Human Genome

Chun-Long Chen,1 Lauranne Duquenne,†1 Benjamin Audit,2,3 Guillaume Guilbaud,4
Aurélien Rappailles,4 Antoine Baker,2,3 Maxime Huvet,†1 Yves d’Aubenton-Carafa,1 Olivier Hyrien,4
Alain Arneodo,²,³ and Claude Thermes*1

1Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France
2Université de Lyon, Lyon, France
3Laboratoire Joliot Curie et Laboratoire de Physique, Ecole Normale Supérieure de Lyon, CNRS, Lyon, France
4Institut de Biologie de l’Ecole Normale Supérieure (IBENS) - CNRS UMR 8197 - Inserm U 1024, Paris, France
†Present address: CNRS UMR 5558, Université Claude Bernard Lyon1, Villeurbanne Cedex, France
‡Present address: Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London, United Kingdom
*Corresponding author: E-mail: thermes@cgm.cnrs-gif.fr.
Associate editor: Koichiro Tamura

Abstract
During evolution, mutations occur at rates that can differ between the two DNA strands. In the human genome, nucleotide substitutions occur at different rates on the transcribed and non-transcribed strands that may result from transcription-coupled repair. These mutational asymmetries generate transcription-associated compositional skews. To date, the existence of such asymmetries associated with replication has not yet been established. Here, we compute the nucleotide substitution matrices around replication initiation zones identified as sharp peaks in replication timing profiles and associated with abrupt jumps in the compositional skew profile. We show that the substitution matrices computed in these regions fully explain the jumps in the compositional skew profile when crossing initiation zones. In intergenic regions, we observe mutational asymmetries measured as differences between complementary substitution rates; their sign changes when crossing initiation zones. These mutational asymmetries are unlikely to result from cryptic transcription but can be explained by a model based on replication errors and strand-biased repair. In transcribed regions, mutational asymmetries associated with replication superimpose on the previously described mutational asymmetries associated with transcription. We separate the substitution asymmetries associated with both mechanisms, which allows us to determine for the first time in eukaryotes, the mutational asymmetries associated with replication and to reevaluate those associated with transcription. Replication-associated mutational asymmetry may result from unequal rates of complementary base misincorporation by the DNA polymerases coupled with DNA mismatch repair (MMR) acting with different efficiencies on the leading and lagging strands. Replication, acting in germ line cells during long evolutionary times, contributed equally with transcription to produce the present abrupt jumps in the compositional skew. These results demonstrate that DNA replication is one of the major processes that shape human genome composition.

Key words: human genome, mutational asymmetries, compositional skew, nucleotide substitutions

Introduction
In prokaryote genomes, single nucleotide substitutions have been shown to occur at different rates on the two DNA strands due to asymmetries intrinsic to the replication and transcription processes (Francino and Ochman 1997; Frank and Lobry 1999). In a number of bacterial genomes, unequal intrastrand frequencies of complementary nucleotides have been associated with replication, the leading replicating strand frequently presenting an excess of G over C and of T over A (Mrazek and Karlin 1998; Rocha et al. 2006). These compositional asymmetries (GC and TA skews defined as $S_{GC} = (G - C)/(G + C)$, $S_{TA} = (T - A)/(T + A)$ and total skew $S = S_{GC} + S_{TA}$) result from neutral nucleotide substitution rates that differ in the leading and lagging replicating strands. This implies that, on one strand, complementary substitution rates differ from each other and that these compositional asymmetries switch sign when crossing the position of the replication origin. For example, if on the leading strand the transition rate $r(A \rightarrow G)$ is larger than $r(T \rightarrow C)$, this will generate, after long evolutionary times, positive values of the skew $S$ on the 3’ side of the origin and negative values on the 5’ side, producing a sharp upward jump of the $S$ profile at the origin position. This property has been used to detect replication origins in bacteria, organelles, and viruses (Mrazek and Karlin 1998; Reyes et al. 1998; Grigoriev 1999; Krishnan et al. 2004; Mackiewicz et al. 2004). In human, few studies have examined substitutions around human replication origins due to a poor knowledge of their positions. Previous studies of the ß-globin locus concluded to an absence of
substitution asymmetries associated with replication (Bulmer 1991; Francino and Ochman 2000). A recent study observed intergenic mutational asymmetries around CpG islands and proposed that they result either from replication originating at CpG islands or from transcription of unknown RNA molecules (Polak and Arndt 2009). Analyses of human genome nucleotide composition exhibited a number (1,546) of upward jumps of the skew $S$ profile (see fig. 1A) and allowed to propose that these upward jumps coincide with replication origins and that their shape mainly results from mutational asymmetries associated with replication (Brodie et al. 2005; Touchon et al. 2005; Audit et al. 2007; Huvet et al. 2007). This was debated in a later study, which suggested that these upward jumps result from mutational asymmetries associated with transcription not with replication (Neculea et al. 2009). Here, we use the replication timing profiles of various human cell types to detect replication initiation zones and show that the upward jumps of the $S$ profile are associated with replication initiation zones with a high likelihood of being active in germ line cells. We then determine and analyze the single nucleotide substitution rates along the human genome and demonstrate that the upward jumps of the $S$ profile result from newly evidenced replication-associated asymmetries superimposed to already known transcription-associated asymmetries. These results establish for the first time DNA replication as a major process driving the evolution of the human genome composition.

