Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

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

Antoine Baker [a,b,1], Samuel Nicolay [c,1], Lamia Zaghloul [a,b], Yves d'Aubenton-Carafa [d], Claude Thermes [d], Benjamin Audit [a,b], Alain Arneodo [a,b,*]

[a] *Université de Lyon, F-69007 Lyon, France*
[b] *Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France*
[c] *Institut de Mathématique, Université de Liège, Grande Traverse 12, 4000 Liège, Belgium*
[d] *Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France*

ARTICLE INFO

Article history:
Received 30 June 2009
Accepted 3 October 2009
Available online 13 October 2009
Communicated by Ginette Saracco

Keywords:
Continuous wavelet transform
DNA sequences
Strand compositional asymmetry
Transcription
Replication
Mammalian genomes

ABSTRACT

During genome evolution, the two strands of the DNA double helix are not subjected to the same mutation patterns. This mutation bias is considered as a by-product of replicative and transcriptional activities. In this paper, we develop a wavelet-based methodology to analyze the DNA strand asymmetry profiles with the specific goal to extract the contributions associated with replication and transcription respectively. In a first step, we use an adapted N-shaped analyzing wavelet to perform a multi-scale pattern recognition analysis of the sum of the TA and GC skews along human chromosomes. This method provides an objective segmentation of the human genome in skew domains of $\simeq 1$ Mbp characteristic size, bordered by two putative replication origins recognized as large amplitude upward jumps in the noisy skew profile. In a second step, we use a least-square fitting procedure to disentangle, in these skew domains, the small-scale (the mean human gene size $\simeq 30$ kbp) square-like transcription component from the global N-shaped component induced by replication. When applying this procedure to the 22 human autosomes, we delineate 678 replication domains of mean length $\bar{L} = 1.2 \pm 0.6$ Mbp spanning 33.8% of the human genome and we predict 1062 replication origins. When investigating the distribution of transcription-associated skew inside the replication N-domains, we reveal some dependence upon the distance to the putative replication origins located at N-domain extremities, the closer the genes to the origin, the larger their transcription bias as the signature of a higher transcriptional activity in the germ-line. As a comparative analysis, we further apply our wavelet-based methodology to skew profiles along the mouse chromosomes. The striking similarity of the results in human and mouse indicates that the remarkable gene organization observed inside the human replication N-domains is likely to be a general feature of mammalian genomes.

[*] Corresponding author at: Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France.
E-mail addresses: antoine.baker@ens-lyon.fr (A. Baker), S.Nicolay@ulg.ac.be (S. Nicolay), lamia.zaghloul@ens-lyon.fr (L. Zaghloul), daubenton@cgm.cnrs-gif.fr (Y. d'Aubenton-Carafa), claude.thermes@cgm.cnrs-gif.fr (C. Thermes), benjamin.audit@ens-lyon.fr (B. Audit), alain.arneodo@ens-lyon.fr (A. Arneodo).

[1] These authors contributed equally to this work.

1063-5203/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.acha.2009.10.003

1. Introduction

Since the pioneering works of J. Morlet and A. Grossmann in the early 1980's [62,64,65], the continuous wavelet transform (WT) has been the subject of considerable theoretical developments and practical applications in a wide variety of fields [3,42,44,49,50,80,97,103,104,125,129,134]. Originally introduced to perform time-frequency analysis, the WT has been early recognized as a mathematical microscope that is well adapted to characterize the scale-invariance properties of fractal objects and to reveal the hierarchy that governs the spatial distribution of the singularities of multifractal measures and functions [2,5,19,68,69,75,76,98,99]. This has led A. Arneodo and collaborators [12,29,109–111] to elaborate on a unified statistical (thermodynamic) description of multifractal distributions including measures and functions, the so-called wavelet transform modulus maxima (WTMM) method. This method relies on the computation of partition functions from the WT skeleton defined by the wavelet transform modulus maxima. This skeleton provides an adaptive space-scale partition of the fractal distribution under study, from which one can extract the $D(h)$ singularity spectrum of Hölder exponent values as the equivalent of a thermodynamic potential (entropy) [12]. We refer the reader to Bacry et al. [29], Jaffard [77,78] and collaborators [79] for rigorous mathematical results and to Hentschel [66] for the theoretical treatment of random multifractal functions. Applications of the WTMM method to 1D signals [9] and its generalization in 2D for image analysis [16,18,45] and in 3D for scalar and vector fields analysis [82–84] have already provided insights into a wide variety of problems [9], in domains as different as fully developed turbulence [20,21,46,105,106,122], hydrology [123,142,143], astrophysics [87], geophysics [17,23,121], econophysics [22,112], fractal growth phenomena [4,6,7,88], medical time series analysis [72,73] and medical and biological image processing [16,38,85,86]. Surprisingly, among these applications, there is one that turns out to be quite fruitful and very promising in regards to its unexpected appropriateness, namely the multi-scale wavelet-based analysis of genomic sequences [9,11,15,25,26].

The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences has been the subject of increasing interest [9,93,130]. During the past fifteen years or so, there has been intense discussion about the existence, the nature and origin of the long-range correlations (LRC) observed in DNA sequences [9,15,28,34,41, 67,81,89,91,92,100,113,117,118,130,144–146]. One of the main obstacles to LRC analysis in DNA sequences is the genuine mosaic structure of these sequences that are well known to be formed of patches of different underlying composition [32, 57,94]. When using the DNA walk representation, these patches appear as trends in the DNA walk landscapes that are likely to break scale-invariance [9,34,41,81,89,113,117,130,144]. Indeed, most of the techniques, *e.g.* the variance method, used in the early studies for characterizing the presence of LRC, were not well-adapted to study non-stationary sequences. There have been some phenomenological attempts to differentiate local patchiness from LRC using *ad hoc* methods such as the so-called "min–max method" [117] and the "detrended fluctuations analysis" [119]. In that context the WT has been early recognized as a well-suited technique that overcomes this difficulty [9,11,15]. By considering analyzing wavelets that make the WT microscope blind to low-frequency trends, any bias in the DNA walk can be removed and the existence of power-law correlations with specific scale invariance properties can be revealed accurately. As a first important result, from a systematic WT analysis of human exons, CDSs and introns, LRC were found in non-coding sequences as well as in regions coding for proteins somehow hidden in their inner codon structure [13]. This observation made rather questionable the model based on genome plasticity proposed at that time to account for the reported absence of LRC in coding sequences [11,15,36, 91,117,130]. An alternative structural interpretation of these LRC has emerged from a comparative multifractal analysis of DNA sequences using structural coding tables based on nucleosome positioning data [25,26]. The application of the WTMM method has revealed that the corresponding DNA chain bending profiles are monofractal (homogeneous) as characterized by a unique Hölder exponent $h = H$ and that there exists two LRC regimes. In the 10–200 bp range, LRC are observed for eukaryotic sequences as quantified by a Hurst exponent value $H \simeq 0.6$ as the signature of the nucleosomal structure. In contrast, for eubacterial sequences, the uncorrelated $H = 0.5$ value is systematically obtained. These LRC were further shown to favor the autonomous formation of small (a few hundred bps) 2D DNA loops and in turn the propensity of eukaryotic DNA to interact with histones to form nucleosomes [139,141]. In addition these LRC might induce some local hyper-diffusion of these loops which would be a very attractive interpretation of the nucleosome repositioning dynamics. Over larger distances ($\geqslant 200$ bp), stronger LRC with $H \simeq 0.8$ seem to exist in any sequence [25,26] as experimentally confirmed by atomic force microscopy imaging of naked DNA molecules deposited onto a mica surface under 2D thermodynamic equilibrium conditions [107]. Furthermore these LRC were recently observed in *S. cerevisiae* nucleosome positioning *in vivo* data suggesting that they are involved in the collective nucleosome organization of the so-called 30 nm chromatin fiber [10,140]. The fact that this second regime of LRC is also present in eubacterial sequences shows that it is likely to be a possible key to the understanding of the structure and dynamics of both eukaryotic and prokaryotic chromatin fibers.

The flowering availability of new fully sequenced genomes has provided the unprecedented opportunity to generalize the application of the WTMM method to genome-wide multifractal sequence analysis when using alternative codings that have a clear functional meaning. According to the second parity rule [40,124], under no strand-bias conditions, each genomic DNA strand should present equimolarities of adenines A and thymines T, and of guanines G and cytosines C [51,95]. Deviations from intrastrand equimolarities have been extensively studied during the past decade and the observed TA and GC skews have been attributed to asymmetries intrinsic to the replication and transcription processes. During these processes, an asymmetry can result if mutational events or repair mechanisms affect the two strands differently. The existence of transcription- and/or replication-associated strand asymmetries has been mainly established for prokaryote, organelle and virus genomes [30,54–56,108,120,133]. For a long time the existence of compositional biases in eukaryotic sequences has been unclear and it is only recently that (i) statistical analyses of eukaryotic gene introns have revealed the presence of transcription-coupled strand asymmetries [63,136,137] and (ii) genome-wide multi-scale analysis of mammalian genomes has clearly shown some departure from the intrastrand equimolarities in intergenic regions and further confirmed the existence of replication-associated strand asymmetries [35,115,138]. In particular the application of the WTMM method to the skew $S = S_{TA} + S_{GC}$ in the human genome [8,116] has revealed the bifractal nature of the corresponding DNA walk landscape which involves two competing scale invariant (from repeat masked distances of 1 kbp up to 40 kbp) components characterized by Hölder exponent $h_1 = 0.78$ and $h_2 = 1$ respectively. The former corresponds to the long-range correlated homogeneous fluctuations previously observed in DNA walks generated with structural codings [25,26]. The latest is associated with the presence of jumps in the original strand asymmetry noisy signal S. A majority of the detected upward (resp. downward) jumps were shown to co-locate with gene transcription start sites (TSS) (resp. transcription termination sites (TTS)). Out of the 20 023 TSS annotated in "refGene", 36% (7728) were delineated within 2 kbp by an upward jump of sufficient amplitude ($\Delta S > 0.1$). This provides a very reasonable estimate of the number of genes expressed in germ-line cells as compared to the 32% experimentally found to be bound to PolII in human embryonic cells [90]. Interestingly, about a third of the detected upward jumps are still observed at scale $\geqslant 200$ kbp as bordering large-scale (mean size $\simeq 1$ Mbp) N-shaped skew domains. As some of these large amplitude upward jumps co-locate with experimentally identified replication origins, these domains were further qualified as replication domains [8,35,70,138]. Recent high resolution replication timing data have confirmed these *in silico* predictions: the putative replication origins at the N-domain extremities are replicating earlier than their surrounding whereas the central regions replicate late in the S-phase [24]. Furthermore, these replication N-domains were shown to be at the heart of a remarkable gene organization along human chromosomes [70].

In regards to the very limited knowledge on replication initiation in mammalian genomes (only about 30 replication origins were experimentally identified when we started this study [35,115,138]), our aim here is to use the WT transform to develop a multi-scale methodology to objectively define replication skew domains thereby predicting (at their edges) replication origins directly from the DNA sequence. With an adequate choice of the analyzing wavelet, we show that the proposed method can be further used to disentangle, in the so-identified skew domains, the contribution coming from transcription (local square-like genic skew component) from the one associated with replication (global N-shaped skew component). Efficient algorithms are implemented and tested on synthetic skew profiles prior to genome-wide analysis of the human and mouse genomes. The paper is organized as follows. In Section 2, we review previous studies of compositional strand asymmetries in eukaryotic genomes that define the state of the art in the domain. In Section 3, we elaborate on our wavelet-based method to detect N-shaped domains in a noisy skew profile. The efficiency and reliability of this method are tested on artificial home made noisy skew signals. Section 4 is devoted to the application of this method to a comparative segmentation of the human and mouse genomes. In Section 5, we further develop this method to achieve the ambitious goal of discriminating between part of the skew associated with transcription and part associated with replication. The results obtained in the human and mouse genomes are compared and further discussed as providing a new vision of gene organization in mammalian genomes which integrates transcription, replication and chromatin structure as coordinated determinants of genome architecture. Concluding remarks and perspectives are given in Section 6.

