# Fractals and Wavelets: what can we learn on transcription and replication from wavelet-based multifractal analysis of DNA sequences?

Alain Arneodo, Benjamin Audit, and Edward-Benedict Brodie of Brodie

*Laboratoire Joliot-Curie and Laboratoire de Physique, ENS-Lyon, CNRS, 46 Allée d'Italie, 69364 Lyon Cedex 07, France*

Samuel Nicolay

*Institut de Mathématique, Université de Liège, Grande Traverse 12, 4000 Liège, Belgium*

Marie Touchon

*Génétique des Génomes Bactériens, Institut Pasteur, CNRS, Paris, France and Atelier de Bioinformatique, Université Pierre et Marie Curie - Paris 6, Paris, France*

Yves d'Aubenton-Carafa, Maxime Huvet, and Claude Thermes

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

(Dated: August 30, 2007)

**Contents**

**Glossary**

**Fractal**

Fractals are complex mathematical objects that are invariant with respect to dilations (**self-similarity**) and therefore do not possess a characteristic length scale. Fractal objects display scale-invariance properties that can either fluctuate from point to point (**multifractal**) or be homogeneous (**monofractal**). Mathematically, these properties should hold over all scales. However, in the real world, there are necessarily lower and upper bounds over which self-similarity applies.

**Wavelet Transform**

The continuous wavelet transform (WT) is a mathematical technique introduced in the early 1980s 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. More specifically, the WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet, by means of translations and dilations.

**Wavelet Transform Modulus Maxima Method**

The WTMM method provides a unified statistical (thermodynamic) description of multifractal distributions including measures and functions. This method relies on the computation of partition functions from the wavelet transform skeleton defined by the wavelet transform modulus maxima (WTMM). This skeleton provides an adaptive space-scale partition of the fractal distribution under study, from which one can extract the $D(h)$ singularity spectrum as the equivalent of a thermodynamic potential (entropy). With some appropriate choice of the analyzing wavelet, one can show that the WTMM method provides a natural generalization of the classical box-counting and structure function techniques.

**Compositional Strand Asymmetry**

The DNA double helix is made of two strands that are maintained together by hydrogen bonds involved in the base-pairing between Adenine (resp. Guanine) on one strand and Thymine (resp. Cytosine) on the other strand. Under no-strand bias conditions, *i.e.* when mutation rates are identical on the two strands, in other words when the two strands are strictly equivalent, one expects equimolarities of adenine and thymine and of guanine and cy-

tosine on each DNA strand, a property named Chargaff's second parity rule. Compositional strand asymmetry refers to deviations from this rule which can be assessed by measuring departure from intrastrand equimolarities. Note that two major biological processes, **transcription** and **replication**, both requiring the opening of the double helix, actually break the symmetry between the two DNA strands and can thus be at the origin of compositional strand asymmetries.

**Eukaryote**

Organisms whose cells contain a nucleus, the structure containing the genetic material arranged into chromosomes. Eukaryotes constitute one of the three domains of life, the two others, called prokaryotes (without nucleus), being the eubacteria and the archaebacteria.

**Transcription**

Transcription is the process whereby the DNA sequence of a gene is enzymatically copied into a complementary messenger RNA. In a following step, **translation** takes place where each messenger RNA serves as a template to the biosynthesis of a specific protein.

**Replication**

DNA replication is the process of making an identical copy of a double-stranded DNA molecule. DNA replication is an essential cellular function responsible for the accurate transmission of genetic information though successive cell generations. This process starts with the binding of initiating proteins to a DNA locus called **origin of replication**. The recruitment of additional factors initiates the bi-directional progression of two replication forks along the chromosome. In eukaryotic cells, this binding event happens at a multitude of replication origins along each chromosome from which replication propagates until two converging forks collide at a **terminus of replication**.

**Chromatin**

Chromatin is the compound of DNA and proteins that forms the chromosomes in living cells. In eukaryotic cells, chromatin is located in the nucleus.

**Histones**

Histones are a major family of proteins found in eukaryotic chromatin. The wrapping of DNA around a core of 8 histones forms a **nucleosome**, the first step of eukaryotic DNA compaction.

## I. OVERVIEW OF THE SUBJECT AND ITS IMPORTANCE

The continuous wavelet transform (WT) is a mathematical technique introduced in signal analysis in th early 1980s [1, 2]. Since then, it has been the subject of considerable theoretical developments and practical applications in a wide variety of fields. The WT has been early recognized as a mathematical microscope that is well adapted to reveal the hierarchy that governs the spatial distribution of singularities of multifractal measures [3–5]. What makes the WT of fundamental use in the present study is that its singularity scanning ability equally applies to singular functions than to singular measures [3–11]. This has led Alain Arneodo and his collaborators [12–16] to elaborate a unified thermodynamic description of multifractal distributions including measures and functions, the so-called Wavelet Transform Modulus Maxima (WTMM) method. By using wavelets instead of boxes, one can take advantage of the freedom in the choice of these "generalized oscillating boxes" to get rid of possible (smooth) polynomial behavior that might either mask singularities or perturb the estimation of their strength $h$ (Hölder exponent), remedying in this way for one of the main failures of the classical multifractal methods (*e.g.* the box-counting algorithms in the case of measures and the structure function method in the case of functions [12, 13, 15, 16]). The other fundamental advantage of using wavelets is that the skeleton defined by the WTMM [10, 11], provides an adaptative space-scale partitioning from which one can extract the $D(h)$ singularity spectrum via the Legendre transform of the scaling exponents $\tau(q)$ ($q$ real, positive as well as negative) of some partition functions defined from the WT skeleton. We refer the reader to Bacry *et al.* [13], Jaffard [17, 18] for rigorous mathematical results and to Hentschel [19] for the theoretical treatment of random multifractal functions.

Applications of the WTMM method to 1D signals have already provided insights into a wide variety of problems [20], *e.g.*, the validation of the log-normal cascade phenomenology of fully developed turbulence [21–24] and of high-resolution temporal rainfall [25, 26], the characterization and the understanding of long-range correlations in DNA sequences [27–30], the demonstration of the existence of causal cascade of information from large to small scales in financial time series [31, 32], the use of the multifractal formalism to discriminate between healthy and sick heartbeat dynamics [33, 34], the discovery of a Fibonacci structural ordering in 1D cuts of diffusion limited aggregates (DLA) [35–38]. The canonical WTMM method has been further generalized from 1D to 2D with the specific goal to achieve multifractal

analysis of rough surfaces with fractal dimensions $D_F$ anywhere between 2 and 3 [39–41]. The 2D WTMM method has been successfully applied to characterize the intermittent nature of satellite images of the cloud structure [42, 43], to perform a morphological analysis of the anisotropic structure of atomic hydrogen ($H_I$) density in Galactic spiral arms [44] and to assist in the diagnosis in digitized mammograms [45]. We refer the reader to Arneodo *et al.* [46] for a review of the 2D WTMM methodology, from the theoretical concepts to experimental applications. In a recent work, Kestener & Arneodo [47] have further extended the WTMM method to 3D analysis. After some convincing test applications to synthetic 3D monofractal Brownian fields and to 3D multifractal realizations of singular cascade measures as well as their random function counterpart obtained by fractional integration, the 3D WTMM method has been applied to dissipation and enstrophy 3D numerical data issued from direct numerical simulations (DNS) of isotropic turbulence. The results so-obtained have revealed that the multifractal spatial structure of both dissipation and enstrophy fields are likely to be well described by a multiplicative cascade process clearly non-conservative. This contrasts with the conclusions of previous box-counting analysis [48] that failed to estimate correctly the corresponding multifractal spectra because of their intrinsic inability to master non-conservative singular cascade measures [47].

For many years, the multifractal description has been mainly devoted to scalar measures and functions. However, in physics as well as in other fundamental and applied sciences, fractals appear not only as deterministic or random scalar fields but also as vector-valued deterministic or random fields. Very recently, Kestener & Arneodo [49, 50] have combined singular value decomposition techniques and WT analysis to generalize the multifractal formalism to vector-valued random fields. The so-called Tensorial Wavelet Transform Modulus Maxima (TWTMM) method has been applied to turbulent velocity and vorticity fields generated in $(256)^3$ DNS of the incompressible Navier-Stokes equations. This study reveals the existence of an intimate relationship $D_{\mathbf{v}}(h+1) = D_{\boldsymbol{\omega}}(h)$ between the singularity spectra of these two vector fields that are found significantly more intermittent that previously estimated from longitudinal and transverse velocity increment statistics. Furthermore, thanks to the singular value decomposition, the TWTMM method looks very promising for future simultaneous multifractal and structural (vorticity sheets, vorticity filaments) analysis of turbulent flows [49, 50].