Materials and Methods

Determination of the Replication Timing Profiles

The replication timing data of Hela cells along the human genome were processed as described in (Chen et al. 2010). The enrichment of sequence reads along the genome was computed in each $S$ phase compartment and the ratio $S_{50}$, defined as the fraction of the $S$ phase (0 computed in each $S$ phase compartment and the ratio $S_{50}$, was computed by linear interpolation of enrichment values in the four $S$ phase compartments. $S_{50}$ values for human embryonic stem cells (ESCs) BGO2, fibroblasts BJ, lymphoblastoid GM06990, and erythroid K562 cells were computed by using the enrichment values in the six compartments of the $S$ phase for each cell line (raw data obtained from Hansen et al. 2010). The replication timing data (log($S/G1$)) of H1 ESC were retrieved from Desprat et al. 2009.

Multiscale Detection of the Peaks in Replication Timing Profiles

The simple intuitive idea allowing for effective detection of peaks in a noisy profile is to delineate positions along the signal that are local extrema (null derivative) and present a strong (negative) curvature (negative second derivative) as expected at the tip of a vertical peak. Here, we used the continuous wavelet transform, which provides a powerful framework for the robust estimation of signal variations (first and second derivative) over any length scales. In these space-scale representations of signal variations, we delineate 1) regions that are local extrema (supplementary fig. 1A and B, Supplementary Material online) and 2) regions presenting a strong concavity (supplementary fig. 1C, Supplementary Material online). They are combined to delineate candidate space-scale regions corresponding to peaks in the timing profile (supplementary fig. 1A, Supplementary Material online). Actually, in order to avoid the confusion between “true” peaks and those induced by the presence of a noisy background, we retain only peaks that correspond to connected candidate space-scale regions with a scale extension greater than 1.74 (see supplementary Methods, Supplementary Material online).

Comparison of S-Jumps With t-Peaks

The $S$-jumps detected in the skew $S$ profile were retrieved from (Touchon et al. 2005). The distance $d$ between each $S$-jump and the closest $t$-peak was computed for each timing profile; we selected the $S$-jumps close to $t$-peaks ($d < 100$ kbp) in $n$, and only $n$, different timing profiles as indicated by the red arrows in figure 1A. Random simulations were performed to evaluate the significance of the results: at each simulation step, the number of $S$-jumps and the distribution of intervals between $S$-jumps were conserved for each chromosome. The distance $d$ between each random $S$-jump position and the closest $t$-peak was computed; the distributions of the numbers of these random $S$-jumps close to $t$-peaks in $n$ timing profiles resulting from 1,000 simulations were identified as normal distributions (Sapiro–Wilk normality test, $P > 0.01$) or Poisson distributions (Goodness-of-fit test, $P > 0.01$) and used to obtain the (expected) mean values and the corresponding one-tailed probability.

Determination of the Substitution Rates

Substitution rates were computed using the four-way catarrhini-specific alignments of *Homo sapiens* (assembly hg18), *Pan troglodytes* (panTro2), *Pongo pygmaeus abelii* (ponAbe2), and *Macaca mulatta* (rheMac2) as described in Chen et al. (2010). These alignments were generated using the EPO (Enredo Pecan Ortheus) pipeline (Paten, Herrero, Beal, et al. 2008; Paten, Herrero, Fitzgerald, et al. 2008) and were retrieved from the Ensembl Genome Brower (http://www.ensembl.org). Annotations of the human genome were retrieved from the University of California–San Cruz Genome Brower (http://genome.ucsc.edu). Substitutions were tabulated on the Watson strand, that is, the upper (S’–3’) strand of double-stranded sequences. For each $S$-jump, the sequence available for the analysis starts at the $S$-jump position and ends at the middle of the segment separating this $S$-jump from the adjacent $S$-jump.

Nucleotide Composition at Equilibrium

We computed the compositional skew profile toward which the sequences are evolving as the result of the observed substitution profiles (four transitions and eight transversions for non CpG substitutions, two transitions...
for CpG substitutions). The method used to compute the nucleotide composition at equilibrium is based on the model of sequence evolution with neighbor-dependent mutations introduced by Arndt et al. (2003). Because complementary substitutions do not occur at the same rate around upward jumps, we took into account the 12 neighbor-independent substitution processes (four transitions and eight transversions). In addition to these 12 rates, we consider the most abundant neighbor-dependent substitution process, namely the deamination of methylated
cytosines in CpG dinucleotides, resulting in TpG (or CpA on the complementary strand). Thanks to these 14 parameters, we are able to express the evolution through time of the 16 dinucleotide frequencies as a system of 16 differential equations (supplementary equation 1, Supplementary Material online). To solve these equations, we used a standard numerical approach, the conjugate gradient method using a Newton-type algorithm (nlm function in R package, http://www.r-project.org/). We computed the dinucleotide frequencies and deduced the nucleotide composition, which allowed us to determine the compositional skew at equilibrium S*.

Properties of Replication Initiation Zones Distributions Among Different Cell Types

We examined the timing profiles of six cell lines including two different ESC lines (BG02 and H1 ESC), a fibroblast, a lymphoid, an erythroid, and an Hela cell lines (Desprat et al. 2009; Chen et al. 2010; Hansen et al. 2010) (Materials and Methods). They all present numerous peaks pointing toward early replication times (fig. 1B–G). The regions located at the peak centers are replicated earlier than the surrounding regions, and thus, harbor replication initiation zones highly active in the corresponding cell type. These peaks (t-peaks) were detected in each replication timing profile using a multiscale methodology (see Materials and Methods). Their mean profiles present symmetrical shapes, and the timing values at the peak centers range from early to late timing values (supplementary fig. 2, Supplementary Material online).