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

2.1. Definitions

Because transcription and replication both require the opening of the DNA double helix, they are likely to induce some departure from intrastrand equimolarities resulting from asymmetry in mutational and repair events on the two strands. As indicators of DNA strand-asymmetry, we will mainly use in this study the TA and GC skews computed in non-overlapping 1 kbp windows [8,14]:

$$S_{TA} = \frac{n_T - n_A}{n_T + n_A}, \qquad S_{GC} = \frac{n_G - n_C}{n_G + n_C}, \tag{1}$$

where n_A, n_C, n_G and n_T are respectively the numbers of A, C, G and T in the windows. Because of the observed correlation between TA and GC skews [8,14], we will mainly consider the total skew:

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

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

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

Asymmetries of substitution rates coupled to transcription have been mainly observed in prokaryotes [30,54,56], with only recent results in eukaryotes. In the human genome, excess of T was early observed in a set of gene introns [48] and

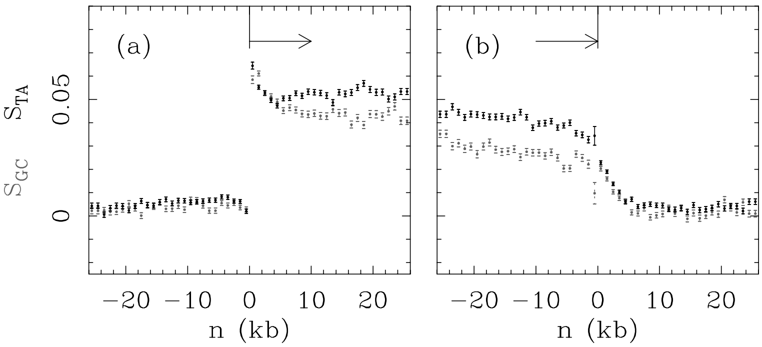

Fig. 1. TA (●) and GC (●) skew profiles in the regions surrounding 5′ and 3′ gene extremities [137]. S_{TA} and S_{GC} were calculated in 1 kbp windows starting from each gene extremities in both directions. In abscissa is reported the distance (n) of each 1 kbp window to the indicated gene extremity; zero values of abscissa correspond to 5′ (a) or 3′ (b) gene extremities. In ordinate is reported the mean value of the skews over 14 854 intron-containing genes for all 1 kbp windows at the corresponding abscissa. Error bars represent the standard error of the means.

some large-scale asymmetry was observed in human sequences but was attributed to replication [128]. More recently, a comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G + T over A + C in the coding strand [63,136,137]. In contrast to the substitution biases observed in bacteria presenting frequently an excess of C → T transitions, these asymmetries are characterized by an excess of purine (A → G) transitions relatively to pyrimidine (T → C) transitions. These might be a by-product of the transcription-associated repair mechanism acting on uncorrelated substitution errors during replication [131]. In a previous study, we have reported definite evidence for nucleotide compositional strand asymmetries in transcribed regions of human sequences [136,137]. First, we have computed the S_{TA} and S_{GC} skews (1) for intronic sequences since, in contrast to exonic sequences, they can be considered as weakly selected sequences. For each gene, we have considered only the intronic sequences without changing their positions relative to the TSS [137]. The distributions of the TA and GC skews, computed on the 14 854 intron-containing genes, present positive mean values for the (+) genes (7058) which correspond to the reference sequence ("+" strand), namely $\bar{S}_{TA} = 4.49 \pm 0.01\%$ and $\bar{S}_{GC} = 3.29 \pm 0.01\%$, and nearly opposed values for the (−) genes (7346) which correspond to the sequence of the other strand ("−" strand). Then we have compared the overall S_{TA} and S_{GC} skew profiles in transcribed regions to those in the neighboring intergenic sequences [136,137]. In Fig. 1 are reported the mean values of S_{TA} and S_{GC} skews for all genes as a function of the distance to the 5′ or 3′ gene ends. At the 5′ gene extremity (Fig. 1(a)), a sharp transition of both skews is observed from close to zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for S_{TA} and between 3 and 5% for S_{GC}. At the 3′ gene extremity (Fig. 1(b)), the TA and GC skews also exhibit a transition from significantly large positive values inside the gene to very small values in untranscribed regions. However, in comparison to the steep transition observed at 5′ end, the 3′ end mean profile presents a slightly smoother transition pattern extending over ∼ 5 kbp and including regions downstream of the 3′ end likely reflecting the fact that transcription continues to some extent downstream of the polyadenylation site. In pluricellular organisms, mutations responsible for the observed biases have occurred in germ-line cells. It could happen that gene 3′ ends annotated in the databank differ from the poly-A sites effectively used in germ-line cells. Such differences would then lead to some broadening of the skew profiles. Overall, the results reported in Fig. 1 suggest that the S_{TA} and S_{GC} are constant along introns. Since introns amount for about 80% of gene sequences, this means that skew profiles induced by the transcription process have a characteristic square-like shape [8,14,136,137]. However, the absence of asymmetries in intergenic regions does not exclude the possibility of additional replication associated biases. Such biases would present opposite signs on leading and lagging strands that would cancel each other in our statistical analysis as a result of the spatial distribution of multiple unknown replication origins.

If there is not doubt that the mean TA and GC skew profiles are different from zero inside the genes likely resulting from transcription-coupled processes, how many genes actually contribute to these mean profiles or in other words, how many genes have biased sequences? Since each square-like skew pattern is edged by one upward and one downward jump, the set of human genes that are significantly biased is expected to contribute to an equal number of $\Delta S > 0$ and $\Delta S < 0$ jumps. This is exactly what we observed when using the WT microscope to detect jumps in the noisy total skew profile S when exploring the range of scales $10 \leqslant a \leqslant 40$ kbp, typical of human gene size [116]. Out of 20 023 TSS, 36% (7228) were found to be delineated within 2 kbp by an upward jump of amplitude $\Delta S > 0.1$. This percentage of biased genes provides a very reasonable estimate of the number of genes expressed in germ-line cells as compared to the 31.9% recently experimentally found to be bound to PolII in human embryonic stem cells [90].

This study of strand asymmetries in intronic sequences has been further extended to evolutionary distant eukaryotes [136]. When appropriately examined, all genomes present transcription-coupled excess of T over A ($S_{TA} > 0$) in the coding strand. In contrast, GC skew is found positive in mammals and plants but negative in invertebrates suggesting different repair mechanisms associated with transcription in vertebrates and invertebrates [135,136].

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

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. According to the so-called “replicon” paradigm derived from prokaryotes [74], this process starts with the binding of some “initiator” protein to a specific “replicator” DNA sequence called *origin of replication*. The recruitment of additional factors initiates the bi-directional progression of two divergent replication forks along the chromosome. One strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand). In eukaryotic cells, this event is initiated at a number of replication origins and propagates until two converging forks collide at a *terminus of replication* [31]. The initiation of different replication origins is coupled to the cell cycle but there is a definite flexibility in the usage of the replication origins at different developmental stages [1,52,58,71, 127]. Also, it can be strongly influenced by the distance and timing of activation of neighboring replication origins, by the transcriptional activity and by the local chromatin structure [1,52,58,127]. Actually, sequence requirements for a replication origin vary significantly between different eukaryotic organisms. In the unicellular eukaryote *S. cerevisiae*, the replication origins spread over 100–150 bp and present some highly conserved motifs [31]. However, among eukaryotes, *S. cerevisiae* seems to be an exception that remains faithful to the replicon model. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp [31]. In multicellular organisms, the nature of initiation sites of DNA replication is even more complex. Metazoan replication origins are rather poorly defined and initiation may occur at multiple sites distributed over a thousand of base pairs [60]. The initiation of replication at random and closely spaced sites was repeatedly observed in *Drosophila* and *Xenopus* early embryo cells, presumably to allow for extremely rapid S phase, suggesting that any DNA sequence can function as a replicator [43,71,126]. A developmental change occurs around midblastula transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites [71,126]. Thus, although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types [33,61]. As recently proposed by many authors [47,101,102], the need to fulfill specific requirements that result from cell diversification may have led multicellular eukaryotes to develop various epigenetic controls over the replication origin selection rather than to conserve specific replication sequence. This might explain that for many years, very few replication origins have been identified in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human when we started this study. Since then, about a few hundred replication origins have been further experimentally identified [39,96,132] on the ENCODE sequences that represent one percent of the human genome only. Along the line of this epigenetic information, one might wonder what can be learned about eukaryotic DNA replication from DNA sequence analysis.

The existence of replication-associated strand asymmetries has been mainly established in bacterial genomes [55,95,108, 120,133]. S_{GC} and S_{TA} skews abruptly switch sign (over few kbp) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This leads to a square-like skew pattern characteristic of the replicon model [74]. However, in eukaryotes, the existence of compositional biases is unclear and most attempts to detect the replication origins from strand compositional asymmetry have been inconclusive. Several studies have failed to show compositional biases related to replication, and analysis of nucleotide substitutions in the region of the *β-globin* replication origin in primates does not support the existence of mutational bias between the leading and the lagging strands [37,53,108]. Other studies have led to rather opposite results. For instance, strand asymmetries associated with replication have been observed in the subtelomeric regions of *S. cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism [59]. In a previous work [35,138], we have investigated the behavior of the (repeat masked) skew profiles around 9 replication origins experimentally identified in the human genome. As shown in Fig. 2(a) for TOP1 replication origin, most of these origins also correspond to rather sharp (over several kbp) transitions from negative to positive S (S_{TA} as well as S_{GC}) skew values that clearly emerge from the noisy background. As shown in Fig. 2(b)–(d), sharp upward jumps of amplitude $\Delta S \geqslant 15\%$, similar to the ones observed for the known replication origins (Fig. 2(a)), seem to exist at many other locations along the human chromosomes. This observation led us to develop an upward jump detection methodology based on the WT microscope [35,138]. By selecting in the WT skeleton, the maxima lines that still exist at scales $a \geqslant 200$ kbp, *i.e.* scales larger than the typical gene size (~ 30 kbp) [35,138], we not only reduce the effect of the noise but we also reduce the contribution of the upward ($5'$ extremity) and downward ($3'$ extremity) jumps associated to the square-like skew pattern induced by transcription (Fig. 1). When applying this wavelet-based method to the 22 human autosomes, retaining as putative replication origins upward jumps with $\Delta S \geqslant 12\%$, *i.e.* with an amplitude much larger than the one induced by transcription at the TSS (Fig. 1(a)), we got a set of 1012 candidates mainly located in regions with $\mathrm{G+C} \leqslant 42\%$ (as seen in Fig. 2(d), in $\mathrm{G+C}$ rich regions the required scale separation between the characteristic replicon and gene sizes is no longer tractable to our multi-scale methodology) [35,138].

But when examining the behavior of the skews at large distance from the replication origins, one does not observe a square-like pattern with upward and downward jumps at the origin and termination positions as expected from the replicon model. Indeed the most striking feature is the fact that in between two neighboring major upward jumps, not only the noisy S profile does not present any comparable downward sharp transition, but it displays a remarkable decreasing linear behavior. At chromosome scale, we thus get jagged S profiles that have the aspect of “factory roofs” [14,35,70,138]. Note that the jagged S profiles shown in Fig. 2(a)–(d) look somehow disordered because of the extreme variability in the distance between two successive upward jumps, from spacing $\sim$ 50–100 kbp ($\sim$ 100–200 kbp for the native sequences)

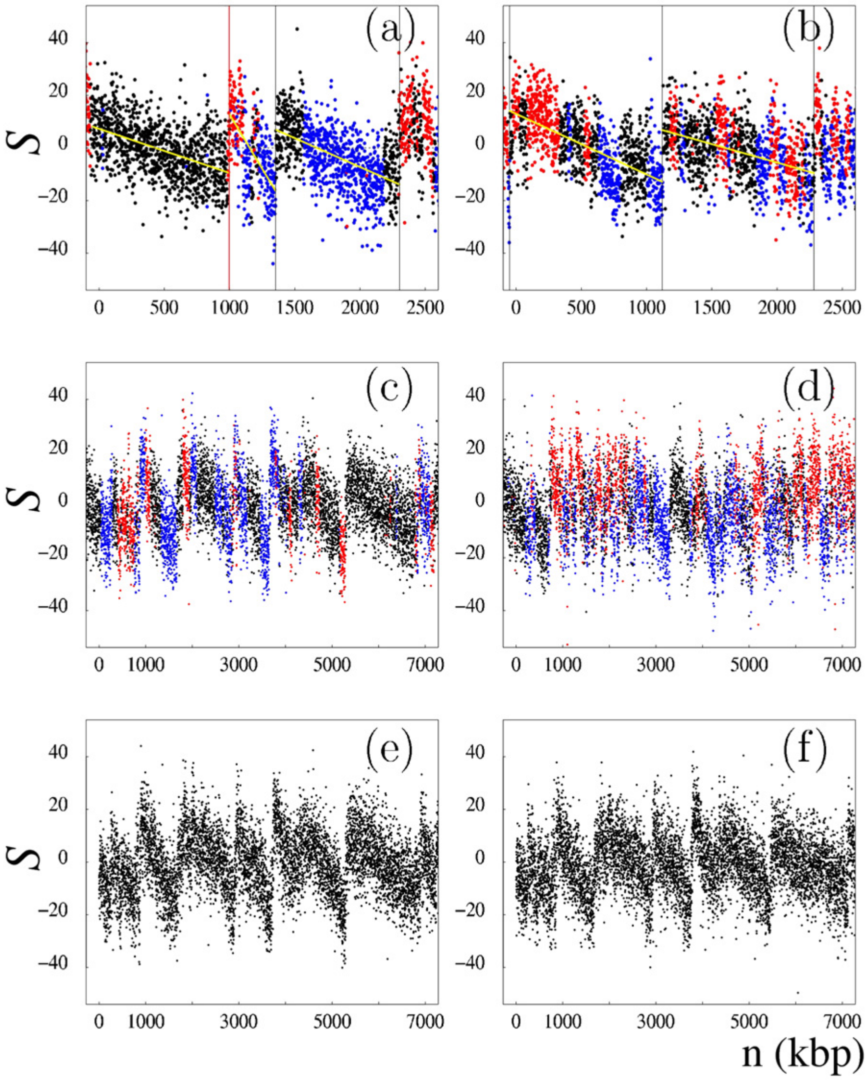

Fig. 2. *S* profiles along mammalian genome fragments (without repeats) [14,138]. (a) Fragment of human chromosome 20 including the TOP1 origin (red vertical line). (b and c) Human chromosome 4 and chromosome 9 fragments, respectively, with low GC content (36%). (d) Human chromosome 22 fragment with larger GC content (48%). In (a) and (b), vertical lines correspond to selected putative origins; yellow lines are linear fits of the *S* values between successive putative origins. Black, intergenic regions; red, (+) genes; blue, (−) genes. Note the fully intergenic regions upstream of TOP1 in (a) and from positions 5290–6850 kbp in (c). (e) Fragment of mouse chromosome 4 homologous to the human fragment shown in (c). (f) Fragment of dog chromosome 5 homologous to the human fragment shown in (c). In (e) and (f), genes are not represented. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mainly in GC rich regions (Fig. 2(d)), up to 1–2 Mbp (∼ 2–3 Mbp for native sequences) (Fig. 2(c)). But what is important to notice is that some of these segments between two successive skew upward jumps are entirely intergenic (Fig. 2(a, c)), clearly suggesting that the observed peculiar N-shape skew profile is characteristic of a strand bias resulting solely from replication [35,70,138]. Importantly, as illustrated in Fig. 2(e, f), the factory-roof pattern is not specific to human sequences but is also conserved in homologous regions of the mouse and dog genomes [138]. Hence, the presence of strand asymmetry in regions that have strongly diverged during evolution further supports the existence of compositional bias associated with replication in mammalian germ-line cells [14,35,70,138].