## II. INTRODUCTION

The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences has been the subject of considerable increasing interest [20, 51, 52]. During the past fifteen years or so, there has been intense discussion about the existence, the nature and the origin of the long-range correlations (LRC) observed in DNA sequences. Different techniques including mutual information functions [53, 54], auto-correlation functions [55, 56], power-spectra [54, 57, 58], "DNA walk" representation [52, 59], Zipf analysis [60, 61] and entropies [62, 63], were used for the statistical analysis of DNA sequences. For years there has been some permanent debate on rather struggling questions like the fact that the reported LRC might be just an artifact of the compositional heterogeneity of the genome organization [20, 27, 52, 55, 56, 64–67]. Another controversial issue is whether or not LRC properties are different for protein-coding (exonic) and non-coding (intronic, intergenic) sequences [20, 27, 52, 54–59, 61, 68]. Actually, there were many objective reasons for this somehow controversial situation. Most of the pioneering investigations of LRC in DNA sequences were performed using different techniques that all consisted in measuring power-law behavior of some characteristic quantity, *e.g.*, the fractal dimension of the DNA walk, the scaling exponent of the correlation function or the power-law exponent of the power spectrum. Therefore, in practice, they all faced the same difficulties, namely finite-size effects due to the finiteness of the sequence [69–71] and statistical convergence issue that required some precautions when averaging over many sequences [52, 65]. But beyond these practical problems, there was also a more fundamental restriction since the measurement of a unique exponent characterizing the global scaling properties of a sequence failed to resolve multifractality [27], and thus provided very poor information upon the nature of the underlying LRC (if they were any). Actually, it can be shown that for a homogeneous (monofractal) DNA sequence, the scaling exponents estimated with the techniques previously mentioned, can all be expressed as a function of the so-called Hurst or roughness exponent $H$ of the corresponding DNA walk landscape [20, 27, 52]. $H = 1/2$ corresponds to classical Brownian, *i.e.* uncorrelated random walk. For any other value of $H$, the steps (increments) are either positively correlated ($H > 1/2$: persistent random walk) or anti-correlated ($H < 1/2$: anti-persistent random walk).

One of the main obstacles to LRC analysis in DNA sequences is the genuine mosaic

structure of these sequences which are well known to be formed of "patches" of different underlying composition [72–74]. When using the "DNA walk" representation, these patches appear as trends in the DNA walk landscapes that are likely to break scale-invariance [20, 52, 59, 64–67, 75, 76]. Most of the techniques, *e.g.* the variance method, used for characterizing the presence of LRC are 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" [59] and the "detrended fluctuation analysis" [77]. In previous works [27, 28], the WT has been emphasized as a well suited technique to overcome this difficulty. 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. In Ref. [78], from a systematic WT analysis of human exons, CDSs and introns, LRC were found in non-coding sequences as well as in coding regions somehow hidden in their inner codon structure. These results made rather questionable the model based on genome plasticity proposed at that time to account for the reported absence of LRC in coding sequences [27, 28, 52, 54, 59, 68]. More recently, some structural interpretation of these LRC has emerged from a comparative multifractal analysis of DNA sequences using structural coding tables based on nucleosome positioning data [29, 30]. The application of the WTMM method has revealed that the corresponding DNA chain bending profiles are monofractal (homogeneous) 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 $H \simeq 0.6$ (but not for eubacterial sequences for which $H = 0.5$) as the signature of the nucleosomal structure. These LRC were 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 [79, 80]. In addition, these LRC might induce some local hyperdiffusion of these loops which would be a very attractive interpretation of the nucleosomal repositioning dynamics. Over larger distances ($\gtrsim$ 200 bp), stronger LRC with $H \simeq 0.8$ seem to exist in any sequence [29, 30]. These LRC are actually observed in the *S. cerevisiae* nucleosome positioning data [81] suggesting that they are involved in the nucleosome organization in the so-called 30 nm chromatin fiber [82]. 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. In regards

to their potential role in regulating the hierarchical structure and dynamics of chromatin, the recent report [83] of sequence-induced LRC effects on the conformations of naked DNA molecules deposited onto mica surface under 2D thermodynamic equilibrium observed by Atomic Force Microscopy (AFM) is a definite experimental breakthrough.

Our purpose here is to take advantage of the availability of fully sequenced genomes to generalize the application of the WTMM method to genome-wide multifractal sequence analysis when using codings that have a clear functional meaning. According to the second parity rule [84, 85], 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 [86, 87]. Deviations from intrastrand equimolarities have been extensively studied during the past decade and the observed skews have been attributed to asymmetries intrinsic to the replication and transcription processes that both require the opening of the double helix. Actually, during these processes mutational events can affect the two strands differently and an asymmetry can result if one strand undergoes different mutations, or is repaired differently than the other strand. The existence of transcription and/or replication associated strand asymmetries has been mainly established for prokaryote, organelle and virus genomes [88–94]. For a long time the existence of compositional biases in eukaryotic genomes has been unclear and it is only recently that (i) the statistical analysis of eukaryotic gene introns have revealed the presence of transcription-coupled strand asymmetries [95–97] and (ii) the genome wide multi-scale analysis of mammalian genomes has clearly shown some departure from intrastrand equimolarities in intergenic regions and further confirmed the existence of replication-associated strand asymmetries [98–100]. In this manuscript, we will review recent results obtained when using the WT microscope to explore the scale invariance properties of the TA and GC skew profiles in the 22 human autosomes [98, 100]. These results will enlighten the richness of informations that can be extracted from these functional codings of DNA sequences including the prediction of 1012 putative human replication origins. In particular, this study will reveal a remarkable human gene organization driven by the coordination of transcription and replication [101].

## III. A WAVELET-BASED MULTIFRACTAL FORMALISM: THE WAVELET TRANSFORM MODULUS MAXIMA METHOD

### A. The continuous wavelet transform

The WT is a space-scale analysis which consists in expanding signals in terms of *wavelets* which are constructed from a single function, the *analyzing wavelet* $\psi$, by means of translations and dilations. The WT of a real-valued function $f$ is defined as [1, 2]:

$$T_\psi[f](x_0,a) = \frac{1}{a}\int_{-\infty}^{+\infty} f(x)\psi(\frac{x-x_0}{a})dx \ , \tag{1}$$

where $x_0$ is the space parameter and $a$ ($>0$) the scale parameter. The analyzing wavelet $\psi$ is generally chosen to be well localized in both space and frequency. Usually $\psi$ is required to be of zero mean for the WT to be invertible. But for the particular purpose of singularity tracking that is of interest here, we will further require $\psi$ to be orthogonal to low-order polynomials [7–16]:

$$\int_{-\infty}^{+\infty} x^m\psi(x)dx \ , \quad 0 \leq m < n_\psi \ . \tag{2}$$

As originally pointed out by Mallat and collaborators [10, 11], for the specific purpose of analyzing the regularity of a function, one can get rid of the redundancy of the WT by concentrating on the WT skeleton defined by its modulus maxima only. These maxima are defined, at each scale $a$, as the local maxima of $|T_\psi[f](x,a)|$ considered as a function of $x$. As illustrated in Fig. 2(e,f), these WTMM are disposed on connected curves in the space-scale (or time-scale) half-plane, called *maxima lines*. Let us define $\mathcal{L}(a_0)$ as the set of all the maxima lines that exist at the scale $a_0$ and which contain maxima at any scale $a \leq a_0$. An important feature of these maxima lines, when analyzing singular functions, is that there is at least one maxima line pointing towards each singularity [10, 11, 16].

### B. Scanning singularities with the wavelet transform modulus maxima

The strength of the singularity of a function $f$ at point $x_0$ is given by the *Hölder* exponent, *i.e.*, the largest exponent such that there exists a polynomial $P_n(x-x_0)$ of order $n < h(x_0)$ and a constant $C > 0$, so that for any point $x$ in a neighborhood of $x_0$, one has [7–11, 13, 16]:

$$|f(x) - P_n(x-x_0)| \leq C|x-x_0|^h \ . \tag{3}$$

If $f$ is $n$ times continuously differentiable at the point $x_0$, then one can use for the polynomial $P_n(x - x_0)$, the order-$n$ Taylor series of $f$ at $x_0$ and thus prove that $h(x_0) > n$. Thus $h(x_0)$ measures how irregular the function $f$ is at the point $x_0$. The higher the exponent $h(x_0)$, the more regular the function $f$.

The main interest in using the WT for analyzing the regularity of a function lies in its ability to be blind to polynomial behavior by an appropriate choice of the analyzing wavelet $\psi$. Indeed, let us assume that according to Eq. (3), $f$ has, at the point $x_0$, a local scaling (Hölder) exponent $h(x_0)$; then, assuming that the singularity is not oscillating [11, 102, 103], one can easily prove that the local behavior of $f$ is mirrored by the WT which locally behaves like [7–18] :

$$T_\psi[f](x_0, a) \sim a^{h(x_0)} \, , a \to 0^+ \, , \tag{4}$$

provided $n_\psi > h(x_0)$, where $n_\psi$ is the number of vanishing moments of $\psi$ (Eq. (2)). Therefore one can extract the exponent $h(x_0)$ as the slope of a log-log plot of the WT amplitude versus the scale $a$. On the contrary, if one chooses $n_\psi < h(x_0)$, the WT still behaves as a power-law but with a scaling exponent which is $n_\psi$ :

$$T_\psi[f](x_0, a) \sim a^{n_\psi} \, , a \to 0^+ \, . \tag{5}$$

Thus, around a given point $x_0$, the faster the WT decreases when the scale goes to zero, the more regular $f$ is around that point. In particular, if $f \in C^\infty$ at $x_0$ ($h(x_0) = +\infty$), then the WT scaling exponent is given by $n_\psi$, *i.e.* a value which is dependent on the shape of the analyzing wavelet. According to this observation, one can hope to detect the points where $f$ is smooth by just checking the scaling behavior of the WT when increasing the order $n_\psi$ of the analyzing wavelet [12–16].

