Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins.

Marie Touchon, Samuel Nicolay, Benjamin Audit, Edward-Benedict Brodie of Brodie, Yves D'Aubenton-Carafa, Alain Arnéodo, Claude Thermes

Major category: BIOLOGICAL SCIENCES

Minor category: EVOLUTION

Replication-associated strand asymmetries in mammalian genomes: towards detection of replication origins

Marie Touchon[*], Samuel Nicolay[†‡], Benjamin Audit[†], Edward-Benedict Brodie of Brodie[†], Yves d'Aubenton-Carafa[*], Alain Arneodo[†] and Claude Thermes[*]

[*]Centre de Génétique Moléculaire (CNRS), Allée de la Terrasse, 91198 Gif-sur-Yvette, France

[†]Laboratoire Joliot Curie et Laboratoire de Physique, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France

[‡] Permanent address: Département de Mathématique, Université de Liège, 12 Grande Traverse, 4000 Liège, Belgique

Abstract

In the course of evolution, mutations do not affect both strands of genomic DNA equally. This mainly results from asymmetric DNA mutation and repair processes associated with replication and transcription. In prokaryotes, prevalence of G over C and T over A is frequently observed in the leading strand. The sign of the resulting TA and GC skews changes abruptly when crossing replication origin and termination sites, producing characteristic step-like transitions. In mammals, transcription-coupled skews have been detected, but so far, no bias has been associated with replication. Here, analysis of intergenic and transcribed regions flanking experimentally-identified human replication origins and the corresponding mouse and dog syntenic regions demonstrates the existence of compositional strand asymmetries associated with replication. Multi-scale analysis of human genome skew profiles reveals numerous transitions that allow us to identify a set of one thousand putative replication initiation zones. Around these putative origins, the skew profile displays a characteristic jagged pattern also observed in mouse and dog genomes. We therefore propose that in mammalian cells, replication termination sites are randomly distributed between adjacent origins. Altogether, these analyses constitute a step toward genome-wide studies of replication mechanisms.

INTRODUCTION

Comprehensive knowledge of genome evolution relies on understanding mutational processes that shape DNA sequences. Nucleotide substitutions do not occur at similar rates and in particular, owing to strand asymmetries of the DNA mutation and repair processes, they can affect each of the two DNA strands differently. Asymmetries of substitution rates coupled to transcription have been observed in prokaryotes (1-3) and in eukaryotes (4-6). Strand asymmetries (*i.e.* $G \neq C$ and $T \neq A$) associated with the polarity of replication have been found in bacterial, mitochondrial and viral genomes where they have been used to detect replication origins (7-9). In most cases, the leading replicating strand presents an excess of G over C and of T over A. Along one DNA strand, the sign of this bias changes abruptly at the replication origin and at the terminus. In eukaryotes, the situation is unclear. Several studies failed to show compositional biases related to replication and analyses of nucleotide substitutions in the region of the *ß-globin* replication origin did not support the existence of mutational bias between the leading and the lagging strands (8, 10, 11). In contrast, strand asymmetries associated with replication were observed in the subtelomeric regions of *Saccharomyces cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism (12). We present here analyses of strand asymmetries flanking experimentally-determined human replication origins, as well as the corresponding mouse and dog syntenic regions. Our results

demonstrate the existence of replication-coupled strand asymmetries in mammalian genomes. Multi-scale analysis of skew profiles of the human genome using the wavelet transform methodology, reveals the existence of numerous putative replication origins associated with randomly distributed termination sites.

Data and Methods

Human replication origins. Nine replication origins were examined, namely those situated near the genes *MCM4* (13), *HSPA4* (14), *TOP1* (15), *MYC* (16), *SCA-7* (17), *AR* (17), *DNMT1* (18), *LaminB2* (19) and *ß-globin* (20).

Sequences. Sequence and annotation data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC) for the human (May 2004), mouse (May 2004) and dog (July 2004) genomes. To delineate the most reliable intergenic regions, transcribed regions were retrieved from "all_mrna", one of the largest sets of annotated transcripts. To obtain intronic sequences, we used the KnownGene annotation (containing only protein-coding transcripts); when several transcripts presented common exonic regions, only common intronic sequences were retained. For the dog genome, only preliminary gene annotations were available, precluding the analysis of intergenic and intronic sequences. To avoid biases intrinsic to repeated elements, all sequences were masked with RepeatMasker, leading to 40-50% sequence reduction.

Strand asymmetries. The TA and GC skews were calculated as $S_{TA} = (T - A)/(T + A)$, $S_{GC} = (G - C)/(G + C)$ and the total skew as $S = S_{TA} + S_{GC}$, in non-overlapping 1 kbp windows (all values are given in percent). The cumulated skew profiles Σ_{TA} and Σ_{GC} were obtained by cumulative addition of the values of the skews along the sequences. To calculate the skews in transcribed regions, only central regions of introns were considered (after removal of 530 nt from each extremity) in order to avoid the skews associated with splicing signals (6). To calculate the skews in intergenic regions, only windows that did not contain any transcribed region were retained. To eliminate the skews associated with promoter signals and with transcription downstream of polyA sites, transcribed sequences were extended by 0.5 kbp and 2 kbp at 5' and 3' extremities, respectively (6).

Sequence alignments. Mouse and dog regions syntenic to the six human regions shown in Fig. 1 were retrieved from UCSC (Human Synteny). Mouse intergenic sequences were individually aligned using PipMaker (21) leading to a total of 150 conserved segments larger than 100 bp (> 70% identity) corresponding to a total of 26 kbp (5.3% of intergenic sequences).