Recent results have shown that the replication timing pattern changes among different cell types in more than half of genome regions (Hansen et al. 2010; Hiratani et al. 2010). We examined to what extent the t-peaks coincide with each other among the six cell types (fig. 1B–G). For each cell type, the t-peaks are significantly close to t-peaks in other timing profiles (at distance d < 100 kbp, P < 10^-16, supplementary fig. 3, Supplementary Material online). We identified the t-peaks in each timing profile that coincide (i.e., are located at distances d < 100 kbp) with n t-peaks in the 5 other timing profiles (table 1A, supplementary table 1, Supplementary Material online).

Among the 1,690 t-peaks of the BG02 timing profile, 962 (57%) coincide (d < 100 kbp) with at least one t-peak (n ≥ 1) in the 5 other timing profiles (table 1A). For example, 128 BG02 t-peaks each coincide with t-peaks in 3 (n = 3) other timing profiles (e.g., in fig. 1B, the t-peak at position 14.2 Mbp coincides with one t-peak in the fibroblast, lymphoid, and Hela profiles). For each n value, the

<table>
<thead>
<tr>
<th>Table 1. Replication Initiation Zones (t-peaks) Common to Different Cell Types.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>(A) t-peaks (BG02)</td>
</tr>
<tr>
<td>Expected</td>
</tr>
<tr>
<td>r</td>
</tr>
<tr>
<td>P value</td>
</tr>
<tr>
<td>(B) t-peaks (BG02)</td>
</tr>
<tr>
<td>Common(Hela) %</td>
</tr>
<tr>
<td>Common(Hela) %, ratio</td>
</tr>
<tr>
<td>Expected</td>
</tr>
<tr>
<td>Expected %</td>
</tr>
<tr>
<td>P value</td>
</tr>
</tbody>
</table>

(a) Normal distributions.
(b) Poisson distributions.

Note.—(A) Peaks of the replication timing profile (t-peaks) of ESC BG02 significantly coincide with t-peaks of five other cell types (fibroblast, lymphoid, erythroid, Hela, and H1 ESC). For each n value are indicated the numbers of BG02 t-peaks close (d < 100 kbp) to the position of t-peaks in n and only n of the 5 other timing profiles; expected, numbers (mean ± SD) of randomized BG02 t-peaks close to n t-peaks in the 5 other timing profiles (1,000 simulations, Materials and Methods); r, ratio between the observed number of BG02 t-peaks close to a number n of t-peaks and the corresponding expected number; P values, probabilities resulting from the simulations.

(B) The BG02 t-peaks common to n among 4 other cell types are also common to an additional cell type (Hela). Line 1, number of BG02 t-peaks close to n t-peaks among 4 other timing profiles (fibroblast, lymphoid, erythroid, H1 ESC; common(Hela) (line 2). For each n value, number of these t-peaks (line 1) that coincide with Hela t-peaks; common(Hela) %, ratio (line 2)/(line 1) in percent; expected, numbers (mean ± SD) of BG02 t-peaks common to the 4 other timing profile that coincide with randomized Hela t-peaks (1000 simulations, Materials and Methods); P values, probabilities resulting from the simulations.
number of these t-peaks is significantly larger than expected by chance (30 ± 5 for n = 3, P < 10⁻⁶, table 1A). The ratio r between the observed and expected numbers of BG02 t-peaks common to n timing profiles strongly increases (from 1.8 for n = 1 to 9.3 for n = 5), as also observed with the t-peaks in the other cell types (fig. 1H, supplementary table 1, Supplementary Material online). This suggests that a t-peak common to many timing profiles is more likely to be also common to an additional timing profile. Indeed, the proportion of the BG02 t-peaks common to n among 4 other timing profiles (H1 ESC, fibroblast, lymphoid, and erythroid) that are also common to an additional timing profile (Hela) strongly increases with n (table 1B). This proportion increases from 15% for n = 0 to 88% for n = 4. This confirms that initiation zones (t-peaks) common to a group of n cell types have a probability to be also active in other cell types, which increases with n.

### Table 2. The Properties of the S-Jumps Distributions Are Similar to Those of t-Peaks.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>n = 0</th>
<th>n = 1</th>
<th>n = 2</th>
<th>n = 3</th>
<th>n = 4</th>
<th>n = 5</th>
<th>n = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S-jumps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected</td>
<td>1,546</td>
<td>683</td>
<td>378</td>
<td>224</td>
<td>136</td>
<td>76</td>
<td>36</td>
<td>13</td>
</tr>
<tr>
<td>r</td>
<td>0.6</td>
<td>1.4</td>
<td>1.4</td>
<td>2.3</td>
<td>3.5</td>
<td>4.2</td>
<td>5.1</td>
<td>6.5</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;10⁻¹⁶(a)</td>
<td>2 × 10⁻¹³(a)</td>
<td>&lt;10⁻¹⁶(a)</td>
<td>&lt;10⁻¹⁶(a)</td>
<td>&lt;10⁻¹⁶(b)</td>
<td>4 × 10⁻¹⁵(b)</td>
<td>8 × 10⁻¹⁸(b)</td>
<td></td>
</tr>
<tr>
<td>t-peaks (BG02)</td>
<td>1,690</td>
<td>857</td>
<td>506</td>
<td>189</td>
<td>106</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common(S-jumps)</td>
<td>1,546</td>
<td>142</td>
<td>140</td>
<td>70</td>
<td>49</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common(S-jumps) %</td>
<td>17</td>
<td>28</td>
<td>37</td>
<td>46</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected</td>
<td>93 ± 9</td>
<td>55 ± 7</td>
<td>20 ± 4</td>
<td>11 ± 3</td>
<td>3.3 ± 1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expected %</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>8 × 10⁻⁷(b)</td>
<td>&lt;10⁻¹⁶(b)</td>
<td>&lt;10⁻¹⁶(b)</td>
<td>&lt;10⁻¹⁶(b)</td>
<td>5 × 10⁻¹⁰(b)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Normal distributions.
(b) Poisson distributions.