**Remark 1**

A very important point (at least for practical purpose) raised by Mallat and Hwang [10] is that the local scaling exponent $h(x_0)$ can be equally estimated by looking at the value of the WT modulus along a maxima line converging towards the point $x_0$. Indeed one can prove that both Eqs. (4) and (5) still hold when following a maxima line from large down to small scales [10, 11].

### C. A wavelet-based multifractal formalism : the wavelet transform modulus maxima method

As originally defined by Parisi & Frisch [104], the multifractal formalism of multi-affine functions amounts to compute the so-called *singularity spectrum* $D(h)$ defined as the Hausdorff dimension of the set where the Hölder exponent is equal to $h$ [12, 13, 16]:

$$D(h) = dim_H\{x \ , \ h(x) = h\} \ , \tag{6}$$

where $h$ can take, *a priori*, positive as well as negative real values (*e.g.*, the Dirac distribution $\delta(x)$ corresponds to the Hölder exponent $h(0) = -1$) [17].

A natural way of performing a multifractal analysis of fractal functions consists in generalizing the "classical" multifractal formalism [105–109] using wavelets instead of boxes. By taking advantage of the freedom in the choice of the "generalized oscillating boxes" that are the wavelets, one can hope to get rid of possible smooth behavior that could mask singularities or perturb the estimation of their strength $h$. But the major difficulty with respect to box-counting techniques [48, 106, 110–112] for singular measures, consists in defining a covering of the support of the singular part of the function with our set of wavelets of different sizes. As emphasized in Refs. [12–16], the branching structure of the WT skeletons of fractal functions in the $(x, a)$ half-plane enlightens the hierarchical organization of their singularities (Figs. 2(e,f)). The WT skeleton can thus be used as a guide to position, at a considered scale $a$, the oscillating boxes in order to obtain a partition of the singularities of $f$. The wavelet transform modulus maxima (WTMM) method amounts to compute the following partition function in terms of WTMM coefficients [12–16]:

$$Z(q, a) = \sum_{l \in \mathcal{L}(a)} \left( \sup_{\substack{(x,a') \in l \\ a' \leq a}} |T_\psi[f](x, a')| \right)^q \ , \tag{7}$$

where $q \in \mathbb{R}$ and the sup can be regarded as a way to define a scale adaptative "Hausdorff-like" partition. Now from the deep analogy that links the multifractal formalism to thermodynamics [12, 113], one can define the exponent $\tau(q)$ from the power-law behavior of the partition function:

$$Z(q, a) \ \sim \ a^{\tau(q)} \ , \ \ a \to 0^+ \ , \tag{8}$$

where $q$ and $\tau(q)$ play respectively the role of the inverse temperature and the free energy. The main result of this wavelet-based multifractal formalism is that in place of the energy

and the entropy (*i.e.* the variables conjugated to $q$ and $\tau$), one has $h$, the Hölder exponent, and $D(h)$, the singularity spectrum. This means that the singularity spectrum of $f$ can be determined from the Legendre transform of the partition function scaling exponent $\tau(q)$ [13, 17, 18]:

$$D(h) = \min_q (qh - \tau(q)) \; . \tag{9}$$

From the properties of the Legendre transform, it is easy to see that *homogeneous* fractal functions that involve singularities of unique Hölder exponent $h = \partial\tau/\partial q$, are characterized by a $\tau(q)$ spectrum which is a *linear* function of $q$. On the contrary, a *nonlinear* $\tau(q)$ curve is the signature of nonhomogeneous functions that exhibit *multifractal* properties, in the sense that the Hölder exponent $h(x)$ is a fluctuating quantity that depends upon the spatial position $x$.

### D. Defining our battery of analyzing wavelets

There are almost as many analyzing wavelets as applications of the continuous WT [3–5, 12–16]. In the present work, we will mainly used the class of analyzing wavelets defined by the successive derivatives of the Gaussian function:

$$g^{(N)}(x) = \frac{d^N}{dx^N} e^{-x^2/2} \; , \tag{10}$$

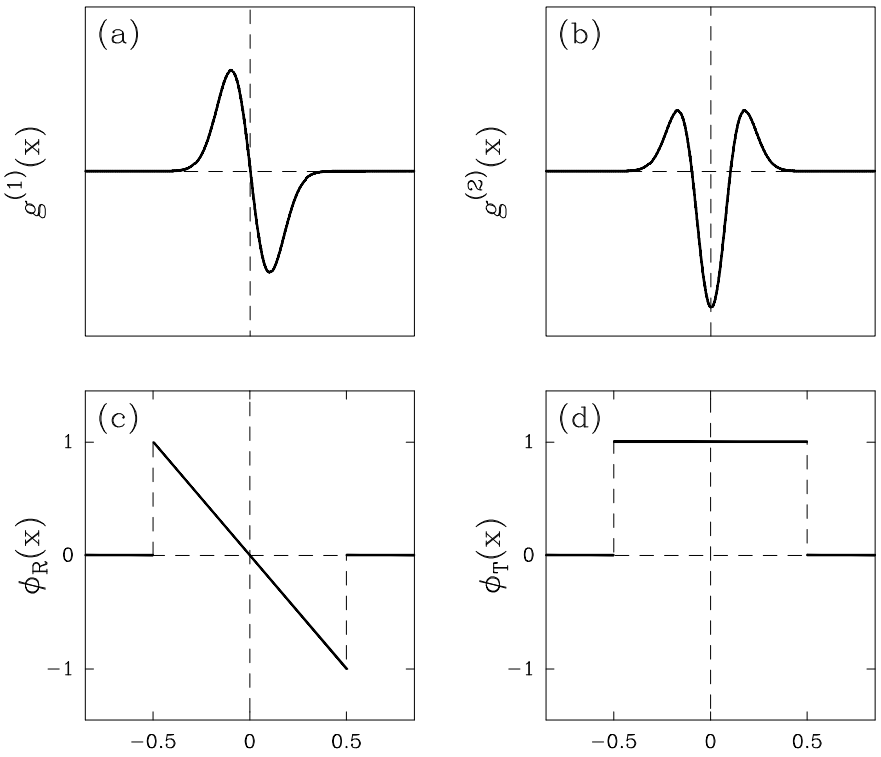
for which $n_\psi = N$ and more specifically $g^{(1)}$ and $g^{(2)}$ that are illustrated in Fig. 1(a,b).

**Remark 2**

The WT of a signal $f$ with $g^{(N)}$ (Eq. (10)) takes the following simple expression:

$$\begin{aligned} T_{g^{(N)}}[f](x,a) &= \frac{1}{a} \int_{-\infty}^{+\infty} f(y) g^{(N)} \left( \frac{y-x}{a} \right) dy, \\ &= a^N \frac{d^N}{dx^N} T_{g^{(0)}}[f](x,a). \end{aligned} \tag{11}$$

Equation (11) shows that the WT computed with $g^{(N)}$ at scale $a$ is nothing but the Nth derivative of the signal $f(x)$ smoothed by a dilated version $g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique [20].




**Fractals and Wavelets. Figure 1**: Set of analyzing wavelets $\psi(x)$ that can be used in Eq. (1). (a) $g^{(1)}$ and (b) $g^{(2)}$ as defined in Eq. (10). (c) $\phi_R$ as defined in Eq. (12), that will be used in Sect. VI to detect replication domains. (d) Box function $\phi_T$ that will be used in Sect. VI to model step-like skew profiles induced by transcription.