2.4. A working model of mammalian “factory roof” skew profiles

According to the results reported just above, we will use as a working model that the overall factory roof profile observed for mammalian genomes actually results from the superposition of two patterns (Fig. 3) [70]. One decreases steadily from the 5′ to the 3′ direction and would be attributable to replication in germ-line cells (Fig. 3(a)). To explain this replication-associated N-shaped pattern, we favor a model in which replication first initiates at origins located at the borders of the N-shaped skew profile domain, followed by successive activations of secondary origins as replication progresses toward

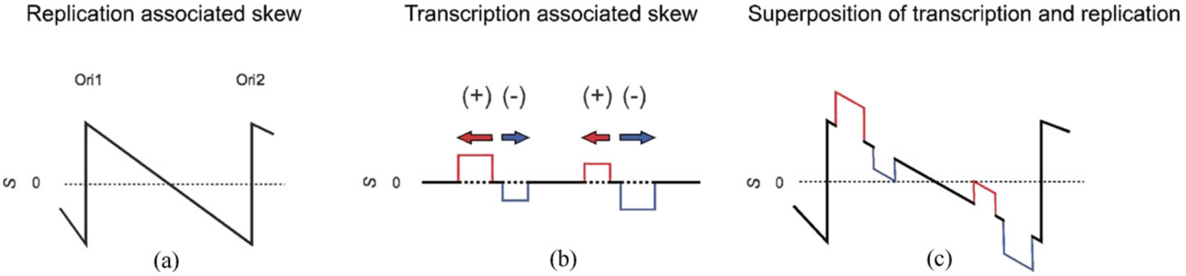
Replication associated skew
Ori1
Ori2
S 0
(a)
Transcription associated skew
(+) (-) (+) (-)
S 0
(b)
Superposition of transcription and replication
S 0
(c)

Fig. 3. (a) N-shaped replication-associated skew profile. (b) Transcription-associated skew profile showing positive square blocks at (+) gene positions and negative square blocks at (−) gene positions. (c) Superimposition of the replication- and transcription-associated skew profiles producing the final factory-roof pattern that defines "N-domains".

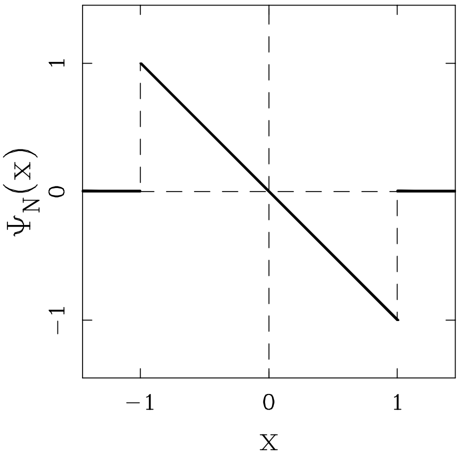

Fig. 4. N-shaped analyzing wavelet $\Psi_N(x)$ defined in (4).

the center of this domain. The linear decline of the skew would reflect a progressive change in the proportion of center- and border-oriented forks that itself reveals the dynamic pattern with which secondary initiations would occur within the domain. The other pattern would result from transcription-associated strand asymmetries that generate square-like profiles corresponding to (+) and (−) genes (Fig. 3(b)). When the two profiles are superimposed, this leads to the factory roof pattern (Fig. 3(c)) [70]. Because the typical gene size ($\sim$ 30 kbp) is much smaller than the characteristic distance between two adjacent putative replication origins, we will define these replication domains as "N-domains" [70] in regards of their overall qualitative N-shape.

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

3.1. The continuous N-let transform

The continuous wavelet transform (WT) is a space-scale analysis which consists in expanding a signal S in terms of wavelets that are constructed from a single function, the analyzing wavelet Ψ, by means of dilations and translations [42,44,62,64,65,103]:

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

where b and a (> 0) are the space and scale parameters respectively. The analyzing wavelet Ψ is generally chosen to be well localized in both space and frequency. Usually Ψ is required to be of zero mean for the WT to be invertible. For the particular purpose of segmenting the human genome, and more generally mammalian genomes, according to the working model described in Fig. 3, we will use in the following an adapted analyzing wavelet called N-let because of its shape that looks like the letter N (Fig. 4) [8,24,70]:

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

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

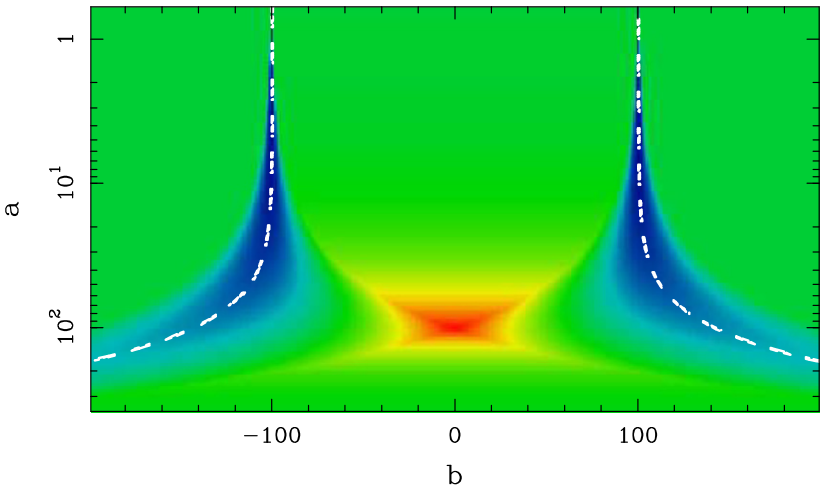

Fig. 5. Space-scale representation of the N-let transform of the skew function $S(x) = -10^{-2}x\chi_{[-100,100]}(x)$ (see (5)). $W_{\Psi_N}[S](b,a)$ is maximum at $(b^*, a^*) = (0, 100)$. $W_{\Psi_N}[S](b,a)$ is color coded from blue (minimum) to red (maximum) through green (intermediate values). The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

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

Along the line of the working model of mammalian "factory roof" skew profiles defined in Fig. 3, let us compute the N-let transform of the following theoretical skew profile [114]:

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

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

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

$$\text{for } \begin{cases} b \leqslant r_1 - a \text{ or } b \geqslant r_2 + a, \\ r_1 - a < b \leqslant r_1 + a \text{ and } b < r_2 - a, \\ b \geqslant r_1 + a \text{ and } b \leqslant r_2 - a, \\ b > r_1 + a \text{ and } r_2 - a < b < r_2 + a, \\ b < r_1 + a \text{ and } b > r_2 - a. \end{cases} \tag{6}$$

As an illustration, a space-scale representation of the N-let transform of $S(x)$ for the following parameters values $\theta = 10^{-2}$, $h = 0$, $r_1 = -100$ and $r_2 = 100$, is shown in Fig. 5. Some 1D cuts corresponding to different scale values $a \leqslant a^*$ are shown in Fig. 6. For a given scale a, $W_{\Psi_N}[S](b,a)$ exhibits a plateau for $b \in [b_1^*(a), b_2^*(a)]$, where $b_1^*(a) = r_1 + a$ and $b_2^*(a) = r_2 - a$. When increasing a, this plateau increases linearly with a while the interval $[b_1^*(a), b_2^*(a)]$ shrinks to zero so that $W_{\Psi_N}[S](b,a)$ presents a local maximum in the (b,a) half-plane at the point (b^*, a^*):

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

where

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

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

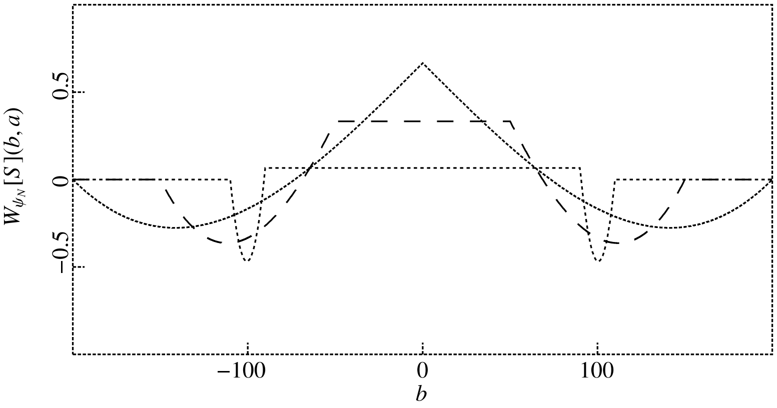

Fig. 6. Horizontal 1D cuts of the N-let transform represented in Fig. 5 for the scales $a = 10$ (dotted line), 50 (dashed line) and 100 (solid line).

3.3. Numerical method

The method we propose to identify replication N-domains in noisy factory roof like profiles (Fig. 2) involves several steps.

- Step 1: Detecting potential N-domain borders.
  We smooth the skew S with a square-like filter of size 20 kbp:
  $$\tilde{S}(x) = \frac{1}{20} S * \chi_{[-10,10]}(x). \tag{9}$$
  The amplitude $\Delta S(x)$ at a point x is defined as:
  $$\Delta S(x) = \tilde{S}(x+20) - \tilde{S}(x-20). \tag{10}$$
  On purpose, we do not take into account the nucleotides that are closer than 10 kbp to the jump location for which a high variability of skew values is observed. Then, along the line of our previous investigation of replication origins in mammalian genomes [8,24,70], the position x_n of an upward jump will be considered as a good candidate for the location of a putative replication origin if it is a local maximum of $\Delta S(x)$ and satisfies the condition:
  $$\tilde{S}(x_n - 20) \leqslant -\epsilon \quad \text{and} \quad \tilde{S}(x_n + 20) \geqslant \epsilon. \tag{11}$$
  This condition not only fixes a threshold ($= 2\epsilon$) in the jump amplitude but it also imposes the fact that a putative replication origin must correspond to a jump from negative to positive skew values. In the present work, according to the histograms of S values obtained from the human and mouse genomes, we will fix the threshold parameter ϵ to:
  $$\epsilon = 3\,10^{-2}. \tag{12}$$
  In this way, for each human and mouse chromosomes, we obtain a dictionary of upward jump locations as potential candidates for N-domain borders.
- Step 2: Associating pairs of borders to define N-domains.
  For each pair of selected upward jumps (x_1, x_2), we determine the position (b_M, a_M) of the local maximum of the N-let transform (*i.e.* red spot in Fig. 5). For each $b \in [x_1, x_2]$, we estimate the range of scales $[a_1(b), a_2(b)]$ over which $W_{\psi_N}[S](b,a)$ behaves linearly with the scale a as predicted by (6). We impose $a_1(b)$ to be larger than 40 (kbp) to minimize bias induced by the noise and the N-let sampling and, using a classical least square fit procedure, we estimate $a_2(b)$. As illustrated in Fig. 7, for the theoretical example (5), in the absence of noise, $a_2(b)$ corresponds to the scale $a_M(b)$ for which $W_{\psi_N}[S](b,a)$ starts decreasing. For the genomic noisy skew profile, the sharp maximum of $W_{\psi_N}[S](b,a)$ *vs* a turns out to be smoothed out so that $a_2(b) \leqslant a_M(b)$. We estimate $a_M(b)$ as the first N-let transform maximum above $a_2(b)$, provided $a_M(b) - a_2(b) \leqslant a_2(b)/10$ (if not we set $a_M(b) = a_2(b)$). Then, b_M is the position corresponding to the maximal value of $a_M(b)$ for $b \in [x_1, x_2]$. Finally, we check the consistency of associating (x_1, x_2) into a N-domain by comparing (b_M, a_M) with (b^*, a^*) (8). The interval $[x_1, x_2]$ is considered as a N-domain candidate if (b_M, a_M) corresponds to the expectation (b^*, a^*) (8) within 10% accuracy and allowing for a maximum of 30 kbp error on domain length. It is retained only if the χ^2 obtained by a linear regression fit of S over $[x_1, x_2]$ is smaller than a critical threshold. We select between overlapping intervals according to their χ^2. For example, if the intervals $I_1 = [x_1, x_2]$, $I_2 = [x_2, x_3]$ and $I_3 = [x_1, x_3]$ are all three good candidates, we will retain either I_1 and I_2 or I_3 according to whether $\chi^2_{I_1} + \chi^2_{I_2} < \chi^2_{I_3}$ or the opposite.
- Step 3: Refining N-domain border locations.
  The dictionary of upward jumps generated in Step 1 contains jumps identified at rather small scales ($\leqslant 20$ kbp) as compared to the mean replication N-domain size we want to detect (see Section 4) and also to the characteristic size

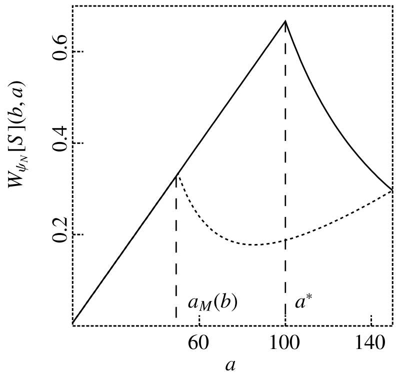

Fig. 7. Vertical 1D cuts of the N-let transform represented in Fig. 5 for $b = 50$ (dotted line) and $b = b^* = 0$ (solid line). The largest value of $a_M(b)$ is obtained for $b = b^*$, $a = a^*$ (8), *i.e.* when the analyzing N-let is positioned at the center of the support of S.