NOTE.—(A) S-jumps are significantly associated with t-peaks. S-jumps significantly coincide with t-peaks of the six cell types (BG02 ESC, fibroblast, lymphoid, erythroid, Hela, and H1 ESC). All, total number of S-jumps in the human genome. For each n value are indicated the numbers of S-jumps close (d < 100 kbp) to the position of t-peaks in n and only n of the 6 timing profiles (see fig. 1A-C). Expected, numbers (mean ± SD) of randomized S-jumps close to n t-peaks of the 6 timing profiles (1,000 simulations, Materials and Methods); r, ratio between the observed number of S-jumps close to a number n of t-peaks and the corresponding expected number; P values, probabilities resulting from the simulations. (B) The BG02 t-peaks common to n among four other cell types are also common to S-jumps. Line 1, number of BG02 t-peaks close to n t-peaks among 4 other timing profiles (fibroblast, lymphoid, erythroid, and H1 ESC); common(S-jumps) (line 2), for each n value, number of these t-peaks (line 1) that coincide with S-jumps; common(S-jumps) %, ratio (line 2)/(line 1) in percent; expected, numbers of BG02 t-peaks common to the 4 other timing profile that coincide with randomized S-jumps ±SD (1,000 simulations, Materials and Methods); P values, probabilities resulting from the simulations.

Upward Jumps of the Skew Profile (S-Jumps) Are Associated With Replication Initiation Zones Potentially Active in Germ Line Cells.

If the S-jumps are associated with replication initiation zones, their distribution is expected to display properties similar to those of the t-peak distributions. We first observe that the S-jumps are significantly close to t-peaks of the 6 timing profiles (at distance d < 100 kbp, fig. 1I; similar distribution as with t-peaks in supplementary fig. 3, Supplementary Material online). Among the 1,546 S-jumps, 863 (set1) coincide (d < 100 kbp) with at least one t-peak in the 6 timing profiles (table 2A). This number is much higher than expected by chance (443, P < 10⁻¹⁵). Among these S-jumps, we distinguish those that coincide with n (1 ≤ n ≤ 6) t-peaks in the 6 timing profiles (table 2A). When n increases, the number of S-jumps associated with t-peaks decreases regularly but is always significantly larger than expected (fig. 1J, in all cases, P value < 8 × 10⁻⁸). Whatever n, the proportions of S-jumps associated with t-peaks in n timing profiles are similar to the proportions of t-peaks in one timing profile associated with t-peaks in n other timing profiles; this is particularly true for the BG02 and Hela timing profiles that both display numbers of t-peaks close to the number of S-jumps (fig. 1H). In addition, the ratio r between the observed and expected S-jumps close to a number n of t-peaks (table 2A) strongly increases with n as previously observed for BG02 t-peaks (table 1A).

We also examined, among the BG02 t-peaks common to n other timing profiles, those that also coincide with S-jumps (table 2B). Their numbers are significantly larger than expected by chance, and their proportion increases with n as observed with Hela t-peaks (compare with table 1B). This confirms that when n increases, the t-peaks common to n timing profiles have a larger probability to be also common to S-jumps, that is, to be associated with initiation zones active in the germ line. Note that the numbers of these t-peaks (table 2B, line 2) are smaller than those obtained with Hela (table 1B, line 2); also, 28 BG02 t-peaks are common to all cell types but only 13 of them coincide with S-jumps (table 2A). This results from the stringent detection process of S-jumps (Brodie et al. 2005; Touchon et al. 2005) that excluded a number of bona fide upward jumps. Indeed, the mean skew profile, in intergenic regions around the 15 t-peaks common to all cell types but that do not coincide with S-jumps, presents a characteristic upward jump similar to that displayed by the set1 S-jumps (as control, the BG02 t-peaks that do not coincide with S-jumps, nor with other t-peaks, display a flat profile; supplementary fig. 4, Supplementary Material online).
Table 3. Substitution Rates Measured on Both Sides of the $S$-jumps.