With the specific goal of disentangling the contributions to the nucleotide composition strand asymmetry coming respectively from transcription and replication processes, we will use in Sect. VI, an adapted analyzing wavelet of the following form (Fig. 1(c)) [101, 114]:

$$\begin{aligned} \phi_R(x) &= -(x-1/2), \qquad \text{for } x \in [-1/2, 1/2] \\ &= 0 \qquad \text{elsewhere.} \end{aligned} \tag{12}$$

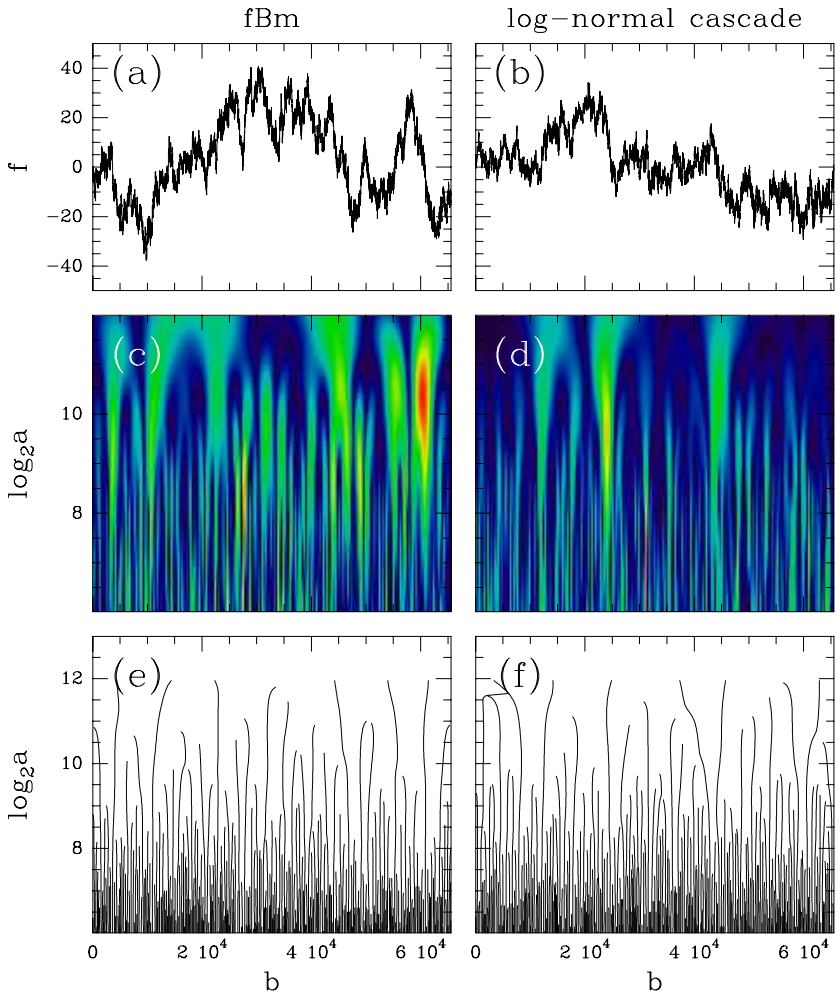
By performing multi-scale pattern recognition in the (space, scale) half-plane with this analyzing wavelet, we will be able to define replication domains bordered by putative replication origins in the human genome and more generally in mammalian genomes [101, 114].

### E. Test applications of the WTMM method on monofractal and multifractal synthetic random signals

This section is devoted to test applications of the WTMM method to random functions generated either by *additive* models like fractional Brownian motions [115] or by *multiplicative* models like random $\mathcal{W}$-cascades on wavelet dyadic trees [21, 22, 116, 117]. For each model, we first wavelet transform 1000 realizations of length $L = 65536$ with the first order ($n_\psi = 1$) analyzing wavelet $g^{(1)}$. From the WT skeletons defined by the WTMM, we compute the mean partition function (Eq. (7)) from which we extract the annealed $\tau(q)$ (Eq. (8)) and, in turn, $D(h)$ (Eq. (9)) multifractal spectra. We systematically test the robustness of our estimates with respect to some change of the shape of the analyzing wavelet, in particular when increasing the number $n_\psi$ of zero moments, going from $g^{(1)}$ to $g^{(2)}$ (Eq. (10)).

*Fractional Brownian signals*

Since its introduction by Mandelbrot and Van Ness [115], the fractional Brownian motion (fBm) $B_H$ has become a very popular model in signal and image processing [16, 20, 39]. In 1D, fBm has proved useful for modeling various physical phenomena with long-range dependence, *e.g.*, "$1/f$" noises. The fBm exhibits a power spectral density $S(k) \sim 1/k^\beta$, where the spectral exponent $\beta = 2H + 1$ is related to the Hurst exponent $H$. fBm has been extensively used as test stochastic signals for Hurst exponent measurements. In Figs 2, 3 and 4, we report the results of a statistical analysis of fBm's using the WTMM method [12–16]. We mainly concentrate on $B_{1/3}$ since it has a $k^{-5/3}$ power-spectrum similar to the spectrum of the multifractal stochastic signal we will study next. Actually, our goal is to demonstrate that, where the power spectrum analysis fails, the WTMM method succeeds in discriminating unambiguously between these two fractal signals. The numerical signals were generated by filtering uniformly generated pseudo-random noise in Fourier space in order to have the required $k^{-5/3}$ spectral density. A $B_{1/3}$ fractional Brownian trail is shown in Fig. 2(a). Fig. 2(c) illustrates the WT coded, independently at each scale $a$, using 256 colors. The analyzing wavelet is $g^{(1)}$ ($n_\psi = 1$). Fig. 3(a) displays some plots of $\log_2 Z(q, a)$ versus $\log_2(a)$ for different values of $q$, where the partition function $Z(q, a)$ has been computed on the WTMM skeleton shown in Fig. 2(e), according to the definition (Eq. (7)). Using a linear regression fit, we then obtain the slopes $\tau(q)$ of these graphs. As shown in Fig. 3(c), when

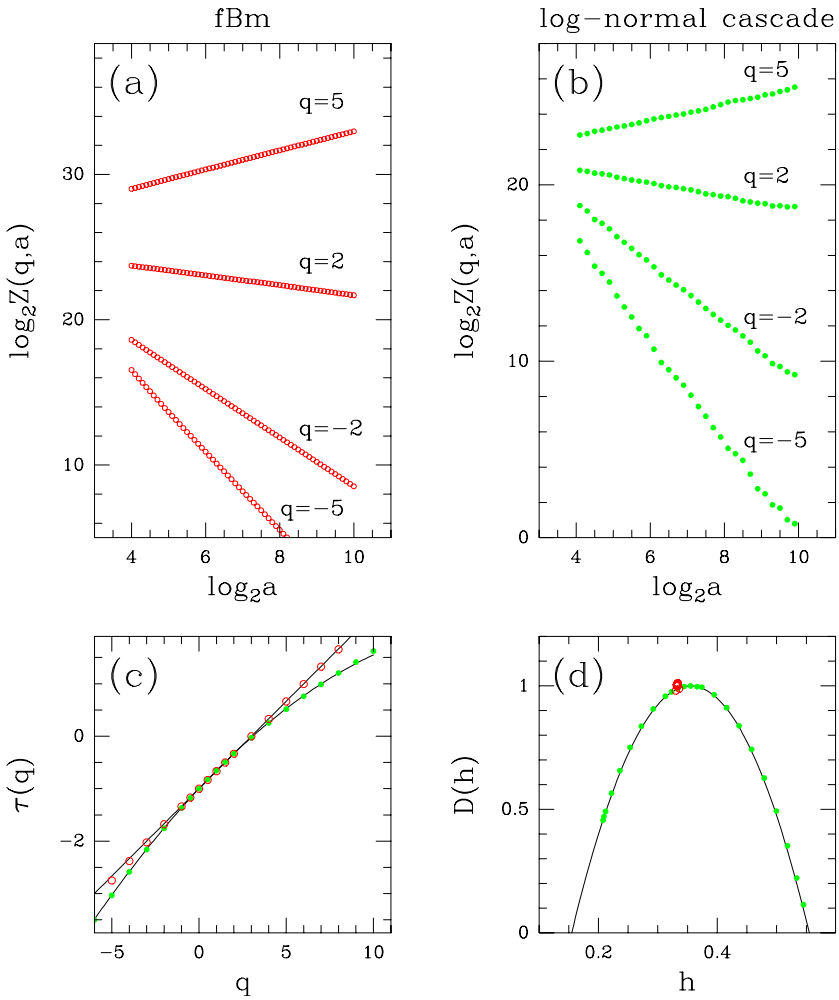



**Fractals and Wavelets. Figure 2**: WT of monofractal and multifractal stochastic signals. *Fractionnal Brownian motion:* (a) a realization of $B_{1/3}$ ($L = 65536$); (c) WT of $B_{1/3}$ as coded, independently at each scale $a$, using 256 colors from black ($|T_\psi| = 0$) to red ($\max_b |T_\psi|$); (e) WT skeleton defined by the set of all the maxima lines. *Log-normal random $\mathcal{W}$-cascades:* (b) a realization of the log-normal $\mathcal{W}$-cascade model ($L = 65536$) with the following parameter values $m = -0.355 \ln 2$ and $\sigma^2 = 0.02 \ln 2$ (see Ref. [116]); (d) WT of the realization in (b) represented with the same color coding as in (c); (f) WT skeleton. The analyzing wavelet is $g^{(1)}$ (see Fig. 1(a)).