Wavelet-based analysis of the human genome. The wavelet transform (WT) methodology is a multi-scale discontinuities tracking technique (22, 23) (for details, see Supplementary material). The main steps involved in detection of jumps were the following. We selected the extrema of the first derivative S' of the skew profile S smoothened at large scale (*i.e.* computed in large windows). The scale 200 kbp was chosen as being just large enough to reduce the contribution of discontinuities associated with transcription (*i.e.* larger than most human genes (24)), yet as small as possible so as to capture most of the contributions associated with replication. In order to delineate the position corresponding to the jumps in the skew S at smaller scale, we then progressively decreased the size of the analyzing window and followed the positions of the extrema of S' across the whole range of scales down to the shortest scale analyzed (the precision was limited by the noisy background fluctuations in the skew profile). As expected, the set of extrema detected by this methodology corresponded to similar numbers of upward and downward jumps. The putative replication origins were then selected among the set of upward jumps on the basis of their ΔS amplitude (see text).

RESULTS AND DISCUSSION

Strand asymmetries associated with replication. We examined the nucleotide strand asymmetries around 9 replication origins experimentally-determined in the human genome (Data and Methods). For most of them, the S skew measured in the regions situated 5' to the origins on the Watson strand (lagging strand) presented negative values that shifted abruptly (over few kbp) to positive values in

regions situated 3' to the origins (leading strand), displaying sharp upward transitions with large ΔS amplitudes as observed in bacterial genomes (7-9) (Fig. 1a). This was particularly clear with the cumulated TA and GC skews that presented decreasing (increasing) profiles in regions situated 5' (3') to the origins, displaying characteristic V-shapes pointing to the initiation zones. These profiles could, at least in part, result from transcription, as shown in previous work (6). To measure compositional asymmetries that would result only from replication, we calculated the skews in intergenic regions on both sides of the origins. The mean intergenic skews shifted from negative to positive values when crossing the origins (Fig. 2). This result strongly suggested the existence of mutational pressure associated with replication, leading to the mean compositional biases S_{TA} = 4.0 ± 0.4% and S_{GC} = 3.0 ± 0.5% (note that the value of the skew could vary from one origin to another, possibly reflecting different initiation efficiencies) (Table 1). In transcribed regions, the S bias presented large values when transcription was co-oriented with replication fork progression ((+) genes on the right, (-) genes on the left), and close to zero values in the opposite situation (Fig. 2). In these regions, the biases associated with transcription and replication added to each other when transcription was co-oriented with replication fork progression, giving the skew S_{Lead}; they subtracted from each other in the opposite situation, giving the skew S_{lag} (Table 1). We could estimate the mean skews associated with transcription by subtracting intergenic skews from S_{Lead} values, giving S_{TA} = 3.6 ± 0.7% and S_{GC} = 3.8 ± 0.9%. These estimations were consistent with those obtained with a large set of human introns S_{TA} = 4.49 ± 0.01% and S_{GC} = 3.29 ± 0.01% in ref. (6), further supporting the existence of replication-coupled strand asymmetries.

	S_{TA}	S_{GC}	S	l	(G+C)%
intergenic (*H.s.*) all	3.9±0.4	3.0±0.4	6.9±0.4	487	42
intergenic (*H.s.*) ncr.	4.0±0.4	3.0±0.5	7.0±0.5	461	42
intergenic (*M.m.*) ncr.	3.6±0.4	2.2±0.5	5.8±0.5	441	42
S_{Lead} (*H.s.* introns)	7.5±0.3	6.8±0.4	14.3±0.4	358	40
S_{lag} (*H.s.* introns)	-1.9±1.0	-0.3±1.4	-2.2±1.3	49	44

Table 1. Strand asymmetries associated with human replication origins. The skews were calculated in the regions flanking the six human replication origins (Fig. 1a) and in the corresponding syntenic regions of the mouse genome. Intergenic sequences were always considered in the direction of replication fork progression (leading strand); they were considered in totality (all) or after elimination of conserved regions (ncr.) between human (*H.s.*) and mouse (*M.m.*) (see Data and Methods). To calculate the mean skew in introns, the sequences were considered on the non-transcribed strand: S_{Lead}, the orientation of transcription was the same as the replication fork progression; S_{lag}, opposite situation. The mean values of the skews S_{TA}, S_{GC} and S are given in % (± SEM); l, total sequence length in kbp.