<table>
<thead>
<tr>
<th>Windows</th>
<th>(+) Introns</th>
<th>Intergenic</th>
<th>(-) Introns</th>
<th>(+) Introns</th>
<th>Intergenic</th>
<th>(-) Introns</th>
</tr>
</thead>
<tbody>
<tr>
<td>A→G</td>
<td>0.3307 ± 0.0079</td>
<td>0.2826 ± 0.0031</td>
<td>0.2134 ± 0.0029</td>
<td>0.3415 ± 0.0042</td>
<td>0.3174 ± 0.0035</td>
<td>0.2409 ± 0.0068</td>
</tr>
<tr>
<td>T→C</td>
<td>0.2620 ± 0.0066</td>
<td>0.3192 ± 0.0034</td>
<td>0.3339 ± 0.0039</td>
<td>0.2164 ± 0.0029</td>
<td>0.2883 ± 0.0032</td>
<td>0.3244 ± 0.0071</td>
</tr>
<tr>
<td>P value</td>
<td>$8 \times 10^{-21}$</td>
<td>$7 \times 10^{-29}$</td>
<td>$&lt;1 \times 10^{-10}$</td>
<td>$&lt;1 \times 10^{-10}$</td>
<td>$&lt;1 \times 10^{-20}$</td>
<td>$&lt;1 \times 10^{-19}$</td>
</tr>
<tr>
<td>C→T</td>
<td>0.3079 ± 0.0060</td>
<td>0.3511 ± 0.0030</td>
<td>0.3188 ± 0.0031</td>
<td>0.3220 ± 0.0030</td>
<td>0.3643 ± 0.0033</td>
<td>0.3392 ± 0.0069</td>
</tr>
<tr>
<td>G→A</td>
<td>0.3354 ± 0.0059</td>
<td>0.3666 ± 0.0032</td>
<td>0.3228 ± 0.0029</td>
<td>0.3183 ± 0.0029</td>
<td>0.3540 ± 0.0030</td>
<td>0.3166 ± 0.0061</td>
</tr>
<tr>
<td>P value</td>
<td>$-0.0275 ± 0.0058$</td>
<td>$-0.0155 ± 0.0029$</td>
<td>$-0.0040 ± 0.0031$</td>
<td>$0.0037 ± 0.0030$</td>
<td>$0.0104 ± 0.0030$</td>
<td>$0.0226 ± 0.0069$</td>
</tr>
<tr>
<td>(++) Introns</td>
<td>476</td>
<td>444</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Measurements are performed in 300 kbp windows (see Materials and Methods). Mean values ± standard error mean (in percent) are computed for the 863 $S$-jumps of set1 (see main text); $\Delta$ is the difference between substitution rates for each pair of complementary substitutions; the statistical significance of $\Delta$ (P value) is computed using a paired t-test; (+) introns are transcribed in the 5′ → 3′ direction of the Watson strand; (-) introns are transcribed in the opposite direction; only windows containing at least 50 kbp of the indicated sequences were retained for analysis; the number of windows retained for each category is indicated in the first line.

The 683 $S$-jumps that do not coincide with t-peaks in any of the 6 timing profiles examined (set0) display a mean S profile similar to that of the other $S$-jumps (fig. 1K). This strongly suggests 1) that the same process has generated the $S$-jumps belonging to both sets and 2) that the set0 $S$-jumps are associated with initiation zones active in germ line cells but not in the other cell types (these “germ line-specific” $S$-jumps represent 44% of all $S$-jumps).

Altogether, these similarities between the $S$-jumps and t-peaks distributions strongly support our hypothesis that the $S$-jumps are associated with initiation zones active in germ line cells. However, because only the set1 $S$-jumps coincide with t-peaks, they will be examined separately from the set0 $S$-jumps in the following sections.

S-Jumps Result From Asymmetric Nucleotide Substitution Patterns

We determined the nucleotide substitution pattern on both sides of each $S$-jump. Substitutions were tabulated in the human lineage since its divergence with chimpanzee using the macaque and orangutan as outgroups (Materials and Methods). Neutral substitution rates were reliably obtained by eliminating coding regions, splice sites, and CpG islands which are unlikely to evolve neutrally. Substitutions were tabulated separately in introns and intergenic regions. The mean values computed for the 863 $S$-jumps of set1 differ in both types of regions (table 3 and fig. 2). In addition, complementary substitution rates differ significantly from each other on each side of the $S$-jump, in intergenic and in intronic regions (except for the A→C and T→G transversions) (fig. 2). The largest differences were observed for the couple of complementary transitions A→G and T→C (fig. 2B). For this couple, the mean difference is positive downstream of $S$-jumps in (+) introns (0.125 ± 0.003%, red), in agreement with a previous study (Green et al. 2003). Surprisingly, this value differs significantly from the asymmetry in (-) introns observed upstream of $S$-jumps (0.069 ± 0.006%). Similarly, the asymmetries in (-) introns (blue) upstream and downstream of $S$-jumps differ from each other. On the same side of the $S$-jumps, asymmetries in (+) and (-) introns do not present opposite values. The asymmetry in (-) introns upstream of $S$-jumps (−0.121 ± 0.003%) has a value opposite to that in (+) introns downstream of $S$-jumps. Similarly, the asymmetry in (-) introns downstream of $S$-jumps (−0.084 ± 0.007%), is opposite to that in (+) introns upstream of $S$-jumps. Asymmetries in intergenic regions present intermediate values (upstream of $S$-jumps: 0.029 ± 0.003%; downstream: −0.037 ± 0.003%). The linear dotted lines, observed for all complementary substitutions (except for the A→C/T→G asymmetries), illustrate a linear relationship between the mean values of the asymmetries in intergenic regions and (+) and (-) introns (fig. 2B):
This relationship is observed on both sides of the S-jumps. Changing from one to the other side of S-jumps corresponds to change $A_{(+)}$ into $-A_{(-)}$ and $A_{\text{int}}$ into $-A_{\text{int}}$ (fig. 2, supplementary fig. 5, Supplementary Material online). Similar patterns are observed when the asymmetries are computed in smaller windows around the S-jumps (supplementary fig. 6A–A'' and B–B'', Supplementary Material online).