plotted versus $q$, the data for the exponents $\tau(q)$ consistently fall on a straight line that is remarkably fitted by the theoretical prediction:

$$\tau(q) = qH - 1 \ , \tag{13}$$

with $H = 1/3$. From the Legendre transform of this linear $\tau(q)$ (Eq. (9)), one gets a $D(h)$




**Fractals and Wavelets. Figure 3**: Determination of the $\tau(q)$ and $D(h)$ multifractal spectra of fBm $B_{1/3}$ (red circles) and log-normal random $\mathcal{W}$-cascades (green dots) using the WTMM method. (a) $\log_2 Z(q,a)$ vs $\log_2 a$: $B_{1/3}$. (b) $\log_2 Z(q,a)$ vs $\log_2 a$: log-normal $\mathcal{W}$-cascades with the same parameters as in Fig. 2b. (c) $\tau(q)$ vs $q$; the solid lines correspond respectively to the theoretical spectra (13) and (16). (d) $D(h)$ vs $h$; the solid lines correspond respectively to the theoretical predictions (14) and (17). The analyzing wavelet is $g^{(1)}$. The reported results correspond to annealed averaging over 1000 realizations of $L = 65536$.

singularity spectrum that reduces to a single point:

$$D(h) = 1 \quad \text{if} \quad h = H \ , \tag{14}$$
$$= -\infty \quad \text{if} \quad h \neq H \ .$$

Thus, as expected theoretically [16, 115], one finds that the fBm $B_{1/3}$ is a nowhere differentiable homogeneous fractal signal with a unique Hölder exponent $h = H = 1/3$. Note that similar good estimates are obtained when using analyzing wavelets of different order ( *e.g.*

$g^{(2)}$), and this whatever the value of the index $H$ of the fBm [12–16].

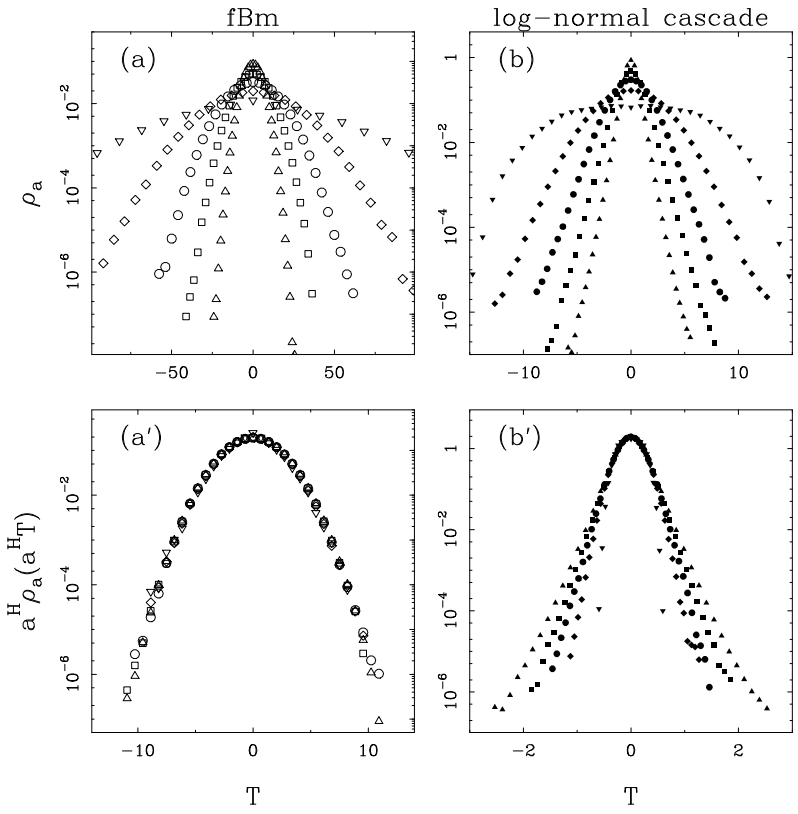
Within the perspective of confirming the monofractality of fBm's, we have studied the probability density function (pdf) of wavelet coefficient values $\rho_a(T_{g^{(1)}}(.,a))$, as computed at a fixed scale $a$ in the fractal scaling range. According to the monofractal scaling properties, one expects these pdfs to satisfy the self-similarity relationship [20, 27, 28]:

$$a^H \rho_a(a^H T) = \rho(T) \ , \tag{15}$$

where $\rho(T)$ is a "universal" pdf (actually the pdf obtained at scale $a = 1$) that does not depend on the scale parameter $a$. As shown in Fig. 4(a,a') for $B_{1/3}$, when plotting $a^H \rho_a(a^H T)$ vs $T$, all the $\rho_a$ curves corresponding to different scales (Fig. 4(a)) remarkably collapse on a unique curve when using a unique exponent $H = 1/3$ (Fig. 4(a')). Furthermore the so-obtained universal curve cannot be distinguished from a parabola in semi-log representation as the signature of the monofractal Gaussian statistics of fBm fluctuations [16, 20, 27].

*Random $\mathcal{W}$-cascades*

Multiplicative cascade models have enjoyed increasing interest in recent years as the paradigm of multifractal objects [16, 19, 48, 105, 107, 108, 118]. The notion of cascade actually refers to a self-similar process whose properties are defined multiplicatively from coarse to fine scales. In that respect, it occupies a central place in the statistical theory of turbulence [48, 104]. Originally, the concept of self-similar cascades was introduced to model multifractal measures (e.g. dissipation or enstrophy) [48]. It has been recently generalized to the construction of scale-invariant signals (e.g. longitudinal velocity, pressure, temperature) using orthogonal wavelet basis [116, 119]. Instead of redistributing the measure over sub-intervals with multiplicative weights, one allocates the wavelet coefficients in a multiplicative way on the dyadic grid. This method has been implemented to generate multifractal functions (with weights $W$) from a given deterministic or probalistic multiplicative process. Along the line of the modelling of fully developed turbulent signals by log-infinitely divisible multiplicative processes [120, 121], we will mainly concentrate here on the log-normal $\mathcal{W}$-cascades in order to calibrate the WTMM method. If $m$ and $\sigma^2$ are respectively the mean and the variance of $\ln W$ (where $W$ is a multiplicative random variable with log-normal probability distribution), then, as shown in Ref. [116], a straitghtforward




**Fractals and Wavelets. Figure 4**: Probability distribution functions of wavelet coefficient values of fBm $B_{1/3}$ (open symbols) and log-normal random $\mathcal{W}$-cascades (filled symbols) with the same parameters as in fig.2(b). (a) $\rho_a$ vs $T_{g^{(1)}}$ for the set of scales $a = 10$ (△), $50$ (□), $100$ (○), $1000$ (◊), $9000$ (▽); (a') $a^H\rho_a(\rho(a^H T_{g^{(1)}}))$ vs $T_{g^{(1)}}$ with $H = 1/3$; The symbols have the same meaning as in (a). (b) $\rho_a$ vs $T_{g^{(1)}}$ for the set of scales $a = 10$ (▲), $50$ (■), $100$ (●), $1000$ (◆), $9000$ (▼); (b') $a^H\rho_a(a^H T_{g^{(1)}})$ vs $T_{g^{(1)}}$ with $H = -m/\ln 2 = 0.355$. The analyzing wavelet is $g^{(1)}$ (Fig. 1(a)).

computation leads to the following $\tau(q)$ spectrum:

$$\begin{aligned} \tau(q) &= -\log_2 < W^q > -1 \ , \ \forall \ q \ \in \mathbb{R} \\ &= -\frac{\sigma^2}{2\ln 2}q^2 - \frac{m}{\ln 2}q - 1 \ , \end{aligned} \tag{16}$$

where $< ... >$ means ensemble average. The corresponding $D(h)$ singularity spectrum is obtained by Legendre transforming $\tau(q)$ (Eq. (9)):

$$D(h) = -\frac{(h + m/\ln 2)^2}{2\sigma^2/\ln 2} + 1 \ . \tag{17}$$

According to the convergence criteria established in Ref. [116], $m$ and $\sigma^2$ have to satisfy the conditions: $m < 0$ and $|m|/\sigma > \sqrt{2\ln 2}$. Moreover, by solving $D(h) = 0$, one gets the following bounds for the support of the $D(h)$ singularity spectrum: $h_{\min} = -\frac{m}{\ln 2} - \frac{\sqrt{2}\sigma}{\sqrt{\ln 2}}$ and $h_{\max} = -\frac{m}{\ln 2} + \frac{\sqrt{2}\sigma}{\sqrt{\ln 2}}$.