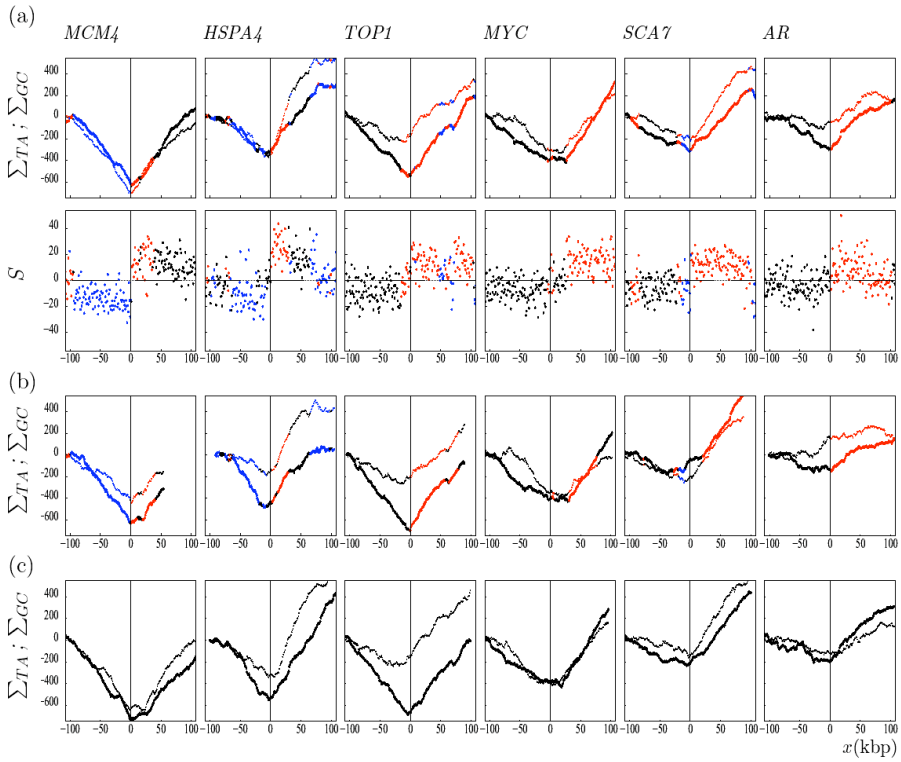

Fig. 1. TA and GC skew profiles around experimentally-determined human replication origins. (a) The skew profiles were determined in 1 kbp windows in regions surrounding (± 100 kbp without repeats) experimentally-determined human replication origins (Data and Methods). First row, TA and GC cumulated skew profiles Σ_{TA} (thick line) and Σ_{GC} (thin line). Second row, skew S calculated in the same regions. The ΔS amplitude associated with these origins, calculated as the difference of the skews measured in 20 kbp windows on both sides of the origins, are: *MCM4* (31%), *HSPA4* (29%), *TOP1* (18%), *MYC* (14%), *SCA7* (38%), *AR* (14%). (b) Cumulated skew profiles calculated in the 6 regions of the mouse genome syntenic to the human regions figured in (a). (c) Cumulated skew profiles in the 6 regions of the dog genome syntenic to human regions figured in (a). Abscissa (x) represents the distance (kbp) of a sequence window to the corresponding origin; ordinate represents the values of S given in percent; red, (+) genes (coding strand identical to the Watson strand); blue, (-) genes (coding strand opposite to the Watson strand); black, intergenic regions; in (c) genes are not represented.

Could the biases observed in intergenic regions result from the presence of as yet undetected genes? Two reasons argued against this possibility. First, we retained as transcribed regions one of the largest sets of transcripts available, resulting in a stringent definition of intergenic regions. Second, several studies have demonstrated the existence of hitherto unknown transcripts in regions where no protein coding genes had been previously identified (25-28). Taking advantage of the set of non-protein-coding RNAs identified in the "H-Inv" database (29), we checked that none of them was present in the intergenic regions studied here. Another possibility was that the skews observed in intergenic regions result from conserved DNA segments. Indeed, comparative analyses have shown the presence of non-genic sequences conserved in human and mouse (30). These could present biased sequences, possibly contributing to the observed intergenic skews. We examined the mouse genome regions syntenic to the six human replication zones (Fig. 1b). Alignment of corresponding intergenic regions revealed the presence of homologous segments, but these accounted for only 5.3 % of all intergenic sequences. Removal of these segments did not change significantly the skew in intergenic regions, therefore eliminating the possibility that intergenic skews are due to conserved sequence elements (Table 1).

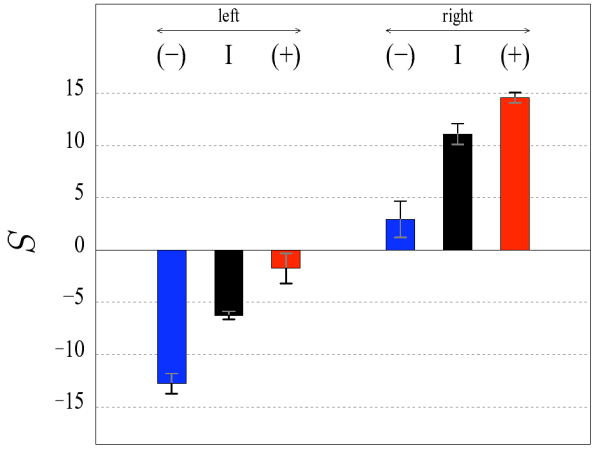

Fig. 2. Skew S in regions situated on both sides of human replication origins. The mean values of S were calculated in intergenic regions and in intronic regions situated 5' (left) and 3' (right) of the six origins analyzed in Fig. 1a; colors are as in Fig. 1; mean values are in percent ± SEM.

Conservation of replication-coupled strand asymmetries in mammalian genomes. We analyzed the skew profiles in DNA regions of mammalian genomes syntenic to the six human origins (Fig. 1). The human, mouse and dog profiles were strikingly similar to each other, suggesting that in mouse and dog, these regions also corresponded to replication initiation zones (indeed, they were very similar in primate genomes). Examination of mouse intergenic regions showed, as for human, significant skew S values with opposite signs on each side of these putative origins, suggesting the existence of a compositional bias associated with replication $S = 5.8 \pm 0.5\%$ (Table 1). Human and mouse intergenic sequences situated at these homologous loci presented significant skews, even though they presented almost no conserved sequence elements. This presence of strand asymmetry in regions that strongly diverged from each other during evolution further supported the existence of compositional bias associated with replication in both organisms: in the absence of such process, intergenic sequences would have lost a significant fraction of their strand asymmetry.

Altogether, these results establish, in mammals, the existence of strand asymmetries associated with replication in germ-line cells. They determine that most replication origins experimentally-detected in somatic cells coincide with sharp upward transitions of the skew profiles. The results also imply that for the majority of experimentally-determined origins, the positions of initiation zones are conserved in mammalian genomes (a recent work confirmed the presence of a replication origin in the mouse *MYC* locus (31)). Among nine human origins examined, three do not present typical V-type cumulated profiles. For the first one (*DNMT1*), the central part of the V-profile is replaced by a large horizontal plateau (several tens of kbp) possibly reflecting the presence of several origins dispersed over the whole plateau. Dispersed origins have been observed for example in the hamster *DHFR* initiation zone (32). By contrast, the skew profiles of the *LaminB2* and *ß-globin* origins present no upward transition suggesting that they might be inactive in germ-line cells, or less active than neighboring origins (data not shown).