We examined whether these substitution rates could have generated the skew pattern observed on both sides of the S-jumps (fig. 3A). We computed the skew $S^*$ that would be produced by the observed substitution rates at equilibrium (i.e., after long evolutionary times; Materials and Methods). The $S^*$ values are strongly correlated with the $S$ values both in intronic and intergenic regions. They change their sign in intergenic regions when crossing the S-jump positions (fig. 3B and C). Similar results were observed when examining the $S$ and $S^*$ profiles around the S-jumps in smaller (50 kbp) windows (supplementary fig. 6C–C'', Supplementary Material online). These data show that the pattern of the skew profile observed in intergenic regions and in the (+) and (−) introns around the S-jump is fully explained by the observed nucleotide substitution asymmetries. The mean $S^*$ values (7.69 ± 0.44% in intergenic regions) are significantly larger than the mean $S$ values (3.72 ± 0.11%) indicating that the compositional equilibrium has not been attained.

$$A_{(+)} - A_{\text{int}} \approx -(A_{(-)} - A_{\text{int}}).$$

**Fig. 2.** Differences between complementary nucleotide substitution rates around S-jumps. (A) Scheme of the analyzed regions; red and blue arrows represent the direction of gene transcription (colors as in fig. 1). (B–G) Differences $\Delta$ between the indicated complementary substitution rates (asymmetries) are shown upstream and downstream of the 863 S-jumps of set1 in colored boxes; red, asymmetries measured in (+) introns, $A_{(+)}$; blue, asymmetries in (−) introns, $A_{(-)}$ black, intergenic regions, $A_{\text{int}}$. $\Delta$ values (in percent ± standard error mean) are computed as in table 3; the dotted lines illustrate the linear relationship: $A_{(+)} - A_{\text{int}} \approx -(A_{(-)} - A_{\text{int}})$.

**Fig. 3.** The compositional skew $S$ is strongly correlated with the skew at equilibrium $S^*$. The skew $S$, (A) and the skew at equilibrium $S^*$, (B) are computed in 300 kbp windows around the 863 S-jumps of set1 (in percent ± standard error mean); to compute $S$, only repeat-masked sequences were considered; colors as in figure 2. (C) Correlation between the skews $S$ and $S^*$ in intergenic regions, upstream (green) and downstream (violet) of the S-jumps; $R = 0.64$, $P < 10^{-16}$ (Pearson). Dotted lines, see figure 2.
Replication Generates Asymmetric Substitution Patterns

What mechanism has generated the mutational strand asymmetries observed around the S-jumps? The equation (1) allows us to propose the following model of superposition of replication- and transcription-associated asymmetries. During DNA replication, the DNA polymerases would generate uncorrected substitution errors at different rates on the two DNA strands. The resulting mutational asymmetries are constant in the vicinity of S-jumps, along genes and intergenic regions, but change their sign when crossing the S-jumps. They are directly observed in intergenic regions (Δ_{int}). In introns, transcription generates mutational asymmetries due to the transcription-coupled repair (TCR) mechanism (Svejstrup 2002) acting on replication-induced mismatches as proposed in Green et al. (2003). The resulting asymmetries, Δ_{trans} in (+) introns and Δ_{trans} in (−) introns, would superimpose to the replication-associated asymmetries to finally generate the observed asymmetries (fig. 4A and B). Following this model, we can deduce the asymmetries associated with replication and transcription (table 4). They all significantly differ from zero except for the A→C/T→G pair. For four pairs, the asymmetries associated with transcription are approximately three times larger than those associated with replication. For the C→T/G→A pair, replication generates an excess of C→T over G→A on the leading strand (Δ_{int} = 0.0130 ± 0.0021), whereas transcription generates an excess of G→A over T→C on the nontemplate strand (Δ_{trans} = − 0.0098 ± 0.0028) (fig. 4A and B). In (+) introns downward of S-jumps, the superimposition of replication and transcription thus leads to ≈0 final value of C→T/G→A asymmetry, and to a large value in (−) introns (fig. 4B). All the observed asymmetries are fully explained in the context of this model.

An alternative model might be proposed based on transcription only. Recent studies have shown that, besides protein coding genes, most genomic DNA is transcribed (Birney et al. 2007). Several types of Pol II noncoding (nc) transcripts have been identified in intergenic regions (Core et al. 2008; Preker et al. 2008; Seila et al. 2008). These types of nc-transcripts might be responsible for the asymmetries in intergenic regions of S-jumps, as suggested in Necsulea et al. (2009). In order to generate asymmetries with the same sign in intergenic regions as in introns diverging from S-jump centre (e.g., A→G/T→C, fig. 2B), these nc transcripts should be mostly oriented divergently from the S-jumps (as (+) introns in fig. 4A). However, to generate the observed C→T/G→A asymmetries, the nc-transcripts should be oriented in the opposite direction (opposed to (+) introns in fig. 4B). This inconsistency allowed us to reject this model based on transcription only, sustaining that different mechanisms have generated the asymmetries observed in intronic and in intergenic regions.