Table 1
Replication N-domains detected in the human and mouse genomes.

	Number	Mean length (Mbp) ± std. dev.	Genome coverage (%)	Mean number of genes	Intergenic coverage (%)	GC (%)
Human	678	0.63 ± 0.33 (masked) 1.19 ± 0.62 (native)	33.8	4.93	57.3 (masked)	40.8
Mouse	587	0.54 ± 0.34 (masked) 0.91 ± 0.62 (native)	22.3	4.13	58.2 (masked)	42.4

of mammalian genes that were shown to induce transcription-associated upward jumps in the skew profile at their promoter (Section 2.2). Consistently with the strategy of detecting putative replication origins pioneered in our previous work [35,138], we take advantage of the space-scale representation provided by the WT to follow from large scales $a \leqslant a_M(\frac{x_1+x_2}{2})$ to small scales, the blue tongues like the ones shown in Fig. 5 that are likely to point to the jump positions at the N-domain extremities. Practically, we identify in the WT skeleton the two nearest WTMM lines that exist at scale $a_M(\frac{x_1+x_2}{2})$ immediately on the left $\mathcal{L}_L$ and the right $\mathcal{L}_R$ of the central point $\frac{(x_1+x_2)}{2}$, and that correspond to the two local negative minima of $W_{\psi_N}[S](b,a)$ when represented versus b in Fig. 6. In regards to the noise amplitude and the mean gene size, we reallocate the N-domain extremities x_1 and x_2, to b_1 and b_2 respectively, where

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

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

3.4. Test application on synthetic skew signals

To test our methodology, we generate synthetic factory roof like profiles of the following simple form [114]:

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

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

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

where $r_j = \sum_{i=1}^{j-1} \rho_i$ and $g(x)$ is a centered white noise. We choose the parameters to get a numerical S profile similar to the ones observed in the human and mouse genomes (see Section 4). We fix $\theta_j \rho_j = 0.14$, $h_j = 0$ an the noise standard deviation $\sigma_g = 0.08$, consistently with the corresponding genomic mean values (see Sections 4 and 5). We generate the length ρ_j of the synthetic N-domains according to a normal law of mean $\bar{\rho} = 550$ kbp and standard deviation $\sigma_\rho = 300$ kbp consistently with the size statistics of the human and mouse masked N-domains (see Table 1). Square-like skew profiles of mean length 30 kbp and amplitude $\Delta S = 0.06$ (resp. $\Delta S = -0.06$) for sense (resp. anti-sense) genes are finally added to mimic the contribution to the skew associated to transcription (Section 2.2).

In Fig. 8 is shown the space-scale representation of a portion of the generated synthetic skew signal provided by the N-let transform. This representation illustrates the essence of our methodology, namely a multi-scale pattern recognition in the N-let transform half-plane. As reported in Fig. 9, out of the generated 2201 N-domains, 1997 are identified by our methodology, *i.e.* more than 90%. When examining the N-domains that were missed, they mainly correspond to small domains of length $\rho \leqslant 200$ kbp for which our method fails because of the lack of scale separation between the three

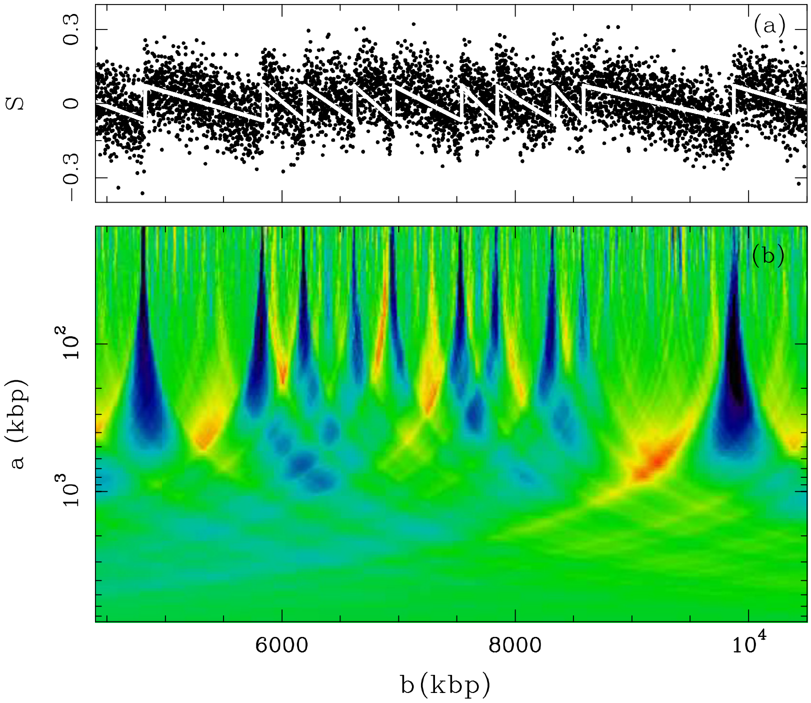

Fig. 8. (a) Synthetic skew profile (14) generated as explained in the text; the gray line represents the succession of deterministic N-shape functions $S_j(x)$. (b) Space-scale representation of the N-let transform of $S(x)$ using the same color coding as in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

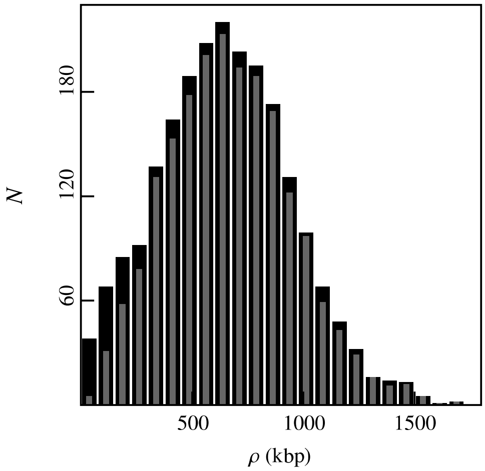

Fig. 9. Histogram of skew N-domain length generated as explained in the text and illustrated in Fig. 8(a): (black) theoretical histogram corresponding to a synthetic skew signal containing 2201 N-domains; (gray) histogram obtained from the 1997 N-domains detected by our N-let based methodology.

components contributing to the total skew, namely the replication- and transcription-associated skews and the noise. When further analyzing the 1997 detected N-domains, we realize that the extremities of these domains are predicted with an accuracy of ∼ 15 kbp which is reasonable in regards to the noise amplitude. As experimented in our pioneering study [35, 138], a better accuracy in upward jump detection can be obtained if one uses a smoother analyzing wavelet than the N-let like the first derivative of the Gaussian function. This leads us to modify Step 3 in our method.

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

When reproducing the test application with this new Step 3, similar efficiency is obtained but with a better accuracy in determining the N-domains edges, the mean error being reduced to 5 kbp [114].

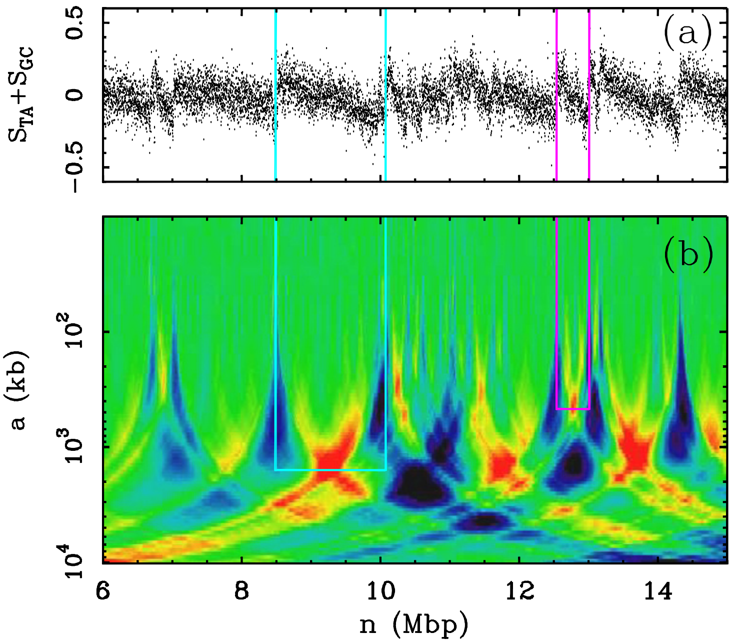

Fig. 10. (a) Skew profile S of a 9 Mbp repeat-masked fragment of human chromosome 21. (b) N-let transform of S using Ψ_N (Fig. 4); $W_{\Psi_N}[S](n,a)$ is color-coded from dark-blue (min; negative values) to red (max; positive values) through green (null values). Light blue and purple lines illustrate the detection of two replication domains of significantly different sizes. Note than in (b), blue cone-shape areas signing upward jumps point at small scale (top) towards the putative replication origins and that the vertical positions of the WT maxima (red areas) corresponding to the two indicated replications domains match the distance between the putative replication origins (1.6 Mbp and 470 kbp respectively). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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

4.1. Human autosomes

When applying the wavelet-based method described in Section 3 to the skew profiles along the 22 human autosomes, we detect 759 N-domain candidates (Fig. 10). Among these domains, we discard 17 domains that contain stretches of N (unknown nucleotides) longer than 100 kbp. We also remove from the remaining N-domains those (64) of length $L >$ 2.8 Mbp whose shape is reminiscent of an N but split in half, leaving in the center a region of null skew whose length increases with domain size. As recently discovered [147], these split-N-domains have a central region corresponding to large heterochromatic gene deserts of homogeneous composition, *i.e.* both a null skew and a constant and low GC content. We end with a library of 678 N-domains bordered by 1062 putative replication origins spanning 33.8% of the genome (Table 1). As shown in Fig. 11, the size of these N-domains ranges from ~ 200 kbp (resp. 100 kbp when masked) to 2.8 Mbp (resp. 1.6 Mbp when masked) with a mean length $\bar{L} = 1.19$ Mbp (resp. 0.63 Mbp when masked). Most of the 1062 putative replication origins at the extremities of the detected replication domains are intergenic (78%) and are located near to a gene promoter more often than would be expected by chance (data not shown). These N-domains contain approximately equal numbers of genes oriented in each direction (1653 (+) genes and 1690 (−) genes) with a mean gene number per domain of 4.93. As observed in [70], gene distributions in the 5′ half of N-domains contain more (+) than (−) genes, and vice-versa for the 3′ half of N-domains. Note that these N-domains have a high intergenic coverage where the skew S is likely to result from replication only. As reported in Table 1 and Fig. 12, most of the detected N-domains are mainly intergenic with a mean (masked) coverage of 57.3%. Indeed only a few N-domains (64/678) have a (masked) intergenic coverage less than 20% (Fig. 12(b)).

4.2. Mouse autosomes

When reproducing the same wavelet-based analysis for the 19 mouse autosomes, 634 N-domains candidates are identified with no domain containing large stretches of unsequenced nucleotides (N). After discarding 47 detected N-domains of length $L > 2.8$ Mbp, we end with a library of 587 N-domains that cover 22.3% of the genome, *i.e.* a percentage slightly smaller than previously obtained for the human genome. This results from the fact that more small domains are detected in the mouse genome (Fig. 10) with a mean native (resp. masked) length of 0.91 Mbp (resp. 0.54 Mbp). Consistently with previous observations that replicon sizes are smaller in GC rich regions (Fig. 2(d)) [8,14,70], the mean GC content of the mouse N-domains (42.4%) is slightly higher than in the human N-domains (40.8%). Despite these slight quantitative differences, most of the features concerning gene organization in human N-domains are also observed in mouse N-domains with

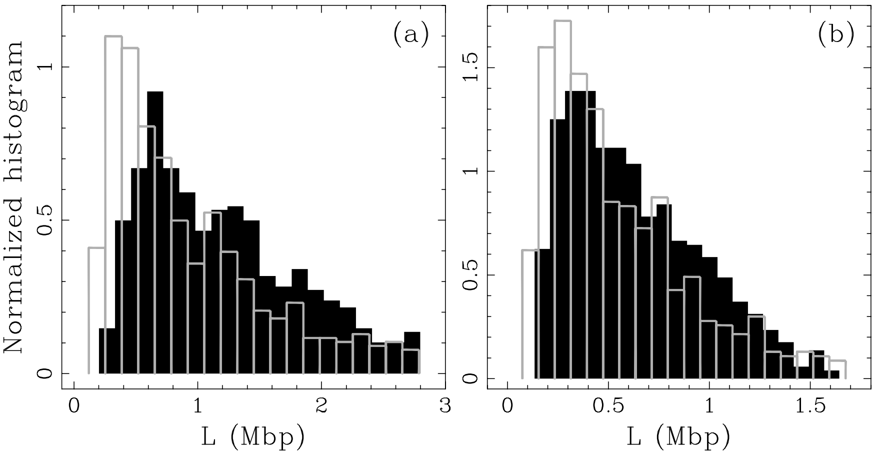

Fig. 11. Normalized histograms of N-domain length detected in the human (black) and mouse (gray) genomes: (a) native sequences; (b) masked sequences.