Detection of putative replication origins. Human experimentally-determined replication origins coincided with large amplitude upward transitions of skew profiles. The corresponding ΔS ranged between 14% and 38% owing to possible different replication initiation efficiencies and/or different contributions of transcriptional biases (Fig. 1a). Are such discontinuities frequent in human sequences, and can they be considered as diagnostic of replication initiation zones? In particular, can they be distinguished from the transitions associated with transcription only? Indeed, strand asymmetries associated with transcription can generate sharp transitions in the skew profile at both gene extremities. These jumps are of same amplitude and of opposite signs, *e.g.* upward (downward) jumps at 5' (3') extremities of (+) genes (6). Upward jumps resulting from transcription only, might thus be confused with upward jumps associated with replication origins.

To address these questions, systematic detection of discontinuities in the S profile was performed with the wavelet transform methodology, leading to a set of 2415 upward jumps and, as expected, to a similar number of downward jumps (see Data and Methods). The distributions of the $|\Delta S|$ amplitude of these jumps were then examined, showing strong differences between upward and downward jumps. For large $|\Delta S|$ values, the number of upward jumps exceeded by far the number of downward jumps (Fig. 3). This excess likely resulted from the fact that, contrasting with prokaryotes where downward jumps result from precisely positioned replication termination, in eukaryotes, termination appears not to occur at specific positions but to be randomly distributed (this point will be detailed in the last section) (33, 34). Accordingly, the small number of downward jumps with large $|\Delta S|$ resulted from transcription, not replication. These jumps were due to highly biased genes that also generated a small number of large amplitude upward jumps, giving rise to false positive candidate replication origins. The number of large downward jumps was thus taken as an estimation of the number of false positives. In a first step, we retained as acceptable a proportion of 33% of false positives. This value resulted from the selection of upward and downward jumps presenting an amplitude $|\Delta S| \geq 12.5\%$, corresponding to a ratio of downward jumps over upward jumps $r = 0.33$. The values of this ratio r were highly variable along the chromosomes (Fig. 3). In G+C-poor regions ($G+C < 37\%$) we observed the smallest r values ($r = 0.15$). In regions with $37\% \leq G+C \leq 42\%$, we obtained $r = 0.24$, contrasting with $r = 0.53$ in regions with $G+C > 42\%$. In these latter regions (accounting for about 40% of the genome) with high gene density and small gene length (24), the skew profiles oscillated rapidly with large upward and downward amplitudes (Fig. 5d) resulting in a too large number of false positives (53%). In a final step, we retained as putative origins upward jumps (with $|\Delta S| \geq 12.5\%$) detected in regions with $G+C \leq 42\%$. This led to a set of 1012 candidates among which we could estimate the proportion of true replication origins to 79% ($r = 0.21$, Fig. 3a).

The mean amplitude of the jumps associated with the 1012 putative origins was 18%, consistent with the range of values observed for the six origins in Fig. 1. Note that these origins were all found in the detection process. In close vicinity of the 1012 putative origins (± 20kbp) most DNA sequences (55 % of the analyzing windows) are transcribed in the same direction as the progression of the replication fork. By contrast, only 7% of sequences are transcribed in the opposite direction (38% are intergenic). These results show that the $|\Delta S|$ amplitude at putative origins mostly results from superposition of biases (*i*) associated with replication and (*ii*) with transcription of the gene proximal to the origin. Whether transcription is co-oriented with replication at larger distances will require further studies.

We then determined the skews of intergenic regions on both sides of these putative origins. As shown in Fig. 4, the mean skew profile calculated in intergenic windows shift abruptly from negative to positive values when crossing the jump positions. To avoid the skews that could result from incompletely annotated gene extremities (*e.g.* 5' and 3' UTRs), 10 kbp sequences were removed at both ends of all annotated transcripts. The removal of these intergenic sequences did not significantly modify the skew profiles indicating that the observed values do not result from transcription. On both sides of the jump, we observed a steady decrease of the bias, with some flattening of the profile close to the transition point. Note that, due to (*i*) the potential presence of signals implicated in replication initiation, and (*ii*) the possible existence of dispersed origins (32), one might question the meaningfulness of this flattening that leads to a significant underestimate of the jump amplitude. As shown in Fig. 4, extrapolating the linear behavior observed at distance from the jump would lead to a skew of 5.3%, a value consistent with the skew measured in intergenic regions around the six origins (7.0 ± 0.5%, Table 1). Overall, the detection of upward jumps with characteristics similar to those of experimentally-determined replication origins and with no downward counterpart, further support the existence, in human chromosomes, of replication-coupled strand asymmetries, leading to the identification of numerous putative replication origins active in germ-line cells.