### Table 4. Substitution Rate Asymmetries Associated With Replication and Transcription.

<table>
<thead>
<tr>
<th></th>
<th>Replication</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>r(A→G) − r(T→C)</td>
<td>0.0329 ± 0.0022</td>
<td>0.0939 ± 0.0029</td>
</tr>
<tr>
<td>P value</td>
<td>3 × 10^{−46}</td>
<td>&lt;1 × 10^{−100}</td>
</tr>
<tr>
<td>r(C→T) − r(G→A)</td>
<td>0.0130 ± 0.0021</td>
<td>−0.0098 ± 0.0028</td>
</tr>
<tr>
<td>P value</td>
<td>6 × 10^{−10}</td>
<td>6 × 10^{−14}</td>
</tr>
<tr>
<td>r(A→T) − r(T→A)</td>
<td>0.0029 ± 0.0007</td>
<td>0.0077 ± 0.0009</td>
</tr>
<tr>
<td>P value</td>
<td>9 × 10^{−5}</td>
<td>8 × 10^{−17}</td>
</tr>
<tr>
<td>r(C→G) − r(G→C)</td>
<td>0.0077 ± 0.0013</td>
<td>0.0227 ± 0.0017</td>
</tr>
<tr>
<td>P value</td>
<td>3 × 10^{−9}</td>
<td>4 × 10^{−19}</td>
</tr>
<tr>
<td>r(A→C) − r(T→G)</td>
<td>0.0014 ± 0.0008</td>
<td>−0.0005 ± 0.0011</td>
</tr>
<tr>
<td>P value</td>
<td>0.09</td>
<td>0.64</td>
</tr>
<tr>
<td>r(G→T) − r(C→A)</td>
<td>0.0039 ± 0.0012</td>
<td>0.0130 ± 0.0015</td>
</tr>
<tr>
<td>P value</td>
<td>1 × 10^{−3}</td>
<td>3 × 10^{−17}</td>
</tr>
</tbody>
</table>

NOTE.—Mutation asymmetries, that is, differences between complementary substitution rates, are computed on both sides of the 863 S-jumps of set1; replication, mean values ± standard error mean (in percent) of intergenic asymmetries on the leading strand (Δ_{int}, fig. 4); transcription, difference between mean values of asymmetries on the leading strand computed in intronic and intergenic regions (Δ_{trans}, fig. 4); P values are computed using one-sample (replication) and two-samples (transcription) t-tests; the number of windows are as in table 3.

**Discussion**

In order to support the hypothesis that S-jumps are associated with replication initiation zones, we took advantage of replication timing profiles determined in several cell types. We examined the properties of the distributions of initiation zones identified as peaks in the timing profiles (t-peaks). Large fractions (30–50%) of these initiation zones are active in only one cell type. The fraction of them common to n cell types decreases progressively when n increases, and the probability that they are also active in other cell types increases with n. Although these properties have been observed in only six cell types, they can be tentatively generalized, allowing us to draw a picture of how replication initiation zones are shared by most cell types.
The distribution of the $S$-jump positions along the genome displays all these properties. We thus consider that among the $S$-jumps, those that coincide with $t$-peaks (set1) are associated with initiation zones potentially active in germ line cells. Importantly, the mutational asymmetries measured around the other $S$-jumps (that do not correspond to any $t$-peak, set0) are similar to those measured around the set1 $S$-jumps, consistently with the similarity between the observed compositional skew $S$ and the skew at equilibrium $S^*$ (supplementary fig. 7, Supplementary Material online). These data sustain that the two sets of $S$-jumps originate from the same mutational mechanisms and strongly suggest that the set0 $S$-jumps also correspond to replication initiation zones potentially active in and specific to this cell line.

The substitution rate asymmetries measured in intronic regions change their sign with the direction of transcription as shown previously (Green et al. 2003; Polak and Arndt 2009), but surprisingly their values also depend on their position on one side or the other of the $S$-jumps (e.g., $(+)$ introns in fig. 2). This observation and the relationship between asymmetries illustrated by equation (1) allowed us to propose a model based on the superimposition of replication- and transcription-associated mutational asymmetries. In this model, intergenic asymmetries result from replication. Recent studies have revealed the presence of Pol II transcripts in intergenic regions (Core et al. 2008; Preker et al. 2008; Seila et al. 2008), but these transcripts are mostly expressed at low levels and are generally short sized. The mutational asymmetries that might be generated by these transcripts would thus poorly contribute to those observed in intergenic regions. However, to explain the asymmetries observed for most pairs of complementary substitutions, the nc-transcripts should be abundant (intergenic asymmetries are equal to $1/4$–$1/3$ of intronic asymmetries, table 3). In addition, they also should cover all regions, intergenic as well as intronic, around the $S$-jumps. Finally, they should be mostly divergent from the $S$-jump centre. So far, no data support these scenarios.

In contrast, the replication-based model proposed here is remarkably consistent with the six pairs of complementary substitution rates determined in intergenic regions as well as in $(+)$ and $(-)$ introns, on each side of the replication initiation zones.

This model allows us to estimate the magnitude of the contributions of replication and transcription to the compositional skew profile around the $S$-jumps. In genes, the mutational asymmetries associated with replication and with transcription contribute to 33% and 67% of the skew, respectively (fig. 3A). If we consider the genes presenting the largest skew values (5% largest values), the contribution of transcription raises to 77%. Because the skew positively correlates with the gene expression level (Majewski 2003), this value reflects the contribution of transcription for the most expressed genes. Overall, given the relative proportions of intergenic and of transcribed sequences around $S$-jumps (supplementary fig. 8, Supplementary Material online), the mutational asymmetry associated with replication contributes to about one half (48%) of the compositional skew, whereas transcription itself contributes to the other half.