In Fig. 2(b) is illustrated a realization of a log-normal $\mathcal{W}$-cascade for the parameter values $m = -0.355\ln 2$ and $\sigma^2 = 0.02\ln 2$. The corresponding WT and WT skeleton as computed with $g^{(1)}$ are shown in Figs 2(d) and 2(f) respectively. The results of the application of the WTMM method are reported in Fig. 3. As shown in Fig. 3(b), when plotted versus the scale parameter $a$ in a logarithmic representation, the annealed average of the partition functions $Z(q, a)$ displays a well defined scaling behavior over a range of scales of about 5 octaves. Note that scaling of quite good quality is found for a rather wide range of $q$ values: $-5 \leq q \leq 10$. When processing to a linear regression fit of the data over the first four octaves, one gets the $\tau(q)$ spectrum shown in Fig. 3(c). This spectrum is clearly a nonlinear function of $q$, the hallmark of multifractal scaling. Moreover, the numerical data are in remarkable agreement with the theoretical quadratic prediction (Eq. (16)). Similar quantitative agreement is observed on the $D(h)$ singularity spectrum in Fig. 3(d) which displays a single humped parabola shape that characterizes intermittent fluctuations corresponding to Hölder exponents values ranging from $h_{\min} = 0.155$ to $h_{\max} = 0.555$. Unfortunately, to capture the strongest and the weakest singularities, one needs to compute the $\tau(q)$ spectrum for very large values of $|q|$. This requires the processing of many more realizations of the considered log-normal random $\mathcal{W}$-cascade. The multifractal nature of log-normal $\mathcal{W}$-cascade realizations is confirmed in Fig. 4(b,b') where the self-similarity relationship (Eq. (15)) is shown not to apply. Actually there does not exist a $H$ value allowing to superimpose onto a single curve the WT pdfs computed at different scales.

The test applications reported in this section demonstrate the ability of the WTMM method to resolve multifractal scaling of 1D signals, a hopeless task for classical power spectrum analysis. They were used on purpose to calibrate and to test the reliability of our methodology, and of the corresponding numerical tools, with respect to finite-size effects and statistical convergence.

## IV. BIFRACTALITY OF HUMAN DNA STRAND-ASYMMETRY PROFILES RESULTS FROM TRANSCRIPTION

During genome evolution, mutations do not occur at random as illustrated by the diversity of the nucleotide substitution rate values [122–125]. This non-randomness is considered as a by-product of the various DNA mutation and repair processes that can affect each of the two DNA strands differently. Asymmetries of substitution rates coupled to transcription have been mainly observed in prokaryotes [88, 89, 91], with only preliminary results in eukaryotes. In the human genome, excess of T was observed in a set of gene introns [126] and some large-scale asymmetry was observed in human sequences but they were attributed to replication [127]. Only recently, a comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G+T over A+C in the coding strand [95–97]. In contrast to the substitution biases observed in bacteria presenting 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-coupled repair mechanism acting on uncorrected substitution errors during replication [128]. In this section, we report the results of a genome-wide multifractal analysis of strand-asymmetry DNA walk profiles in the human genome [129]. This study is based on the computation of the TA and GC skews in non-overlapping 1 kbp windows:

$$S_{\mathrm{TA}} = \frac{n_{\mathrm{T}} - n_{\mathrm{A}}}{n_{\mathrm{T}} + n_{\mathrm{A}}}, \qquad S_{\mathrm{GC}} = \frac{n_{\mathrm{G}} - n_{\mathrm{C}}}{n_{\mathrm{G}} + n_{\mathrm{C}}}, \tag{18}$$

where $n_{\mathrm{A}}$, $n_{\mathrm{C}}$, $n_{\mathrm{G}}$ and $n_{\mathrm{T}}$ are respectively the numbers of A, C, G and T in the windows. Because of the observed correlation between the TA and GC skews, we also considered the total skew

$$S = S_{\mathrm{TA}} + S_{\mathrm{GC}}. \tag{19}$$

From the skews $S_{\rm TA}(n)$, $S_{\rm GC}(n)$ and $S(n)$, obtained along the sequences, where $n$ is the position (in kbp units) from the origin, we also computed the cumulative skew profiles (or skew walk profiles):

$$\Sigma_{\rm TA}(n) = \sum_{j=1}^{n} S_{\rm TA}(j), \qquad \Sigma_{\rm GC}(n) = \sum_{j=1}^{n} S_{\rm GC}(j), \tag{20}$$

and

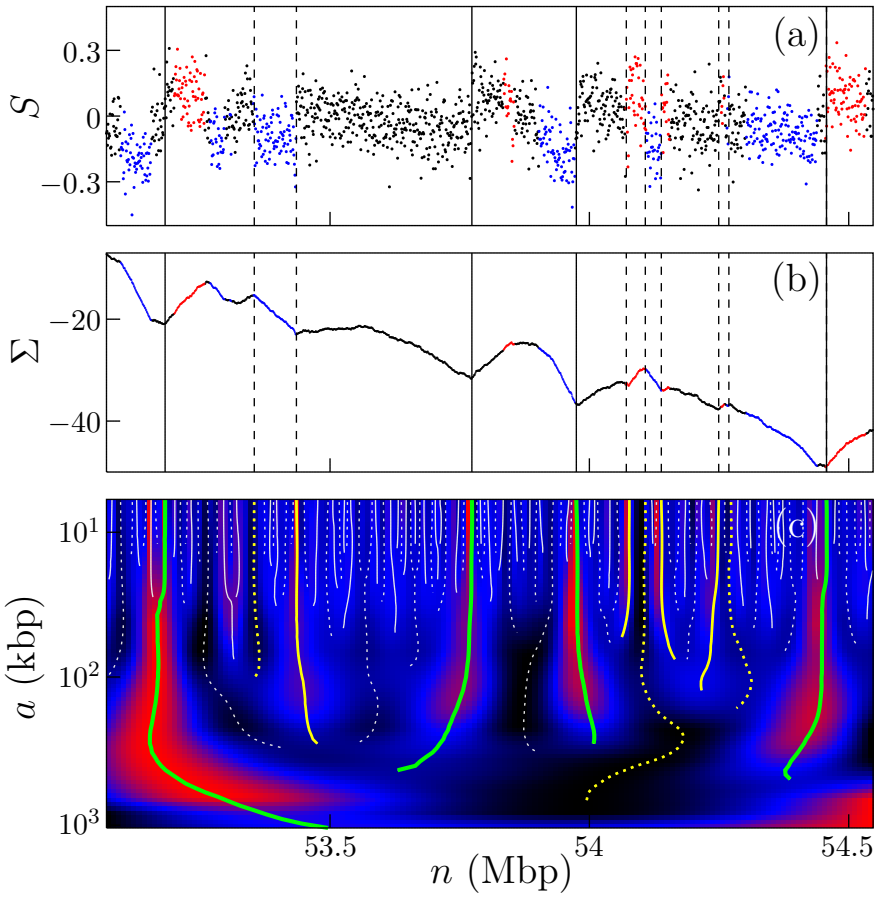
$$\Sigma(n) = \sum_{j=1}^{n} S(j). \tag{21}$$

Our goal is to show that the skew DNA walks of the 22 human autosomes display an unexpected (with respect to previous monofractal diagnosis [27–30]) bifractal scaling behavior in the range 10 to 40 kbp as the signature of the presence of transcription-induced jumps in the LRC noisy $S$ profiles. Sequences and gene annotation data ("refGene") were retrieved from the UCSC Genome Browser (May 2004). We used RepeatMasker to exclude repetitive elements that might have been inserted recently and would not reflect long-term evolutionary patterns.