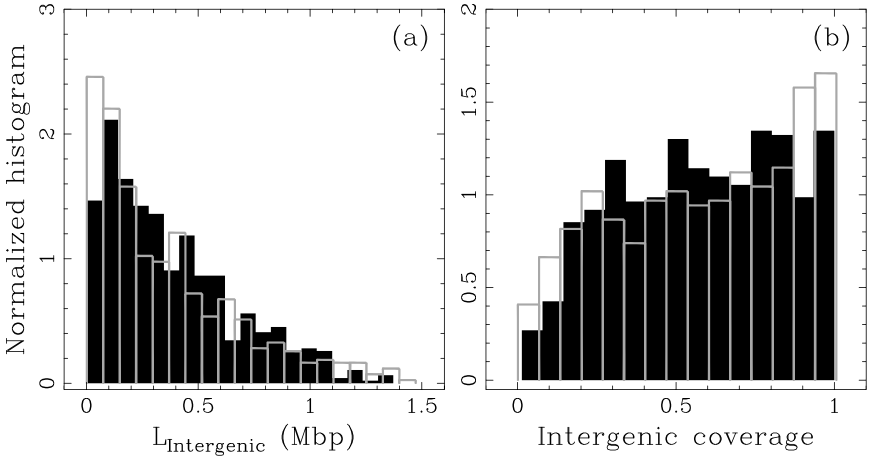

Fig. 12. Normalized histograms of intergenic regions in N-domains detected in human (black) and mouse (gray) genomes: (a) masked intergenic length; (b) coverage by masked intergenic regions per N-domain.

a mean number of genes of 4.13 and globally a similar number of genes oriented in each direction (1220 (+) genes and 1204 (−) genes). Like in the human autosomes a majority of the 994 putative replication origins that border the 587 mouse N-domains are intergenic (71%). Importantly the relative coverage of these N-domains by intergenic regions is important (58.2%) and statistically very similar to what is observed with the human autosomes (Fig. 12).

5. Disentangling transcription- and replication-associated strand asymmetries

5.1. Method

Our method to disentangle transcription and replication skews is based on the working model shown in Fig. 3. When superimposing the N-shaped replication profile and the transcription square-like skew profiles, we get the following theoretical skew profile in a replication N-domain:

$$S(x') = S_R(x') + S_T(x') = -2\delta\left(x' - \frac{1}{2}\right) + h + \sum_{\text{genes}} c_g \chi_g(x'), \tag{16}$$

where position x' within the domain has been rescaled between 0 and 1, h and δ (> 0) are parameters that define the replication bias ($S_R^{5'} = h + \delta$ at the 5′ N-domain extremity and $S_R^{3'} = h - \delta$ at the 3′ extremity), χ_g is the characteristic function for the gth gene that belongs to the N-domain (1 when the points is in the gene and 0 elsewhere) and c_g is its transcriptional bias calculated on the (+) strand (likely to be positive for (+) genes and negative for (−) genes). For each N-domain detected as explained in Section 3 (see Fig. 10), we use a general least-square fit procedure to estimate, from the noisy S profile the parameters δ, h and each of the c_g's. The resulting χ^2 value is then used to select the N-domain candidates where the skew is well described by (16). As illustrated in Fig. 13 for a fragment of human chromosome 6 that

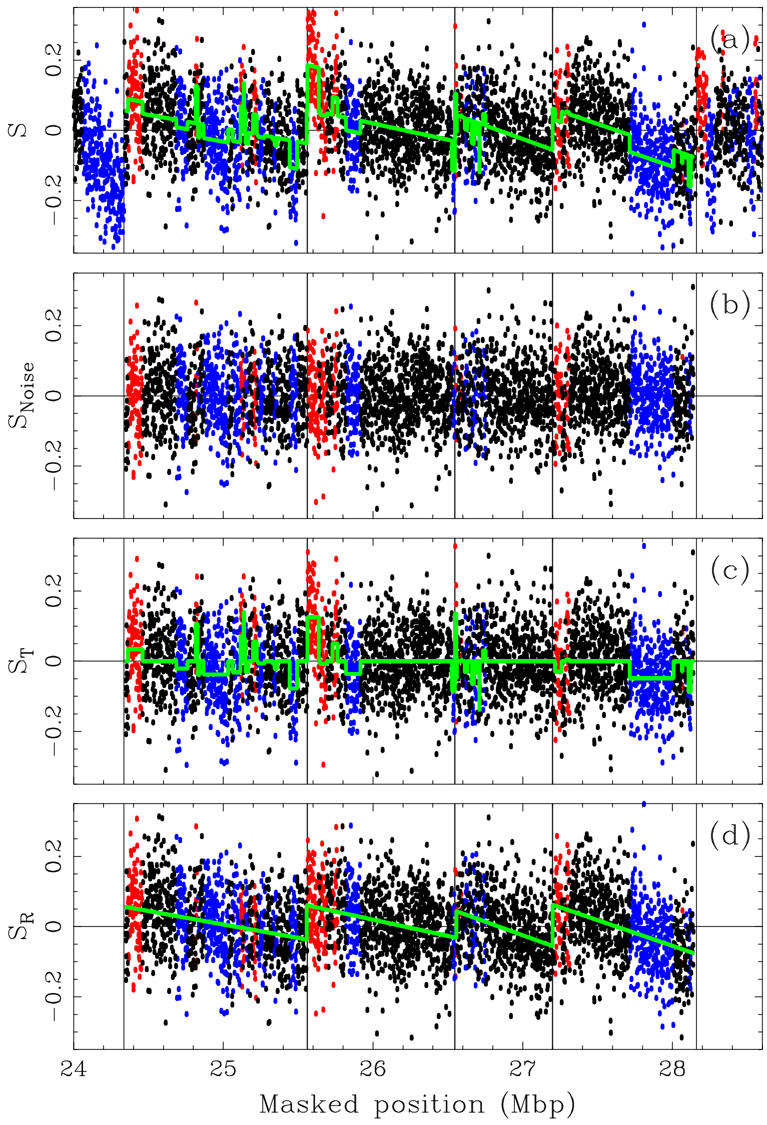

Fig. 13. (a) Skew profile S of a 4.3 Mbp repeat-masked fragment of human chromosome 6; each point corresponds to a 1 kbp window: red, (+) genes; blue, (−) genes; black, intergenic regions (the color is defined by majority rule); the estimated skew profile (16) is shown in green; vertical lines corresponds to the locations of 5 putative replication origins that delimit 4 adjacent domains identified by the wavelet-based methodology. (b) Noise component S_{Noise} obtained by subtracting the estimated total skew (green line in (a)) from the original skew profile in (a). (c) Transcription-associated skew S_{T} obtained by subtracting the estimated replication-associated profile (green lines in (d)) from the original S profile in (a); the estimated transcription step-like profile (third term of the *rhs* of (16)) is shown in green. (d) Replication-associated skew S_R obtained by subtracting the estimated transcription step-like profile (green lines in (c)) from the original S profile in (a); the estimated replication serrated profile (first two terms on the *rhs* of (16)) is shown in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

contains 4 adjacent replication domains (Fig. 13(a)), this method provides a very efficient way to disentangle the square-like transcription skew component (Fig. 13(c)) from the N-shaped component induced by replication (Fig. 13(d)).

Remark. In the least-square fit procedure, we fix the variance σ^2 of the Gaussian noise to the variance $\sigma^2 = \frac{1}{2}\sigma^2_{\delta S}$ computed in each N-domain from the probability distribution function of the skew increments $\delta S(n)/\sqrt{2} = [S(n+1) - S(n)]/\sqrt{2}$. As quantitatively verified *a posteriori* (Fig. 14), this variance directly estimated from the total skew S (Fig. 13(a)) is a very good approximation of the noise component of the skew once subtracted our model skew profile (Fig. 13(b)).

5.2. Human autosomes

Among the 678 N-domains detected in the 22 human autosomes, our disentangling methodology fails to provide satisfactory results (prohibitive too large χ^2 values) for 14 domains only. We have checked that the main reason for which

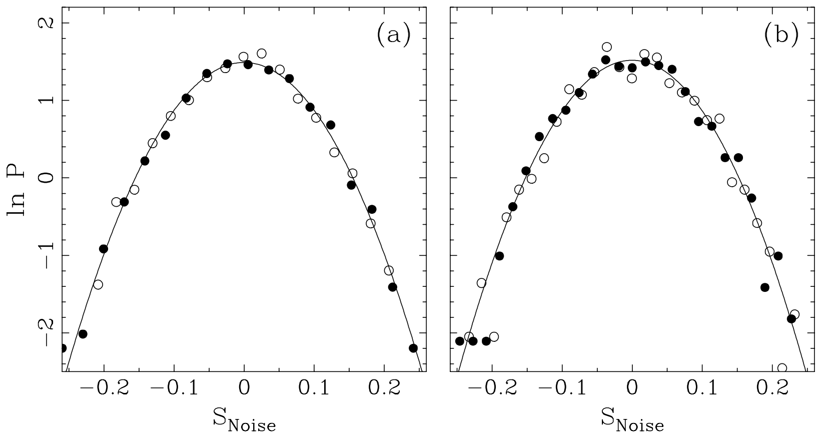

Fig. 14. Semi-log representation of normalized histograms of skew values in: (a) an N-domain of length $L = 2.6$ Mbp (1.5 Mbp masked) in the human genome and (b) an N-domain of length $L = 2.3$ Mbp (1.3 Mbp masked) in the mouse genome. The symbols correspond to the histogram of the skew component S_{Noise} (●) and of the total skew increments $\delta S/\sqrt{2}$ (○). The continuous parabola corresponds to Gaussian distributions of standard deviation $\sigma = 0.090$ (a) and 0.087 (b).

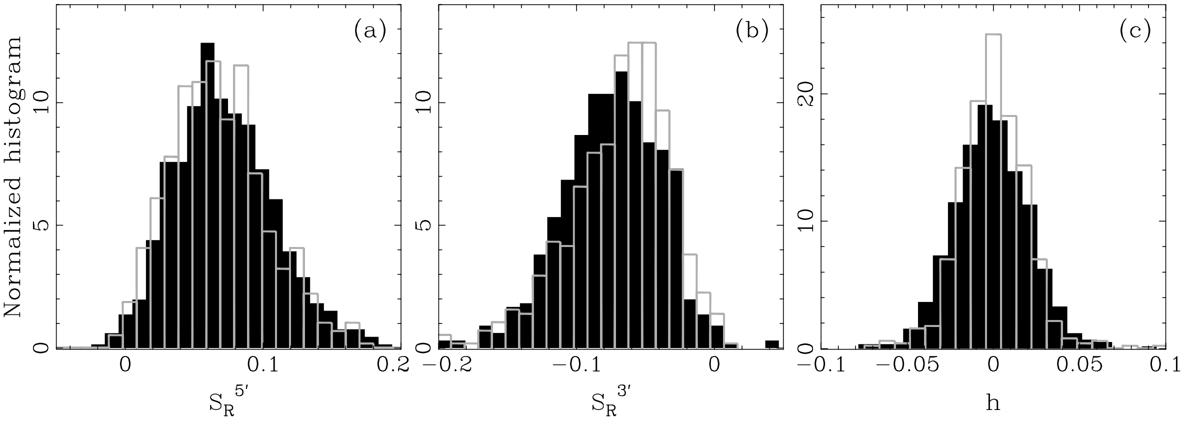

Fig. 15. Normalized histograms of replication parameters $S_R^{5'}$ (a), $S_R^{3'}$ (b) and h (c) estimated in 664 N-domains identified in the 22 human autosomes (black) and in 585 N-domains detected in the 19 mouse autosomes (gray) (see Table 2).

Table 2
Mean replication parameters (in percent ± SEM) computed with our wavelet-based disentangling method from human and mouse skew profiles (see (16)).

	Number of N-domains	$\bar{S}_R^{5'}$	$\bar{S}_R^{3'}$	$\bar{h}$
Human	664	7.2 ± 0.1	-7.4 ± 0.1	$(-9.5 \pm 8.4)\,10^{-2}$
Mouse	585	6.8 ± 0.2	-6.8 ± 0.2	$(-1.8 \pm 7.9)\,10^{-2}$

our working hypothesis (16) does not apply is the fact that some regions present anomalous high amplitude skew values. Hence, in the following, statistical analysis is performed on 664 N-domains.