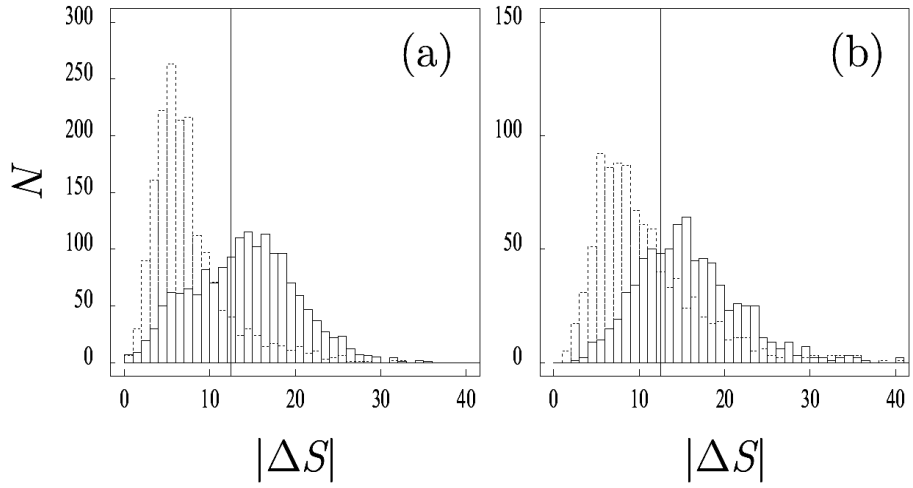

Fig. 3. Histograms of the $|\Delta S|$ amplitudes of the jumps in the S profile. Using the wavelet transform, a set of 5101 discontinuities was detected (2415 upward jumps and 2686 downward jumps, Data and Methods). The $|\Delta S|$ amplitude was calculated as in Fig. 1a. (a) $|\Delta S|$ distributions of the jumps presenting $G+C < 42\%$, corresponding to 1647 upward jumps and 1755 downward jumps; the threshold $|\Delta S| \geq 12.5\%$ (vertical line) corresponded to 1012 upward jumps that were retained as putative replication origins, and to 211 downward jumps ($r = 0.21$). (b) $|\Delta S|$ distributions of the jumps presenting $G+C > 42\%$, $|\Delta S| \geq 12.5\%$ corresponding to 528 upward jumps and 280 downward jumps ($r = 0.53$). The $G+C$ content was measured in the 100 kbp window surrounding the jump position. Upward jumps (black); downward jumps (dots); abscissa represents the values of the $|\Delta S|$ amplitudes calculated in percent.

Random replication termination in mammalian cells. In bacterial genomes, the skew profiles present upward and downward jumps at origin and termination positions, respectively, separated by constant S values (7-9). Contrasting with this step-like shape, the S profiles of intergenic regions surrounding putative origins did not present downward transitions, but decreased progressively in the 5' to 3' direction on both sides of the upward jump (Fig. 4). This pattern was typically found along S profiles of large genome regions showing sharp upward jumps connected to each other by segments of steadily decreasing skew (Fig. 5 a-c). The succession of these segments, presenting variable lengths, displayed a jagged motif reminiscent of the shape of "factory roofs" which was observed around the experimentally-determined human origins (Fig. 5a and data not shown), as well as around a number of putative origins (Fig. 5 b, c). Some of these segments were entirely intergenic (Fig. 5 a, c), clearly illustrating the particular profile of a strand bias resulting solely from replication. In most other cases, we observed the superposition of this replication profile and of the transcription profile of (+) and (-) genes, appearing as upward and downward blocks standing out from the replication pattern (Fig. 5c). Overall, this jagged pattern could not be explained by transcription only, but was perfectly explained by termination sites more or less homogeneously distributed between successive origins. Although some replication terminations have been found at specific sites in *S. pombe* (35), they occur randomly between active origins in *S. cerevisiae* and in *Xenopus* egg extracts (33, 34). Our results indicate that this property can be extended to replication in human germ-line cells. According to our results, we propose a scenario of replication termination relying on the existence of numerous termination sites distributed along the sequence (Fig. 6). For each termination site (used in a small proportion of cell cycles), strand asymmetries associated with replication will generate a skew profile with a downward jump at the position of termination and upward jumps at the positions of the adjacent origins, separated by constant values (as in bacteria). Various termination positions will correspond to elementary skew profiles (Fig. 6, first column).

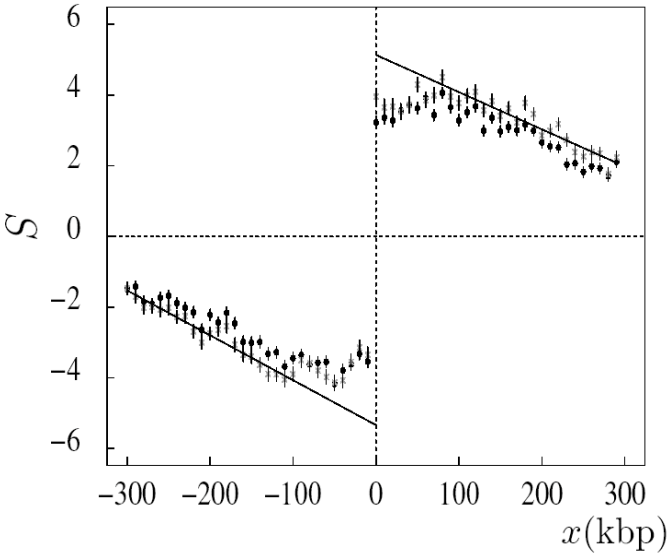

Fig. 4. Mean skew profile of intergenic regions around putative replication origins. The skew *S* was calculated in 1 kbp windows (Watson strand) around the position (± 300 kbp without repeats) of the 1012 upward jumps (Fig. 3); 5' and 3' transcript extremities were extended by 0.5 and 2 kbp, respectively (full circles) or by 10 kbp at both ends (stars) (Data and Methods). Abscissa represents the distance (kbp) to the corresponding origin; ordinate represents the skews calculated for the windows situated in intergenic regions (mean values for all discontinuities and for ten consecutive 1 kbp window positions); the skews are given in percent (vertical bars, SEM). The lines correspond to linear fits of the values of the skew (stars) for $x < -100$ kbp and $x > 100$ kbp.