### A. Revealing the bifractality of human skew DNA walks with the WTMM method

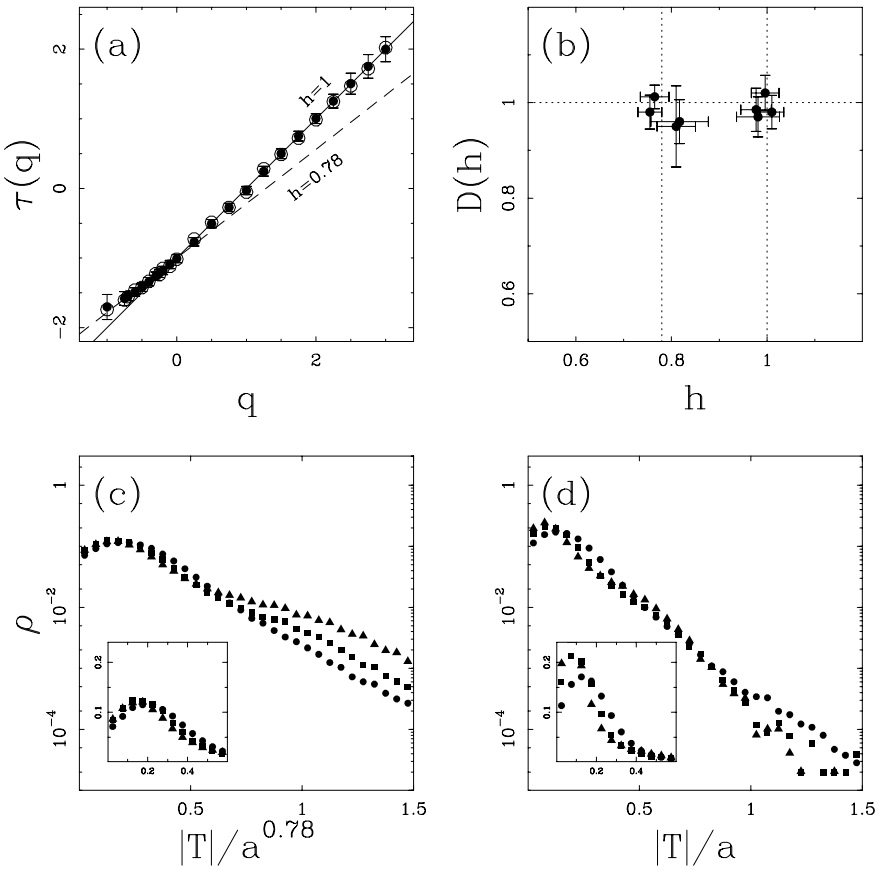
As an illustration of our wavelet-based methodology, we show in Fig. 5 the $S$ skew profile of a fragment of human chromosome 6 (Fig. 5(a)), the corresponding skew DNA walk (Fig. 5(b)) and its space-scale wavelet decomposition using the Mexican hat analyzing wavelet $g^{(2)}$ (Fig. 1(b)). When computing $Z(q,a)$ (Eq. (7)) from the WT skeletons of the skew DNA walks $\Sigma$ of the 22 human autosomes, we get convincing power-law behavior for $-1.5 \leq q \leq 3$ (data not shown). In Fig. 6(a) are reported the $\tau(q)$ exponents obtained using a linear regression fit of $\ln Z(q,a)$ vs $\ln a$ over the range of scales 10 kbp $\leq a \leq$ 40 kbp. All the data points remarkably fall on two straight lines $\tau_1(q) = 0.78q - 1$ and $\tau_2(q) = q - 1$ which strongly suggests the presence of two types of singularities $h_1 = 0.78$ and $h_2 = 1$, respectively on two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ with the same Haussdorf dimension $D = -\tau_1(0) = -\tau_2(0) = 1$, as confirmed when computing the $D(h)$ singularity spectrum in Fig. 6(b). This observation means that $Z(q,a)$ can be split in two parts [12, 16]:

$$Z(q,a) = C_1(q) a^{qh_1 - 1} + C_2(q) a^{qh_2 - 1} \ , \tag{22}$$




**Fractals and Wavelets. Figure 5**: (a) Skew profile $S(n)$ (Eq. (19)) of a repeat-masked fragment of human chromosome 6; red (resp. blue) 1 kbp window points correspond to (+) genes (resp. (−) genes) lying on the Watson (resp. Crick) strand; black points to intergenic regions. (b) Cumulated skew profile $\Sigma(n)$ (Eq. (21)). (c) WT of $\Sigma$; $T_{g^{(2)}}(n,a)$ is coded from black (min) to red (max); the WT skeleton defined by the maxima lines is shown in solid (resp. dashed) lines corresponding to positive (resp. negative) WT values. For illustration yellow solid (resp. dashed) maxima lines are shown to point to the positions of 2 upward (resp. 2 downward) jumps in $S$ (vertical dashed lines in (a) and (b)) that coincide with gene transcription starts (resp. ends). In green are shown maxima lines that persist above $a \geq 200$ kbp and that point to sharp upward jumps in $S$ (vertical solid lines in (a) and (b)) that are likely to be the locations of putative replication origins (see Sect. V) [98, 100]; note that 3 out of those 4 jumps are co-located with transcription start sites [129].

where $C_1(q)$ and $C_2(q)$ are prefactors that depend on $q$. Since $h_1 < h_2$, in the limit $a \mapsto 0^+$, the partition function is expected to behave like $Z(q,a) \sim C_1(q)a^{qh_1-1}$ for $q > 0$ and like $Z(q,a) \sim C_2(q)a^{qh_2-1}$ for $q < 0$, with a so-called phase transition [12, 16] at the critical value $q_c = 0$. Surprisingly, it is the contribution of the weakest singularities $h_2 = 1$ that controls the scaling behavior of $Z(q,a)$ for $q > 0$ while the strongest ones $h_1 = 0.78$ actually dominate




**Fractals and Wavelets. Figure 6**: Multifractal analysis of $\Sigma(n)$ of the 22 human (filled symbols) and 19 mouse (open circle) autosomes using the WTMM method with $g^{(2)}$ over the range 10 kbp $\leq a \leq$ 40 kbp [129]. (a) $\tau(q)$ vs $q$. (b) $D(h)$ vs $h$. (c) WTMM pdf: $\rho$ is plotted versus $|T|/a^H$ where $H = h_1 = 0.78$, in semi-log representation; the inset is an enlargement of the pdf central part in linear representation. (d) Same as in (c) but with $H = h_2 = 1$. In (c) and (d), the symbols correspond to scales $a = 10$ (•), 20 (■) and 40 (▲) kbp.

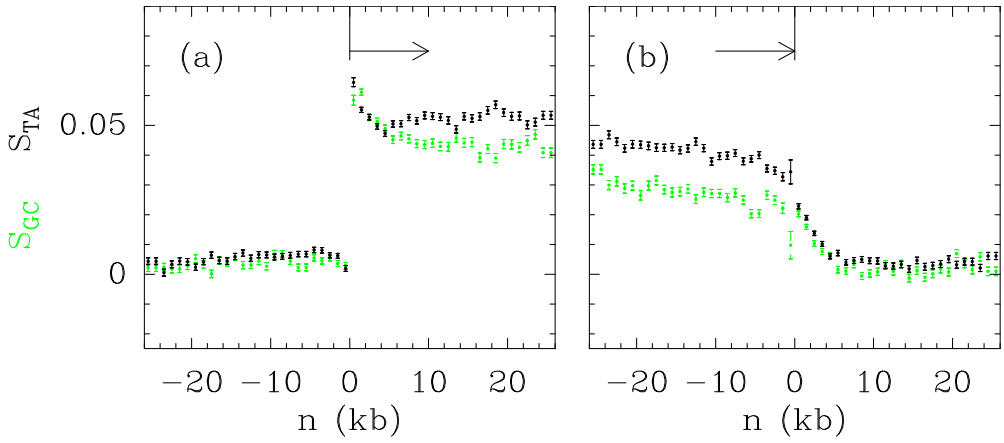
for $q < 0$ (Fig. 6(a)). This inverted behavior originates from finite (1 kbp) resolution which prevents the observation of the predicted scaling behavior in the limit $a \mapsto 0^+$. The prefactors $C_1(q)$ and $C_2(q)$ in Eq. (22) are sensitive to (i) the number of maxima lines in the WT skeleton along which the WTMM behave as $a^{h_1}$ or $a^{h_2}$ and (ii) the relative amplitude of these WTMM. Over the range of scales used to estimate $\tau(q)$, the WTMM along the maxima lines pointing (at small scale) to $h_2 = 1$ singularities are significantly larger than those along the maxima lines associated to $h_1 = 0.78$ (see Figs. 6(c,d)). This implies that the larger $q > 0$, the stronger the inequality $C_2(q) \gg C_1(q)$ and the more pronounced the relative contribution of the second term in the r.h.s. of Eq. (22). On the opposite for $q < 0$,

$C_1(q) \gg C_2(q)$ which explains that the strongest singularities $h_1 = 0.78$ now control the scaling behavior of $Z(q, a)$ over the explored range of scales.

In Figs. 6(c,d) are shown the WTMM pdfs computed at scales $a$ =10, 20 and 40 kbp after rescaling by $a^{h_1}$ and $a^{h_2}$ respectively. We note that there does not exist a value of $H$ such that all the pdfs collapse on a single curve as expected from Eq. (15) for monofractal DNA walks. Consistently with the $\tau(q)$ data in Fig. 6(a) and with the inverted scaling behavior discussed above, when using the two exponents $h_1 = 0.78$ and $h_2 = 1$, one succeeds in superimposing respectively the central (bump) part (Fig. 6(c)) and the tail (Fig. 6(d)) of the rescaled WTMM pdfs. This corroborates the bifractal nature of the skew DNA walks that display two competing scale-invariant components of Hölder exponents: (i) $h_1 = 0.78$ corresponds to LRC homogeneous fluctuations previously observed over the range 200 bp $\lesssim a \lesssim$ 20 kbp in DNA walks generated with structural codings [29, 30] and (ii) $h_2 = 1$ is associated to convex $\vee$ and concave $\wedge$ shapes in the DNA walks $\Sigma$ indicating the presence of discontinuities in the derivative of $\Sigma$, *i.e.*, of jumps in $S$ (Fig. 5(a,b)). At a given scale $a$, according to Eq. (11), a large value of the WTMM in Fig. 5(c) corresponds to a strong derivative of the smoothed $S$ profile and the maxima line to which it belongs is likely to point to a jump location in $S$. This is particularly the case for the colored maxima lines in Fig. 5(c): upward (resp. downward) jumps (Fig. 5(a)) are so-identified by the maxima lines corresponding to positive (resp. negative) values of the WT.