5.2.1. Replication bias

In Fig. 15 are reported the results of our estimate of the replication parameters $S_R^{5'} = h + \delta$ (Fig. 15(a)), $S_R^{3'} = h - \delta$ (Fig. 15(b)), and h (Fig. 15(c)). The normalized histogram of the offset parameter h (vertical shift of the N profile) is symmetric (Fig. 15(c)), with a mean value $\bar{h} = (-9.5 \pm 8.4)\,10^{-4}$ (Table 2), that cannot be distinguished from zero. This means that the replication N-shaped profile is mainly observed centered at zero with equal statistical probability of upward and downward vertical shift by a few percent. The histograms of replication bias at the 5′ (Fig. 15(a)) and 3′ (Fig. 15(a)) N-domain extremities are quite symmetric one from each other with mean values $\bar{S}_R^{5'} = 7.2 \pm 0.1\%$ and $\bar{S}_R^{3'} = -7.4 \pm 0.1\%$, as expected from an antisymmetric N-shaped skew pattern of zero mean. Altogether these results provide some estimate of the mean replication bias $\bar{\delta} = 7.3 \pm 0.1\%$, which corresponds to an upward jump of mean amplitude $\simeq 14.6\%$ in the skew profile at the putative replication origins that border the replication N-domains. These estimations are consistent with those obtained in our pioneering study of large amplitude upward jumps in human skew profiles [35,138].

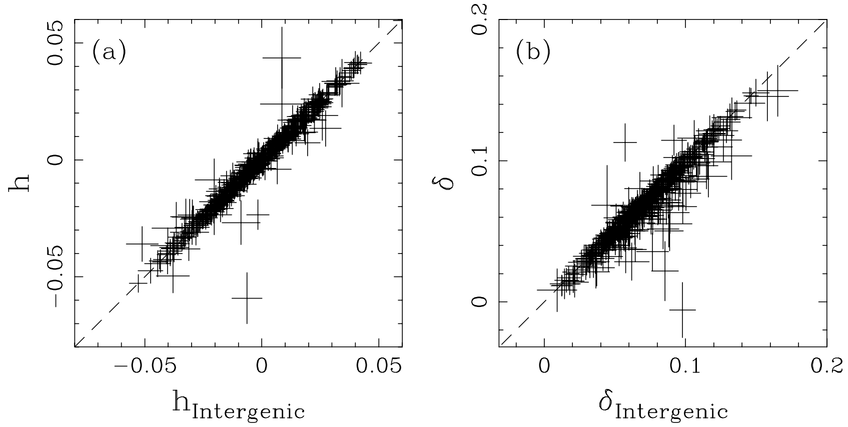

Fig. 16. Test of the consistency of our disentangling methodology: for each N-domain detected in the 22 human autosomes, the replication parameters computed directly from the intergenic skew only are plotted versus the corresponding parameters derived from the method described in Section 5.1. (a) h *vs* $h_{\text{intergenic}}$; (b) δ *vs* $\delta_{\text{intergenic}}$. For clarity, only (437/664) N-domains containing more than 200 kbp of intergenic masked sequences are represented such that the error bars remain small enough.

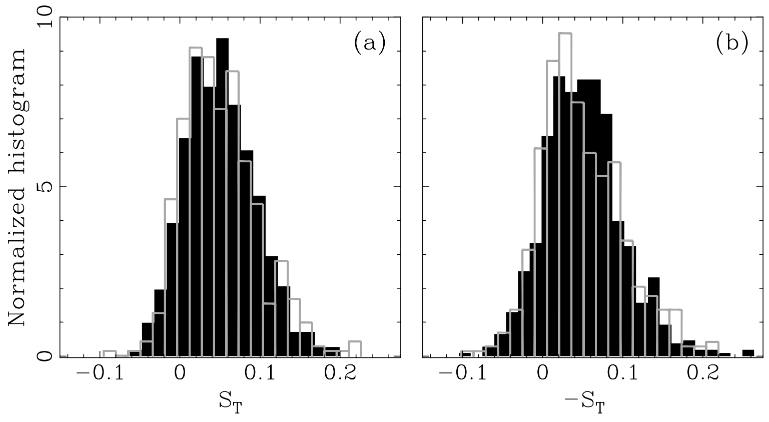

Fig. 17. Normalized histograms of transcription bias S_{T} computed as explained in the main text for human (black) and mouse (gray) genes of length $\geqslant 20$ kbp. (a) (+) Genes; (b) (−) genes (see Table 3).

Table 3

Mean transcription skews (in percent ± SEM) computed with our wavelet-based disentangling method from human and mouse skew profiles (see (16)). $\bar{S}_{\text{T}}^{(+)}$, $\bar{S}_{\text{T}}^{(-)}$, $\bar{S}_{\text{T}}^{R+}$, and $\bar{S}_{\text{T}}^{R-}$ are the mean transcription skews computed for (+), (−), $R+$, and $R-$ genes of length $\geqslant 20$ kbp. For $R+$ and $R-$ genes, the transcription skew is computed on the gene strand, and therefore expected to be positive (see Fig. 1).

	$\bar{S}_{\text{T}}^{(+)}$ $(n^{(+)})$	$\bar{S}_{\text{T}}^{(-)}$ $(n^{(-)})$	$\bar{S}_{\text{T}}^{R+}$ (n^{R+})	$\bar{S}_{\text{T}}^{R-}$ (n^{R-})
Human	5.1 ± 0.2 (726)	-5.2 ± 0.2 (731)	5.3 ± 0.1 (942)	4.8 ± 0.3 (309)
Mouse	5.1 ± 0.2 (467)	-4.9 ± 0.2 (481)	5.1 ± 0.2 (561)	5.1 ± 0.4 (213)

As a test of the reliability of our methodology, we compare in Fig. 16 the results of our estimates of the replication parameters h and δ for each N-domain to the corresponding values obtained directly from some fit of the S profile when considering only the intergenic regions where the observed skew is supposed to result from replication only. As observed in Fig. 16(a) for the parameter h and in Fig. 15(b) for the parameter δ, a large majority of points fall, up to the numerical uncertainty, on the diagonal. Thus, except for a small percentage of N-domains where intergenic coverage (Fig. 12) is too small to allow us to compute the replication parameters directly from the intergenic skew profile, the estimates reported in Table 2 are quite consistent with the skew values observed genome wide in intergenic sequences.

5.2.2. Transcription bias

To estimate the mean transcription bias of the genes belonging to a given N-domain, we consider the transcription-associated skew S_{T} obtained after subtracting the estimated N-shaped replication profile as illustrated in Fig. 13(c). Then as first proposed in [136], the transcription skew of the genes is measured by averaging S_{T} over intron sequences after removing 490 bp at each intron extremities in order to get rid of the contribution to the skew coming from splicing signals. In Fig. 17 and Table 3 are reported the transcription bias so estimated for 726 (+) genes and 731 (−) genes of length

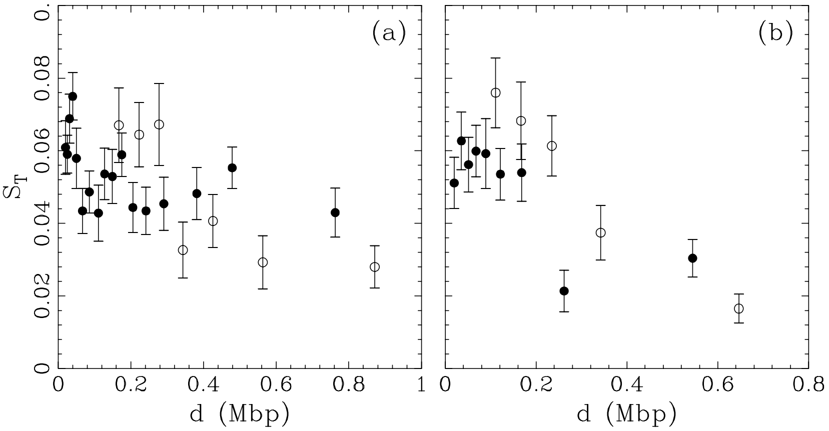

Fig. 18. Mean transcription skew computed on the gene strand for $R+$ (●) and $R-$ (○) genes of length ⩾ 20 kbp as a function of the (non-masked) distance to the closest N-domain border: (a) Human genes; (b) mouse genes (see Table 3). Each point corresponds to average over 40 gene promoter positions and transcriptional skews; vertical error bars represent SEM.

⩾ 20 kbp so that the total intronic coverage is enough to ensure convergence in the estimate of the gene transcription skew. The histograms of S_T values for (+) and (−) genes are remarkably symmetric with means $\bar{S}_T^{(+)} = 5.1 \pm 0.2\%$ and $\bar{S}_T^{(-)} = -5.2 \pm 0.2\%$ respectively. This confirms that the local genic contribution of transcription to the total skew is of the same magnitude (∼ 5%) than the contribution induced by replication (∼ 7.5%) which can be seen *a posteriori* as a justification of the need to develop a methodology capable to disentangle these two skew components. This suggests that previous estimates [136,137] of total transcription bias (∼ 8%, see Fig. 1) were biased by the presence of a replication-associated skew. According to the remarkable gene organization evidenced in replication N-domains in [70], we further focus on $R+$ genes ((+) genes in the 5′ N-domain half and (−) genes in the 3′ N-domain half) that are transcribed in the same direction as the putative replication fork progression and $R-$ genes ((−) genes in the 5′ N-domain half and (+) genes in the 3′ N-domain half) that are transcribed in the opposite direction. Close to the bordering putative origins, $R+$ genes are more abundant and longer than $R-$ genes. As reported in Table 3, $R+$ and $R-$ genes have similar transcriptional skew $\bar{S}_T^{R+} = 5.3 \pm 0.1\%$ and $\bar{S}_T^{R-} = 4.8 \pm 0.3\%$. Furthermore, as shown in Fig. 18(a), when plotting their transcriptional skew versus their distance to their proximal N-domain border, we observe some significant decrease for both $R+$ and $R-$ genes from values ∼ 7% close to the putative replication origin to values ∼ 3% ($R-$) and ∼ 4% ($R+$). Thus genes lying in a close neighborhood of the replication origins are more biased by transcription than central genes, a result which seems to be quite consistent with the recent observation that these putative replication origins correspond to particular chromatin regions permissive to transcription, that may have been imprinted in the DNA sequence during evolution [27]. In particular, if according to the size of the error bars, we cannot conclude about the significance of the damped oscillations observed for the $R+$ transcription bias in Fig. 18(a), they strikingly resemble to those displayed by the mean density profile of PolII binding at gene promoter (±2 kbp around TSS) when plotted versus the distance to the closest N-domain border [27].

5.3. Mouse autosomes

Among the 587 N-domains detected in the 19 mouse autosomes, only 2 are discarded by our disentangling methodology as incompatible with our working model (16). The results presented in this section will thus correspond to a statistical analysis performed on 585 N-domains.

5.3.1. Replication bias

As shown in Fig. 15, the histograms of replication parameters ($S_R^{5'}$, $S_R^{3'}$, h) values computed with our methodology cannot be statistically distinguished from the ones previously obtained for the human autosomes. The detected N-domains have a zero mean replication skew ($\bar{h} = (-1.8 \pm 7.9)\,10^{-4}$) and an antisymmetric shape with $\bar{S}_R^{5'} = 6.8 \pm 0.2\%$ and $\bar{S}_R^{3'} = -6.8 \pm 0.2\%$ (Table 2) corresponding to an upward jump in the mouse skew profile, at the detected putative replication origins, of characteristic amplitude ∼ 13.6% similar to the ∼ 14.4% previously observed for the human autosomes.

5.3.2. Transcription bias

Similarly, the estimates of the transcription bias for sense (Fig. 17(a)) and anti-sense (Fig. 17(b)) mouse genes are in remarkable agreement with those obtained for human genes. As reported in Table 3, we get the following mean values $\bar{S}_T^{(+)} = 5.1 \pm 0.2\%$, $\bar{S}_T^{(-)} = -4.9 \pm 0.2\%$ for sense and anti-sense genes respectively, and $\bar{S}_T^{R+} = 5.1 \pm 0.2\%$, $\bar{S}_T^{R-} = 5.1 \pm 0.4\%$ for the transcription bias for $R+$ and $R-$ genes. Interestingly, the transcription bias for $R+$ and $R-$ genes decrease from values ∼ 6–8% close to the N-domain borders down to values ∼ 2% when going away from the N-domain center (Fig. 18(b)). Again we recover some feature previously evidenced with human genes, except that no (damped) oscillatory behavior is observed in the decrease of S_{T}^{R+} as a function of the distance to the nearest N-domain border. This difference will be the subject of further investigation via a specific comparison of the replication and transcription biases of orthologuous genes in the human and mouse genomes and also of their relative positioning inside the replication N-domains.

6. Conclusions

In this paper, we developed a multi-scale methodology based on the continuous wavelet transform with an adapted (N-shaped) analyzing wavelet. The implementation of this method allowed us to perform genome wide analysis of DNA strand asymmetry profiles with the specific goal to extract from these intrinsically noisy profiles, the contributions associated with replication and transcription respectively. The application of this methodology to the 22 human and 19 mouse autosomes provides reliable estimates of the replication and transcription skews for each individual gene that belongs to a replication N-domain. The quantitative agreement obtained between the human and mouse skew estimates strongly suggest that both the replication and transcription components of the skew inside the N-domains are conserved in mammalian genomes. These results open new perspectives in genomic and post-genomic data analysis. In particular, the possibility to correlate separately the replication-associated and transcription-associated skew components to gene expression data (expression level, expression breath, ...), open chromatin markers (DNase I cleavage, DNA hypomethylation, nucleosome free regions) and other epigenetic data (histone variants, histone tail modifications, ...) is likely to shed a new light on chromatin-mediated regulation of replication and transcription processes. The joint analysis of gene transcription bias and expression data in germ-line cells is in current progress.