Addition of these profiles will generate the intermediate profile (second column) and further addition of many elementary skews will generate the final profile (third column). In a simple picture, we can suppose that termination occurs with constant probability at any position on the sequence. This can result from the binding of some termination factor at any position between successive origins, leading to a homogeneous distribution of termination sites during successive cell cycles. The final skew profile is then a linear segment decreasing between successive origins (Fig. 6, third column, black line). In a more elaborate scenario, termination would take place when two replication forks collide. This would also lead to various termination sites, but the probability of termination would then be maximum at the middle of the segment separating neighboring origins, and decrease towards extremities. Considering that firing of replication origins occurs during time intervals of the S phase (36) could result in some flattening of the skew profile at the origins, as sketched in Fig. 6 (third column, grey curve). In the present state, our results clearly support the hypothesis of random replication termination in mammalian cells, but further analyses will be necessary to determine what scenario is precisely at work.

Importantly, the "factory roof" pattern was not specific to human sequences, but it was also observed in numerous regions of the mouse and dog genomes (*e.g.* Fig. 5 e, f)) indicating that random replication termination is a common feature of mammalian germ-line cells. Moreover, this pattern was displayed by a set of one thousand upward transitions, each flanked on each side by DNA segments of appriximately 300 kbp (without repeats), which can be roughly estimated to correspond to 20-30% of the human genome. In these regions, characterized by low and medium G+C contents, the skew profiles revealed a portrait of germ-line replication, consisting of putative origins separated by long DNA segments of about 1-2 Mbp long. Although such segments are much larger than could be expected from the classical view of $\approx$ 50-300 kbp long replicons (37), they are not incompatible with estimations showing that replicon size can reach up to 1 Mbp (38, 39) and that replicating units in meiotic chromosomes are much longer than those engaged in somatic cells (40). Finally, it is not unlikely that in G+C-rich (gene-rich) regions, replication origins would be closer to each other than in other regions, further explaining the greater difficulty in detecting origins in these regions.

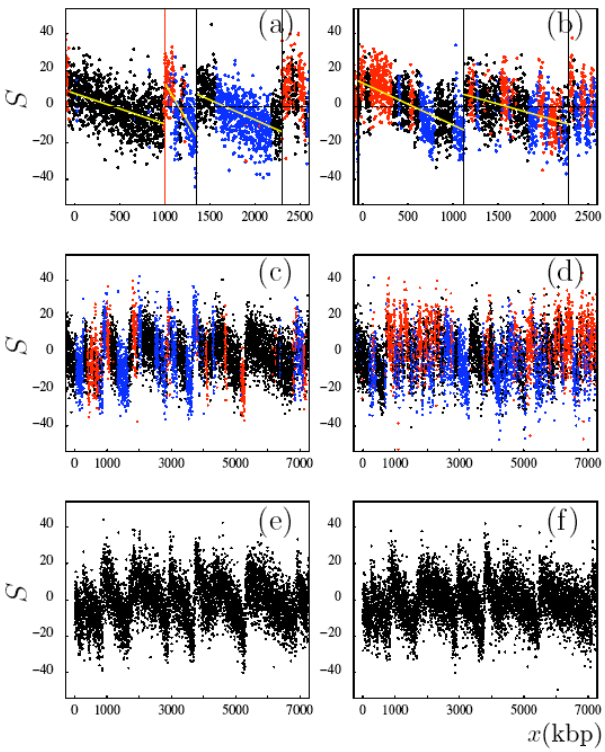

Fig. 5. S profiles along mammalian genome fragments. (a) Fragment of chr. 20 including the *TOP1* origin (red vertical line); (b), (c), chr. 4 and chr. 9 fragments, respectively, with low $G+C$ content (36%); (d) chr. 22 fragment with larger $G+C$ content (48%). In (a) and (b), vertical lines correspond to selected putative origins; yellow lines, linear fits of the S values between successive putative origins. Black, intergenic regions; red, (+) genes; blue, (-) genes; note the fully intergenic regions upstream of *TOP1* in (a) and from positions 5290 to 6850 kbp in (c). (e) Fragment of mouse chr. 4 syntenic to the human fragment shown in (c); (f) fragment of dog chr. 5 syntenic to the human fragment shown in (c); in (e) and (f), genes are nor represented.

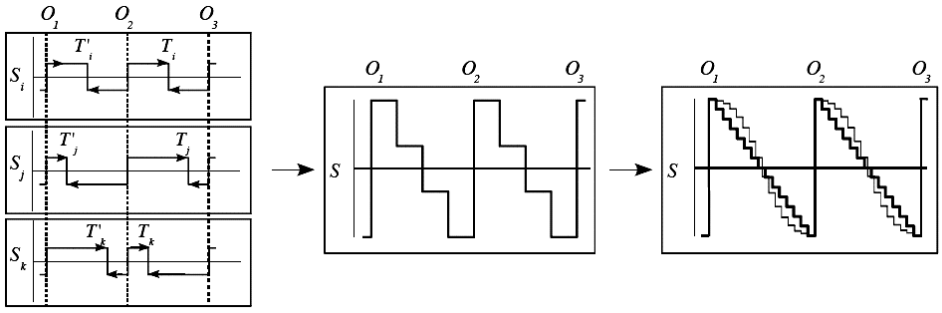

Fig. 6. Model of replication termination. Schematic representation of the skew profiles associated with three replication origins O_1, O_2, O_3; we suppose that these are adjacent, bidirectional origins with similar replication efficiency; abscissa represent the sequence position; ordinate represent the S values (arbitrary units); upward (downward) steps correspond to origin (termination) positions; for convenience the termination sites are symmetric relative to O_2. First column, three different termination positions T_i, T_j, T_k, leading to elementary skew profiles S_i, S_j, S_k; second column, superposition of these 3 profiles; third column, superposition of a large number of elementary profiles leading to the final "factory roof" pattern. Simple model: termination occurs with equal probability on both sides of the origins leading to the linear profile (3rd column, thick line). Alternative model:

replication termination is more likely to occur at lower rates close to the origins, leading to a flattening of the profile (3rd column, grey line).