Accuracy of replication mostly results from the fidelity of DNA polymerases due to the selectivity of the nucleotide selection step followed by the exonucleolytic proofreading step (Brutlag and Kornberg 1972; Bertram et al. 2010) and from the DNA MMR step that suppresses large proportions of mismatched base pairs (Kunkel and Erie 2005). Mutational asymmetries can be generated during replication of a DNA strand by incorporation of wrong nucleotides by the corresponding DNA polymerase (pole for the leading strand and polô for the lagging strand; Kunkel and Burgers 2008) if the rate of a base misincorporation (e.g., G opposite a template T, noted $A\rightarrow G$) differs from the rate of the complementary misincorporation (C opposite to a A, noted $T\rightarrow C$). However, if the level of this asymmetry (difference between misincorporation rates) is the same on the leading and the lagging strands, no mutational asymmetry will finally be generated. For example, the excess of $A\rightarrow G$ over $T\rightarrow C$ substitutions observed on the leading strand in intergenic regions (table 4) can be generated only if:

$$[(A\rightarrow G) - (T\rightarrow C)]_{leading} > [(A\rightarrow G) - (T\rightarrow C)]_{lagging}$$

(2)

where $(A\rightarrow G)_{leading}$ is the rate of $A\rightarrow G$ base misincorporation generated during the synthesis of the leading strand.

First, in vitro studies of the pole and polô fidelities have shown that both polymerases generate excess of $A\rightarrow G$ over $T\rightarrow C$ error rates (Schmitt et al. 2009; Korona et al. 2010). Second, MMR is more efficient on the lagging strand than on the leading strand (Pavlov et al. 2003). Higher reduction of the mispair densities on the lagging strand than on the leading strand would reduce the value of $[(A\rightarrow G) - (T\rightarrow C)]_{lagging}$, contributing to fulfill inequality equation (2). This model relies on different rates of complementary errors by pole and polô as well as on MMR acting on the resulting mismatches with different efficiencies on the leading and lagging strands. In intergenic regions, uncorrected mismatches will not be transcribed and will usually reach the next replication cycle; half of them will be ultimately inherited by the daughter cells. Larger mutational asymmetries are observed in transcribed regions (fig. 2 and table 4). This may result from the mechanism of TCR acting on replication-induced mismatches, as in the model proposed by (Green et al. 2003). The MutSô MMR complex will recognize the mispaired bases, and the TCR machinery will restore a proper base pair using the coding strand as template (Svejstrup 2002). If the misincorporated base is in the coding strand, this will generate a mutated base pair. The large excess of $A\rightarrow G$ over $T\rightarrow C$ substitutions observed in introns would result from a predominantly efficient recognition of G–T mispairs by MutSô (Jiricny 1998).

The model of superimposition of replication- and transcription-associated mutational asymmetries can explain...
all asymmetries of complementary substitutions (table 4). According to this model, for the pair of C → T/G → A transitions, the asymmetries associated with replication and with transcription present opposite signs (fig. 2E and table 4). In this case, a specific mutational mechanism can be proposed. During advance of the replication fork, the discontinuous synthesis of Okazaki fragments temporarily leaves the lagging strand template in single-stranded state (Okazaki et al. 1968). Exposure of cytosines in single-stranded conformation leads to large increase of their spontaneous deamination into uracil (Frederico et al. 1990) which pairs with adenine during replication and generates C → T transitions on the leading strand, as observed in bacteria (Frank and Lory 1999) and mitochondria (Reyes et al. 1998). In parallel, replication errors followed by MMR would produce misincorporations in the nascent DNA strands fulfilling:

\[
[(G \rightarrow A) - (C \rightarrow T)]_{\text{leading}} > [(G \rightarrow A) - (C \rightarrow T)]_{\text{lagging}}.
\] (3)

In intergenic regions, unrepaired misincorporations would generate an excess of G → A over C → T in the leading strand, but deamination of cytosines would compensate for this excess, finally leading to the replication-associated excess of C → T over G → A transition rates, \(A_{\text{int}} = 0.0130 \pm 0.0021\) (table 4). In (+) introns, TCR acting on the mismatches due to the unrepaired misincorporations would increase the corresponding excess of G → A over C → T in the leading strand and produce the transcription-associated excess of G → A over C → T substitution rates, \(A_{\text{trans}} = -0.0098 \pm 0.0028\) (table 4 and fig. 4B). The superimposition of replication- and transcription-associated asymmetries would finally produce the almost null asymmetry observed in (+) introns on the leading strand \(A = 0.0037 \pm 0.0030\), table 3).

In conclusion, we demonstrate for the first time the existence of replication-associated mutational asymmetries in a eukaryotic organism and reevaluate the transcription-associated asymmetries obtained in previous studies (Green et al. 2003). The conservation of human 5-jumps previously observed in the dog and mouse skew profiles (Touchon et al. 2005), raises the possibility that replication-associated mutational asymmetries are conserved among mammals. These data show that in addition to transcription, replication is a major mechanism that shapes human genome composition.

**Supplementary Material**

Supplementary methods, equations, figures 1 – 8, and table 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We would like to thank R. Scott Hansen for providing us with replication timing density data; Gregory Nuel and Stéphane Robin for invaluable help in mathematical analyses, Didier Casane for helpful discussions. This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Agence Nationale de la Recherche (NT05-3_41825) and grants from the Association pour la Recherche sur le Cancer, the Ligue Contre le Cancer (Comité de Paris), and the Fondation pour la Recherche Médicale (équipe labellisée) to O.H.

**References**