Acknowledgments

This work was supported by the Conseil Régional Rhônes-Alpes (Emergence 2005) and the Agence Nationale de la Recherche (ANR), project HUGOREP (NT05-3-41825).

References

[1] M. Anglana, F. Apiou, A. Bensimon, M. Debatisse, Dynamics of DNA replication in mammalian somatic cells: Nucleotide pool modulates origin choice and interorigin spacing, Cell 114 (3) (2003) 385–394.

[2] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, E. Freysz, G. Grasseau, J.-F. Muzy, B. Pouligny, Wavelet transform of fractals, in: Wavelets and Applications, Springer, Berlin, 1992, pp. 286–352.

[3] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, J.-F. Muzy, Ondelettes, Multifractales et Turbulences : de l'ADN aux Croissances Cristallines, Diderot Editeur, Arts et Sciences, Paris, 1995.

[4] A. Arneodo, F. Argoul, E. Bacry, J.-F. Muzy, M. Tabard, Golden mean arithmetic in the fractal branching of diffusion-limited aggregates, Phys. Rev. Lett. 68 (23) (1992) 3456–3459.

[5] A. Arneodo, F. Argoul, J. Elezgaray, G. Grasseau, Wavelet transform analysis of fractals: Application to nonequilibrium phase transitions, in: G. Turchetti (Ed.), Nonlinear Dynamics, World Scientific, Singapore, 1989, pp. 130–180.

[6] A. Arneodo, F. Argoul, J.-F. Muzy, M. Tabard, Structural 5-fold symmetry in the fractal morphology of diffusion-limited aggregates, Physica A 188 (1–3) (1992) 217–242.

[7] A. Arneodo, F. Argoul, J.-F. Muzy, M. Tabard, Uncovering Fibonacci sequences in the fractal morphology of diffusion-limited aggregates, Phys. Lett. A 171 (1–2) (1992) 31–36.

[8] A. Arneodo, B. Audit, E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, Y. d'Aubenton-Carafa, M. Huvet, C. Thermes, Fractals and wavelets: What can we learn on transcription and replication from wavelet-based multifractal analysis of DNA sequences?, in: Encyclopedia of Complexity and System Science, 2008, in press.

[9] A. Arneodo, B. Audit, N. Decoster, J.-F. Muzy, C. Vaillant, Wavelet based multifractal formalism: Application to DNA sequences, satellite images of the cloud structure and stock market data, in: The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes, Springer Verlag, Berlin, 2002, pp. 26–102.

[10] A. Arneodo, B. Audit, C. Faivre-Moskalenko, J. Moukhtar, C. Vaillant, F. Argoul, Y. d'Aubenton-Carafa, C. Thermes, From DNA sequence to chromatin organization: The fundamental role of genomic long-range correlations, Bull. Acad. Roy. Belg. Mém. Cl. Sci. Collect. 8 Sér. 3 XXVIII (2049) (2008).

[11] A. Arneodo, E. Bacry, P.V. Graves, J.-F. Muzy, Characterizing long-range correlations in DNA sequences from wavelet analysis, Phys. Rev. Lett. 74 (16) (1995) 3293–3296.

[12] A. Arneodo, E. Bacry, J.-F. Muzy, The thermodynamics of fractals revisited with wavelets, Physica A 213 (1–2) (1995) 232–275.

[13] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J.-F. Muzy, C. Thermes, Nucleotide composition effects on the long-range correlations in human genes, Eur. Phys. J. B 1 (2) (1998) 259–263.

[14] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E.-B. Brodie of Brodie, S. Nicolay, P. St-Jean, C. Thermes, M. Touchon, C. Vaillant, DNA in chromatin: From genome-wide sequence analysis to the modeling of replication in mammals, Adv. Chem. Phys. 135 (2007) 203–252.

[15] A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P.V. Graves, J.-F. Muzy, C. Thermes, Wavelet based fractal analysis of DNA sequences, Physica D 96 (1–4) (1996) 291–320.

[16] A. Arneodo, N. Decoster, P. Kestener, S.G. Roux, A wavelet-based method for multifractal image analysis: From theoretical concepts to experimental applications, Adv. Imaging Electr. Phys. 126 (2003) 1–92.

[17] A. Arneodo, N. Decoster, S.G. Roux, Intermittency, log-normal statistics, and multifractal cascade process in high-resolution satellite images of cloud structure, Phys. Rev. Lett. 83 (6) (1999) 1255–1258.

[18] A. Arneodo, N. Decoster, S.G. Roux, A wavelet-based method for multifractal image analysis. I. Methodology and test applications on isotropic and anisotropic random rough surfaces, Eur. Phys. J. B 15 (3) (2000) 567–600.

[19] A. Arneodo, G. Grasseau, M. Holschneider, Wavelet transform of multifractals, Phys. Rev. Lett. 61 (20) (1988) 2281–2284.

[20] A. Arneodo, S. Manneville, J.-F. Muzy, Towards log-normal statistics in high Reynolds number turbulence, Eur. Phys. J. B 1 (1) (1998) 129–140.

[21] A. Arneodo, S. Manneville, J.-F. Muzy, S.G. Roux, Revealing a lognormal cascading process in turbulent velocity statistics with wavelet analysis, Philos. Trans. R. Soc. Lond. A 357 (1760) (1999) 2415–2438.

[22] A. Arneodo, J.-F. Muzy, D. Sornette, “Direct” causal cascade in the stock market, Eur. Phys. J. B 2 (2) (1998) 277–282.

[23] J. Arrault, A. Arneodo, A. Davis, A. Marshak, Wavelet based multifractal analysis of rough surfaces: Application to cloud models and satellite data, Phys. Rev. Lett. 79 (1) (1997) 75–78.

[24] B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, DNA replication timing data corroborate in silico human replication origin predictions, Phys. Rev. Lett. 99 (24) (2007) 248102.

[25] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.-F. Muzy, A. Arneodo, Long-range correlations in genomic DNA: A signature of the nucleosomal structure, Phys. Rev. Lett. 86 (11) (2001) 2471–2474.

[26] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Long-range correlations between DNA bending sites: Relation to the structure and dynamics of nucleosomes, J. Mol. Biol. 316 (4) (2002) 903–918.

[27] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells, Nucl. Acids Res. 37 (18) (2009) 6064–6075.

[28] M.Y. Azbel', Universality in a DNA statistical structure, Phys. Rev. Lett. 75 (1) (1995) 168–171.

[29] E. Bacry, J.-F. Muzy, A. Arneodo, Singularity spectrum of fractal signals from wavelet analysis: Exact results, J. Stat. Phys. 70 (1993) 635–674.

[30] A. Beletskii, A. Grigoriev, S. Joyce, A.S. Bhagwat, Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome, J. Mol. Biol. 300 (5) (2000) 1057–1065.

[31] S.P. Bell, A. Dutta, DNA replication in eukaryotic cells, Annu. Rev. Biochem. 71 (2002) 333–374.

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

[33] J.A. Bogan, D.A. Natale, M.L. Depamphilis, Initiation of eukaryotic DNA replication: Conservative or liberal?, J. Cell. Physiol. 184 (2) (2000) 139–150.

[34] B. Borštnik, D. Pumpernik, D. Lukman, Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences, Europhys. Lett. 23 (6) (1993) 389–394.

[35] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, From DNA sequence analysis to modeling replication in the human genome, Phys. Rev. Lett. 94 (24) (2005) 248103.

[36] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, Phys. Rev. E 51 (5) (1995) 5084–5091.

[37] M. Bulmer, Strand symmetry of mutation rates in the beta-globin region, J. Mol. Evol. 33 (4) (1991) 305–310.

[38] L.B. Caddle, J.L. Grant, J. Szatkiewicz, J. van Hase, B.-J. Shirley, J. Bewersdorf, C. Cremer, A. Arneodo, A. Khalil, K.D. Mills, Chromosome neighborhood composition determines translocation outcomes after exposure to high-dose radiation in primary cells, Chrom. Res. 15 (8) (2007) 1061–1073.

[39] J.-C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, M.-N. Prioleau, Genome-wide studies highlight indirect links between human replication origins and gene regulation, Proc. Natl. Acad. Sci. USA 105 (41) (2008) 15837–15842.

[40] E. Chargaff, Structure and function of nucleic acids as cell constituents, Fed. Proc. 10 (3) (1951) 654–659.

[41] C.A. Chatzidimitriou-Dreismann, D. Larhammar, Long-range correlations in DNA, Nature 361 (6409) (1993) 212–213.

[42] J. Combes, A. Grossmann, P. Tchamitchian (Eds.), Wavelets, Springer, Berlin, 1989.

[43] D. Coverley, R.A. Laskey, Regulation of eukaryotic DNA replication, Annu. Rev. Biochem. 63 (1994) 745–776.

[44] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.

[45] N. Decoster, S.G. Roux, A. Arneodo, A wavelet-based method for multifractal image analysis. II. Applications to synthetic multifractal rough surfaces, Eur. Phys. J. B 15 (4) (2000) 739–764.

[46] J. Delour, J.-F. Muzy, A. Arneodo, Intermittency of 1D velocity spatial profiles in turbulence: A magnitude cumulant analysis, Eur. Phys. J. B 23 (2) (2001) 243–248.

[47] C. Demeret, Y. Vassetzky, M. Méchali, Chromatin remodelling and DNA replication: From nucleosomes to loop domains, Oncogene 20 (24) (2001) 3086–3093.

[48] L. Duret, Evolution of synonymous codon usage in metazoans, Curr. Opin. Genet. Dev. 12 (6) (2002) 640–649.

[49] G. Erlebacher, M. Hussaini, L. Jameson (Eds.), Wavelets: Theory and Applications, Oxford University Press, Oxford, 1996.

[50] M. Farge, J. Hunt, J. Vassilicos (Eds.), Wavelets, Fractals and Fourier, Clarendon Press, Oxford, 1996.

[51] J.W. Fickett, D.C. Torney, D.R. Wolf, Base compositional structure of genomes, Genomics 13 (4) (1992) 1056–1064.

[52] D. Fisher, M. Méchali, Vertebrate HoxB gene expression requires DNA replication, EMBO J. 22 (14) (2003) 3737–3748.

[53] M.P. Francino, H. Ochman, Strand symmetry around the beta-globin origin of replication in primates, Mol. Biol. Evol. 17 (3) (2000) 416–422.

[54] M.P. Francino, H. Ochman, Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences, Mol. Biol. Evol. 18 (6) (2001) 1147–1150.

[55] A.C. Frank, J.R. Lobry, Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms, Gene 238 (1) (1999) 65–77.

[56] J.M. Freeman, T.N. Plasterer, T.F. Smith, S.C. Mohr, Patterns of genome organization in bacteria, Science 279 (1998) 1827.

[57] K. Gardiner, Base composition and gene distribution: Critical patterns in mammalian genome organization, Trends Genet. 12 (12) (1996) 519–524.

[58] S.A. Gerbi, A.K. Bielinsky, DNA replication and chromatin, Curr. Opin. Genet. Dev. 12 (2) (2002) 243–248.

[59] A. Gierlik, M. Kowalczuk, P. Mackiewicz, M.R. Dudek, S. Cebrat, Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome?, J. Theor. Biol. 202 (4) (2000) 305–314.

[60] D.M. Gilbert, Making sense of eukaryotic DNA replication origins, Science 294 (5540) (2001) 96–100.

[61] D.M. Gilbert, In search of the holy replicator, Nat. Rev. Mol. Cell Biol. 5 (10) (2004) 848–855.

[62] P. Goupillaud, A. Grossmann, J. Morlet, Cycle-octave and related transforms in seismic signal analysis, Geoexploration 23 (1984) 85–102.

[63] P. Green, B. Ewing, W. Miller, P.J. Thomas, E.D. Green, Transcription-associated mutational asymmetry in mammalian evolution, Nat. Genet. 33 (4) (2003) 514–517.

[64] A. Grossmann, J. Morlet, Decomposition of Hardy functions into square integrable wavelets of constant shape, SIAM J. Math. Anal. 15 (1984) 723–736.

[65] A. Grossmann, J. Morlet, Decomposition of functions into wavelets of constant shape and related transforms, in: L. Streit (Ed.), Mathematics and Physics, Lectures on Recent Results, World Scientific, 1985, pp. 135–165.

[66] H.G.E. Hentschel, Stochastic multifractality and universal scaling distributions, Phys. Rev. E 50 (1) (1994) 243–261.

[67] H. Herzel, I. Große, Measuring correlations in symbol sequences, Physica A 216 (4) (1995) 518–542.

[68] M. Holschneider, On the wavelet transform of fractal objects, J. Stat. Phys. 50 (1988) 963–993.

[69] M. Holschneider, P. Tchamitchian, Régularité locale de la fonction non-différentiable de Riemann, in: P.G. Lemarié (Ed.), Les Ondelettes en 1989, Springer, Berlin, 1990, pp. 102–124.

[70] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Human gene organization driven by the coordination of replication and transcription, Genome Res. 17 (9) (2007) 1278–1285.

[71] O. Hyrien, M. Méchali, Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos, EMBO J. 12 (12) (1993) 4511–4520.

[72] P.C. Ivanov, L.A. Amaral, A.L. Goldberger, S. Havlin, M.G. Rosenblum, Z.R. Struzik, H.E. Stanley, Multifractality in human heartbeat dynamics, Nature 399 (6735) (1999) 461–465.