In conclusion, analyses of strand asymmetries demonstrate the existence of mutational pressure acting asymmetrically on the leading and lagging strands during successive replicative cycles of mammalian germ-line cells. Analyses of the sequences of human replication origins show that most of these origins, determined experimentally in somatic cells, are likely to be active also in germ-line cells. In addition, the results reveal that the positions of these origins are conserved in mammalian genomes. Finally, multi-scale studies of skew profiles allow us to identify a large number (1012) of putative replication initiation zones and provide a genome-wide picture of replication initiation and termination in germ-line cells.

Acknowledgements

This work was supported by the ACI IMPBIO 2004, the Centre National de la Recherche Scientifique (CNRS), the French Ministère de l'Education et de la Recherche and the PAI Tournesol. We thank O. Hyrien for very helpful discussions.

References

1. Freeman, J. M., Plasterer, T. N., Smith, T. F. & Mohr, S. C. (1998) *Science* **279,** 1827-1830.
2. Beletskii, A., Grigoriev, A., Joyce, S. & Bhagwat, A. S. (2000) *J. Mol. Biol.* **300,** 1057-1065.
3. Francino, M. P. & Ochman, H. (2001) *Mol. Biol. Evol.* **18,** 1147-1150.
4. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. (2003) *Nat. Genet.* **33,** 514-517.
5. Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2003) *FEBS Lett.* **555,** 579-582.
6. Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2004) *Nucleic Acids Res.* **32,** 4969-4978.
7. Lobry, J. R. (1996) *Mol. Biol. Evol.* **13,** 660-665.
8. Mrazek, J. & Karlin, S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3720-3725.
9. Tillier, E. R. & Collins, R. A. (2000) *J. Mol. Evol.* **50,** 249-257.
10. Bulmer, M. (1991) *J. Mol. Evol.* **33,** 305-310.
11. Francino, M. P. & Ochman, H. (2000) *Mol. Biol. Evol.* **17,** 416-422.
12. Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M. R. & Cebrat, S. (2000) *J. Theor. Biol.* **202,** 305-314.
13. Ladenburger, E. M., Keller, C. & Knippers, R. (2002) *Mol. Cell. Biol.* **22,** 1036-1048.
14. Taira, T., Iguchi-Ariga, S. M. & Ariga, H. (1994) *Mol. Cell. Biol.* **14,** 6386-6897.
15. Keller, C., Ladenburger, E. M., Kremer, M. & Knippers, R. (2002) *J. Biol. Chem.* **277,** 31430-31440.
16. Vassilev, L. & Johnson, E. M. (1990) *Mol. Cell. Biol.* **10,** 4899-4904.
17. Nenguke, T., Aladjem, M. I., Gusella, J. F., Wexler, N. S. & Arnheim, N. (2003) *Hum. Mol. Genet.* **12,** 1021-1028.
18. Araujo, F. D., Knox, J. D., Ramchandani, S., Pelletier, R., Bigey, P., Price, G., Szyf, M. & Zannis-Hadjopoulos, M. (1999) *J. Biol. Chem.* **274,** 9335-9341.
19. Giacca, M., Zentilin, L., Norio, P., Diviacco, S., Dimitrova, D., Contreas, G., Biamonti, G., Perini, G., Weighardt, F., Riva, S., *et al.* (1994) *Proc. Natl. Acad. Sci. USA* **91,** 7119-7123.
20. Kitsberg, D., Selig, S., Keshet, I. & Cedar, H. (1993) *Nature* **366,** 588-590.
21. Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000) *Genome Res.* **10,** 577-586.
22. Arneodo, A., Audit, B., Decoster, N., Muzy, J. F. & Vaillant, C. (2002) in *The Science of Disaster* (Springer, Berlin), pp. 27-102.
23. Nicolay, S., Brodie of Brodie, E. B., Touchon, M., d'Aubenton Carafa, Y., Thermes, C. & Arneodo, A. (2004) *Physica A* **342,** 270-280.

24. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860-921.
25. Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002) *Science* **296,** 916-919.
26. Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J. D. & Wang, S. M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 12257-12262.
27. Rinn, J. L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N. M., Hartman, S., Harrison, P. M., Nelson, F. K., Miller, P., Gerstein, M., *et al.* (2003) *Genes Dev.* **17,** 529-540.
28. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004) *Genome Res.* **14,** 331-342.
29. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., *et al.* (2004) *PLoS Biol* **2,** e162.
30. Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V., *et al.* (2002) *Nature* **420,** 578-582.
31. Girard-Reydet, C., Gregoire, D., Vassetzky, Y. & Mechali, M. (2004) *Gene* **332,** 129-138.
32. Vassilev, L. T., Burhans, W. C. & DePamphilis, M. L. (1990) *Mol. Cell. Biol.* **10,** 4685-4689.
33. Santamaria, D., Viguera, E., Martinez-Robles, M. L., Hyrien, O., Hernandez, P., Krimer, D. B. & Schvartzman, J. B. (2000) *Nucleic Acids Res.* **28,** 2099-2107.
34. Little, R. D., Platt, T. H. & Schildkraut, C. L. (1993) *Mol. Cell. Biol.* **13,** 6600-6613.
35. Codlin, S. & Dalgaard, J. Z. (2003) *EMBO J.* **22,** 3431-3440.
36. White, E. J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E. J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 17771-17776.
37. Huberman, J. A. & Riggs, A. D. (1968) *J. Mol. Biol.* **32,** 327-341.
38. Yurov, Y. B. & Liapunova, N. A. (1977) *Chromosoma* **60,** 253-267.
39. Berezney, R., Dubey, D. D. & Huberman, J. A. (2000) *Chromosoma* **108,** 471-484.
40. Callan, H. G. (1972) *Proc. R. Soc. Lond.* 181, 19-41.

Supplementary material

Detection of jumps in skew profiles using the continuous wavelet transform.

For effective detection of jumps or discontinuities, the simple intuitive idea is that these jumps are points of strong variation in the signal that can be detected as maxima of the modulus of the (regularized) first derivative of the signal. In order to avoid confusion between "true" maxima of the modulus and maxima induced by the presence of a noisy background, the rate of signal variation has to be estimated using a sufficiently large number of signal samples. This can be achieved using the continuous wavelet transform (WT) that provides a powerful framework for the estimation of signal variations over different length scales. The WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet, by means of dilations and translations (22, 23). When using the first derivative of the Gaussian function, namely $g^{(1)}(x) = -\mathrm{d}g^{(0)}(x)/\mathrm{d}x$, with $g^{(0)}(x) = e^{-x^2/2}$, then the WT of the skew profile *S* takes the following expression:

$$T_{g^{(1)}}[S](x,a) = \frac{1}{a}\int_{-\infty}^{+\infty} S(y) g^{(1)}\left(\frac{y-x}{a}\right) dy = \frac{\mathrm{d}}{\mathrm{d}x}\, g_a^{(0)} * S\,(x,a) \quad , \qquad (1)$$

where *x* and *a* (> 0) are the space and scale parameters, respectively. Equation (1) shows that the WT computed with $g^{(1)}$ is the derivative of the signal *S* smoothened by a dilated version $g_a^{(0)}(x) = g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique (22, 23). The basic principle of the detection of jumps in the skew profiles with the WT is illustrated in Figure S1. From equation (1), it is obvious that at any fixed scale *a*, a large value of the modulus of the WT coefficient corresponds to a large value of the derivative of the skew profile smoothened at that scale. In particular, jumps manifest as local maxima of the WT modulus as illustrated for three different scales in Figure S1 (middle). The main issue when dealing with noisy signals like the skew profile in Figure S1 (top) is to distinguish between the local WT modulus maxima (WTMM) associated with the jumps and those induced by the noise. In this respect, the freedom in the choice of the smoothing scale *a* is fundamental, since the noise amplitude is reduced when increasing the smoothing scale, while an isolated jump contributes equally at all scales. As shown in Figure S1 (bottom), our methodology consists in computing the WT skeleton defined by the set of maxima lines obtained by connecting the WTMM across scales. Then, we select a scale *a* large enough to reduce the effect of the noise, yet small enough to take into account the typical distance between jumps. The maxima lines that exist at that scale are likely to point to jump positions at small scale. The detected jump locations are estimated as the positions at scale 20 kbp of the so-selected maxima lines. According to equation (1), upward (resp. downward) jumps are identified by the maxima lines corresponding to positive (resp. negative) values of the WT as illustrated in Figure S1 (bottom) by the black (resp. red) lines. For the considered fragment of human chromosome 12, we have thus identified 7 upward and 8 downward jumps. The amplitude of the WTMM actually measures the relative importance of the jumps compared to the overall signal. The black dots in Figure S1 (middle) correspond to the 5 WTMM of largest amplitude ($|\Delta S| \geq 12.5\%$); it is clear that the associated maxima lines point to the 5 major jumps in the skew profile. Note that these are 5 upward jumps with no downward counterpart and that they have been reported as 5 putative replication origins.

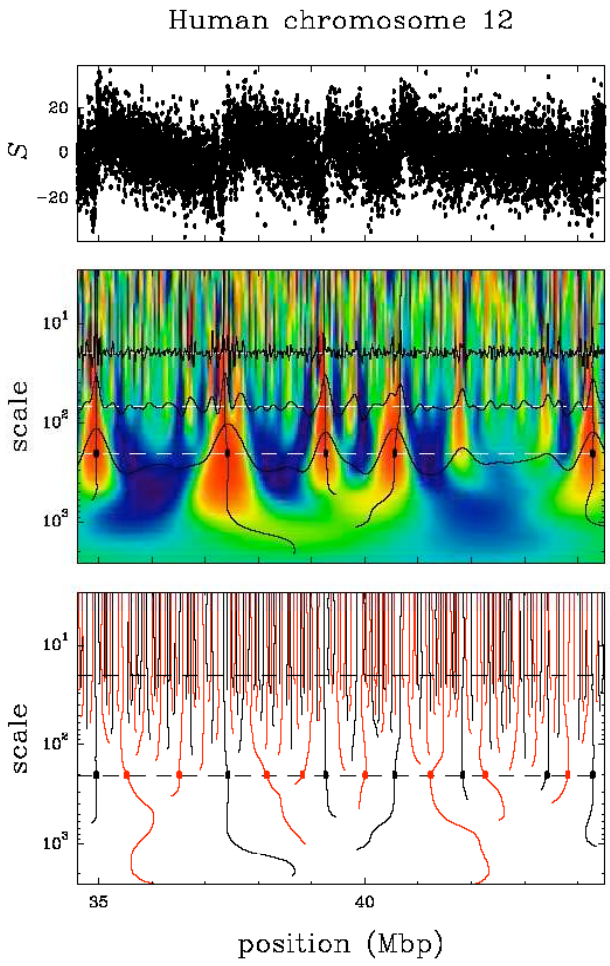

Figure S1: (top) skew profile of a fragment of human chromosome 12; (middle), WT of S using $g^{(1)}$; $T_{g^{(1)}}[S](x,a)$ is coded from black (min) to red (max); three cuts of the WT at constant scale a = a* = 200 kbp, 70 kbp and 20 kbp are superimposed together with five maxima lines identified as pointing to upward jumps in the skew profile; (bottom) WT skeleton defined by the maxima lines in black (resp. red) when corresponding to positive (resp. negative) values of the WT.