[73] P.C. Ivanov, M.G. Rosenblum, C.K. Peng, J. Mietus, S. Havlin, H.E. Stanley, A.L. Goldberger, Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis, Nature 383 (6598) (1996) 323–327.

[74] F. Jacob, S. Brenner, F. Cuzin, On the regulation of DNA replication in bacteria, Cold Spring Harb. Symp. Quant. Biol. 28 (1963) 329–342.

[75] S. Jaffard, Hölder exponents at given points and wavelet coefficients, C. R. Acad. Sci. Paris Sér. I 308 (4) (1989) 79–81.

[76] S. Jaffard, Pointwise smoothness, two-microlocalization and wavelet coefficients, Publ. Mat. 35 (1991) 155–168.

[77] S. Jaffard, Multifractal formalism for functions, Part I: Results valid for all functions, SIAM J. Math. Anal. 28 (4) (1997) 944–970.

[78] S. Jaffard, Multifractal formalism for functions, Part II: Self-similar functions, SIAM J. Math. Anal. 28 (4) (1997) 971–998.

[79] S. Jaffard, B. Lashermes, P. Abry, Wavelet leaders in multifractal analysis, in: T. Qian, M.I. Vai, Y. Xu (Eds.), Wavelet Analysis and Applications, Birkhäuser Verlag, Basel, Switzerland, 2006, pp. 219–264.

[80] S. Jaffard, Y. Meyer, R. Ryan (Eds.), Wavelets: Tools for Science and Technology, SIAM, Philadelphia, 2001.

[81] S. Karlin, V. Brendel, Patchiness and correlations in DNA sequences, Science 259 (5095) (1993) 677–680.

[82] P. Kestener, A. Arneodo, Three-dimensional wavelet-based multifractal method: The need for revisiting the multifractal description of turbulence dissipation data, Phys. Rev. Lett. 91 (19) (2003) 194501.

[83] P. Kestener, A. Arneodo, Generalizing the wavelet-based multifractal formalism to random vector fields: Application to three-dimensional turbulence velocity and vorticity data, Phys. Rev. Lett. 93 (4) (2004) 044501.

[84] P. Kestener, A. Arneodo, A multifractal formalism for vector-valued random fields based on wavelet analysis: Application to turbulent velocity and vorticity 3D numerical data, Stoch. Environ. Res. Risk Assess. 22 (2007) 421.

[85] P. Kestener, J.-M. Lina, P. Saint-Jean, A. Arneodo, Wavelet-based multifractal formalism to assist in diagnosis in digitized mammograms, Image Anal. Stereol. 20 (2001) 169–174.

[86] A. Khalil, J. Grant, L. Caddle, E. Atzema, K. Mills, A. Arneodo, Chromosome territories have a highly non-spherical morphology and non-random positioning, Chrom. Res. 15 (2007) 889–916.

[87] A. Khalil, G. Joncas, F. Nekka, P. Kestener, A. Arneodo, Morphological analysis of H_I features. II. Wavelet-based multifractal formalism, Astrophys. J. Suppl. Ser. 165 (2) (2006) 512–550.

[88] A. Kuhn, F. Argoul, J.-F. Muzy, A. Arneodo, Structural-analysis of electroless deposits in the diffusion-limited regime, Phys. Rev. Lett. 73 (22) (1994) 2998–3001.

[89] D. Larhammar, C.A. Chatzidimitriou-Dreismann, Biological origins of long-range correlations and compositional variations in DNA, Nucl. Acids Res. 21 (22) (1993) 5167–5170.

[90] T.I. Lee, R.G. Jenner, L.A. Boyer, M.G. Guenther, S.S. Levine, R.M. Kumar, B. Chevalier, S.E. Johnstone, M.F. Cole, K. Isono, H. Koseki, T. Fuchikami, K. Abe, H.L. Murray, J.P. Zucker, B. Yuan, G.W. Bell, E. Herbolsheimer, N.M. Hannett, K. Sun, D.T. Odom, A.P. Otte, T.L. Volkert, D.P. Bartel, D.A. Melton, D.K. Gifford, R. Jaenisch, R.A. Young, Control of developmental regulators by polycomb in human embryonic stem cells, Cell 125 (2) (2006) 301–313.

[91] W. Li, Generating non trivial long-range correlations and $1/f$ spectra by replication and mutation, Int. J. Bifurc. Chaos 2 (1992) 137–154.

[92] W. Li, The study of correlation structures of DNA sequences: A critical review, Comput. Chem. 21 (4) (1997) 257–271.

[93] W.T. Li, T.G. Marr, K. Kaneko, Understanding long-range correlations in DNA-sequences, Physica D 75 (1–3) (1994) 392–416.

[94] W. Li, G. Stolovitzky, P. Bernaola-Galván, J.L. Oliver, Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes, Genome Res. 8 (9) (1998) 916–928.

[95] J.R. Lobry, Properties of a general model of DNA evolution under no-strand-bias conditions, J. Mol. Evol. 40 (3) (1995) 326–330.

[96] I. Lucas, A. Palakodeti, Y. Jiang, D.J. Young, N. Jiang, A.A. Fernald, M.M. Le Beau, High-throughput mapping of origins of replication in human cells, EMBO Rep. 8 (8) (2007) 770–777.

[97] S. Mallat, A Wavelet Tour in Signal Processing, Academic Press, New York, 1998.

[98] S. Mallat, W. Hwang, Singularity detection and processing with wavelets, IEEE Trans. Info. Theory 38 (2) (1992) 617–643.

[99] S. Mallat, S. Zhong, Characterization of signals from multiscale edges, IEEE Trans. Patt. Recogn. Mach. Intell. 14 (7) (1992) 710–732.

[100] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev. E 52 (3) (1995) 2939–2950.

[101] A.J. McNairn, D.M. Gilbert, Epigenomic replication: Linking epigenetics to DNA replication, Bioessays 25 (7) (2003) 647–656.

[102] M. Méchali, DNA replication origins: From sequence specificity to epigenetics, Nat. Rev. Genet. 2 (8) (2001) 640–645.

[103] Y. Meyer (Ed.), Wavelets and Applications, Springer, Berlin, 1992.

[104] Y. Meyer, S. Roques (Eds.), Progress in Wavelets Analysis and Applications, Éditions Frontières, Gif-sur-Yvette, 1993.

[105] N. Mordant, J. Delour, E. Léveque, A. Arneodo, J.-F. Pinton, Long-time correlations in Lagrangian dynamics: A key to intermittency in turbulence, Phys. Rev. Lett. 89 (25) (2002) 254502.

[106] N. Mordant, J. Delour, E. Léveque, O. Michel, A. Arneodo, J.-F. Pinton, Lagrangian velocity fluctuations in fully developed turbulence: Scaling, intermittency and dynamics, J. Stat. Phys. 113 (5–6) (2003) 701–717.

[107] J. Moukhtar, E. Fontaine, C. Faivre-Moskalenko, A. Arneodo, Probing persistence in DNA curvature properties with atomic force microscopy, Phys. Rev. Lett. 98 (17) (2007) 178101.

[108] J. Mrázek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, Proc. Natl. Acad. Sci. USA 95 (7) (1998) 3720–3725.

[109] J.-F. Muzy, E. Bacry, A. Arneodo, Wavelets and multifractal formalism for singular signals: Application to turbulence data, Phys. Rev. Lett. 67 (25) (1991) 3515–3518.

[110] J.-F. Muzy, E. Bacry, A. Arneodo, Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method, Phys. Rev. E 47 (2) (1993) 875–884.

[111] J.-F. Muzy, E. Bacry, A. Arneodo, The multifractal formalism revisited with wavelets, Int. J. Bifurc. Chaos 4 (1994) 245–302.

[112] J.-F. Muzy, D. Sornette, J. Delour, A. Arneodo, Multifractal returns and hierarchical portfolio theory, Quant. Finance 1 (2001) 131–148.

[113] S. Nee, Uncorrelated DNA walks, Nature 357 (6378) (1992) 450.

[114] S. Nicolay, Analyse des séquences d'ADN par la transformée en ondelettes : extraction d'informations structurelles, dynamiques et fonctionnelles, Ph.D. thesis, University of Liège, Belgium, 2006.

[115] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Low frequency rhythms in human DNA sequences: A key to the organization of gene location and orientation?, Phys. Rev. Lett. 93 (10) (2004) 108101.

[116] S. Nicolay, E.-B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo, Bifractality of human DNA strand-asymmetry profiles results from transcription, Phys. Rev. E 75 (3) (2007) 032902.

[117] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-range correlations in nucleotide sequences, Nature 356 (6365) (1992) 168–170.

[118] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, H.E. Stanley, Finite-size effects on long-range correlations: Implications for analysing DNA sequences, Phys. Rev. E 47 (5) (1993) 3730–3733.

[119] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E 49 (2) (1994) 1685–1689.

[120] E.P. Rocha, A. Danchin, A. Viari, Universal replication biases in bacteria, Mol. Microbiol. 32 (1) (1999) 11–16.

[121] S.G. Roux, A. Arneodo, N. Decoster, A wavelet-based method for multifractal image analysis. III. Applications to high-resolution satellite images of cloud structure, Eur. Phys. J. B 15 (4) (2000) 765–786.

[122] S. Roux, J.-F. Muzy, A. Arneodo, Detecting vorticity filaments using wavelet analysis: About the statistical contribution of vorticity filaments to intermittency in swirling turbulent flows, Eur. Phys. J. B 8 (2) (1999) 301–322.

[123] S. Roux, V. Venugopal, K. Fineberg, A. Arneodo, E. Foufoula-Georgiou, Evidence for inherent nonlinearity in temporal rainfall, Adv. Water Resour. 32 (2009) 41–48.

[124] R. Rudner, J.D. Karkas, E. Chargaff, Separation of B. subtilis DNA into complementary strands. 3. Direct analysis, Proc. Natl. Acad. Sci. USA 60 (3) (1968) 921–922.

[125] M. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, L. Raphael (Eds.), Wavelets and Their Applications, Jones and Barlett, Boston, 1992.

[126] T. Sasaki, T. Sawado, M. Yamaguchi, T. Shinomiya, Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster, Mol. Cell. Biol. 19 (1) (1999) 547–555.

[127] D. Schübeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, M. Groudine, Genome-wide DNA replication profile for Drosophila melanogaster: A link between transcription and replication timing, Nat. Genet. 32 (3) (2002) 438–442.

[128] C. Shioiri, N. Takahata, Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry, J. Mol. Evol. 53 (4–5) (2001) 364–376.

[129] B. Silverman, J. Vassilicos (Eds.), Wavelets: The Key to Intermittent Information?, Oxford University Press, Oxford, 2000.

[130] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, S.M. Ossadnik, C.-K. Peng, M. Simons, Fractal landscapes in biological systems, Fractals 1 (3) (1993) 283–301.

[131] J.Q. Svejstrup, Mechanisms of transcription-coupled DNA repair, Nat. Rev. Mol. Cell Biol. 3 (1) (2002) 21–29.

[132] The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (7146) (2007) 799–816.

[133] E.R. Tillier, R.A. Collins, The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes, J. Mol. Evol. 50 (3) (2000) 249–257.

[134] B. Torresani, Analyse Continue par Ondelettes, Éditions de Physique, Les Ulis, 1998.

[135] M. Touchon, Biais de composition chez les mammifères : rôle de la transcription et de la réplication, Ph.D. thesis, University Denis Diderot, Paris VII, France, 2005.

[136] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes, Nucl. Acids Res. 32 (17) (2004) 4969–4978.

[137] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, C. Thermes, Transcription-coupled TA and GC strand asymmetries in the human genome, FEBS Lett. 555 (3) (2003) 579–582.

[138] M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes, Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins, Proc. Natl. Acad. Sci. USA 102 (28) (2005) 9836–9841.

[139] C. Vaillant, B. Audit, A. Arneodo, Thermodynamics of DNA loops with long-range correlated structural disorder, Phys. Rev. Lett. 95 (6) (2005) 068101.

[140] C. Vaillant, B. Audit, A. Arneodo, Experiments confirm the influence of genome long-range correlations on nucleosome positioning, Phys. Rev. Lett. 99 (21) (2007) 218103.

[141] C. Vaillant, B. Audit, C. Thermes, A. Arneodo, Formation and positioning of nucleosomes: Effect of sequence-dependent long-range correlated structural disorder, Eur. Phys. J. E 19 (3) (2006) 263–277.

[142] V. Venugopal, S.G. Roux, E. Foufoula-Georgiou, A. Arneodo, Revisiting multifractality of high-resolution temporal rainfall using a wavelet-based formalism, Water Resour. Res. 42 (6) (2006) W06D14.

[143] V. Venugopal, S.G. Roux, E. Foufoula-Georgiou, A. Arneodo, Scaling behavior of high resolution temporal rainfall: New insights from a wavelet-based cumulant analysis, Phys. Lett. A 348 (3–6) (2006) 335–345.

[144] G.M. Viswanathan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA, Physica A 249 (1–4) (1998) 581–586.

[145] R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett. 68 (25) (1992) 3805–3808.

[146] R.F. Voss, Long-range fractal correlations in DNA introns and exons, Fractals 2 (1) (1994) 1–6.

[147] L. Zaghloul, A. Arneodo, C. Thermes, B. Audit, Large replication domains with homogeneous composition delimit heterochromatic gene deserts, in preparation.