### B. Transcription-induced step-like skew profiles in the human genome

In order to identify the origin of the jumps observed in the skew profiles, we have performed a systematic investigation of the skews observed along 14854 intron containing genes [96, 97]. In Fig. 7 are reported the mean values of $S_{\rm TA}$ and $S_{\rm GC}$ skews for all genes as a function of the distance to the 5'- or 3'- end. At the 5' gene extremities (Fig. 7(a)), a sharp transition of both skews is observed from about zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for $\bar{S}_{\rm TA}$ and between 3 and 5% for $\bar{S}_{\rm GC}$. At the gene 3'- extremities (Fig. 7(b)), the TA and GC skews also exhibit transitions from significantly large values in transcribed regions to very small values in untranscribed regions. However, in comparison to the steep transitions observed at 5'- ends, the 3'- end profiles present a slightly smoother transition pattern extending over $\sim$5 kbp

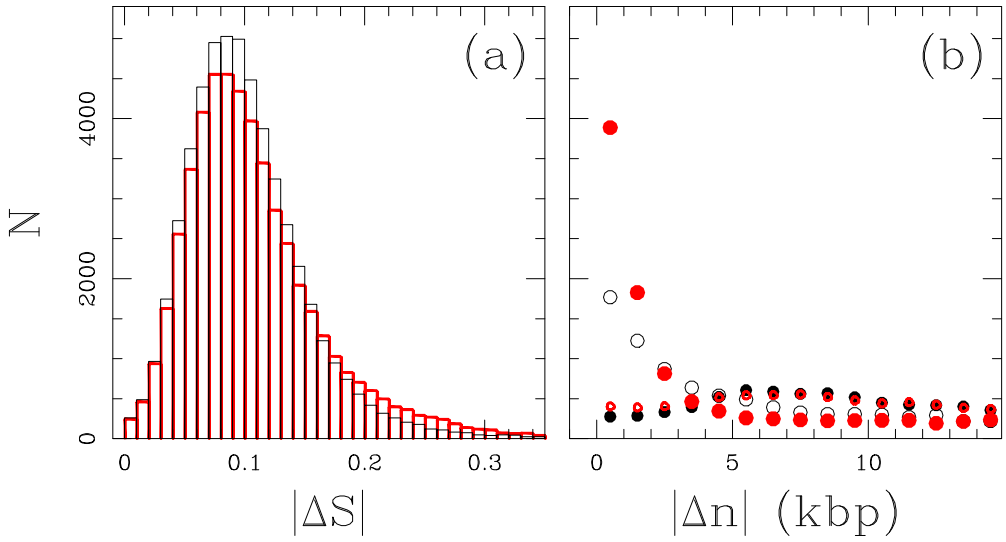



**Fractals and Wavelets. Figure 7**: TA (•) and GC (•) skew profiles in the regions surrounding 5' and 3' gene extremities [96]. $S_{\rm TA}$ and $S_{\rm GC}$ were calculated in 1 kbp windows starting from each gene extremities in both directions. In abscissa is reported the distance (n) of each 1 kbp 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 our set of 14854 intron-containing genes for all 1 kbp windows at the corresponding abscissa. Error bars represent the standard error of the means.

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 are expected to have mostly 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 the germ-line cells. Such differences would then lead to some broadening of the skew profiles.

### C. From skew multifractal analysis to gene detection

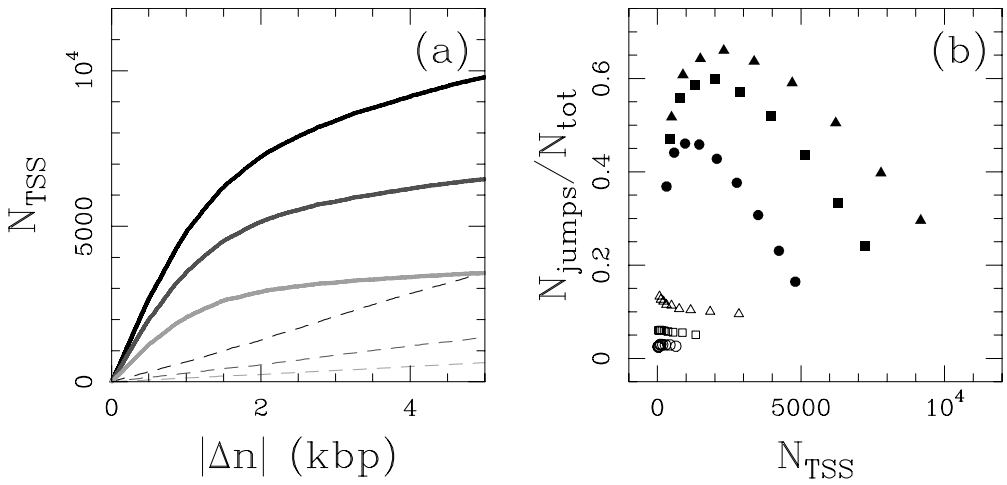
In Fig. 8 are reported the results of a statistical analysis of the jump amplitudes in human $S$ profiles [129]. For maxima lines that extend above $a^*$ =10 kbp in the WT skeleton (see Fig. 5(c)), the histograms obtained for upward and downward variations are quite similar, especially their tails that are likely to correspond to jumps in the $S$ profiles (Fig. 8(a)). When computing the distance between upward or downward jumps ($|\Delta S| \geq 0.1$) to the closest transcription start (TSS) or end (TES) sites (Fig. 8(b)), we reveal that the number of upward jumps in close proximity ($|\Delta n| \lesssim 3$ kpb) to TSS over-exceeds the number of such jumps close to TES. Similarly, downward jumps are preferentially located at TES. These observations




**Fractals and Wavelets. Figure 8**: Statistical analysis of skew variations at the singularity positions determined at scale 1 kbp from the maxima lines that exist at scales $a \geq 10$ kbp in the WT skeletons of the 22 human autosomes [129]. For each singularity, we computed the variation amplitudes $\Delta S = \bar{S}(3') - \bar{S}(5')$ over two adjacent 5 kbp windows, respectively in the $3'$ and $5'$ directions and the distances $\Delta n$ to the closest TSS (resp. TES). (a) Histograms $N(|\Delta S|)$ for upward ($\Delta S > 0$, red) and downward ($\Delta S < 0$, black) skew variations. (b) Histograms of the distances $\Delta n$ of upward (red) or downward (black) jumps with $|\Delta S| \geq 0.1$ to the closest TSS (•, •) and TES (○, ○).

are consistent with the step-like shape of skew profiles induced by transcription: $S > 0$ (resp. $S < 0$) is constant along a (+) (resp. (−)) gene and $S = 0$ in the intergenic regions (Fig. 7) [96]. Since a step-like 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 even number of $\Delta S > 0$ and $\Delta S < 0$ jumps when exploring the range of scales $10 \lesssim a \lesssim 40$ kbp, typical of human gene size. Note that in Fig. 8(a), the number of sharp upward jumps actually slightly exceeds the number of sharp downward jumps, consistently with the experimental observation that whereas TSS are well defined, TES may extend over 5 kbp resulting in smoother downward skew transitions (Fig.7(b)). This TES particularity also explains the excess of upward jumps found close to TSS as compared to the number of downward jumps close to TES (Fig. 8(b)).

In Fig. 9(a), we report the analysis of the distance of TSS to the closest upward jump [129]. For a given upward jump amplitude, the number of TSS with a jump within $|\Delta n|$ increases faster than expected (as compared to the number found for randomized jump positions) up to $|\Delta n| \simeq 2$ kbp. This indicates that the probabilty to find an upward jump within




**Fractals and Wavelets. Figure 9**: (a) Number of TSS with an upward jump within $|\Delta n|$ (abscissa) for jump amplitudes $\Delta S > 0.1$ (black), 0.15 (dark grey) and 0.2 (light grey). Solid lines correspond to true jump positions while dashed lines to the same analysis when jump positions were randomly drawn along each chromosome [129]. (b) Among the $N_{\rm tot}(\Delta S^*)$ upward jumps of amplitude larger than some threshold $\Delta S^*$, we plot the proportion of those that are found within 1 kbp (•), 2 kbp (■) or 4 kbp (▲) of the closest TSS *vs* the number $N_{\rm TSS}$ of the so-delineated TSS. Curves were obtained by varying $\Delta S^*$ from 0.1 to 0.3 (from right to left). Open symbols correspond to similar analyses performed on random upward jump and TSS positions.

a gene promoter region is significantly larger than elsewhere. For example, out of 20023 TSS, 36% (7228) are delineated within 2 kbp by a jump with $\Delta S > 0.1$. This provides a very reasonable estimate for the number of genes expressed in germline cells as compared to the 31.9% recently experimentally found to be bound to Pol II in human embryonic stem cells [130].

Combining the previous results presented in Figs. 8(b) and 9(a), we report in Fig. 9(b) an estimate of the efficiency/coverage relationship by plotting the proportion of upward jumps ($\Delta S > \Delta S^*$) lying in TSS proximity as a function of the number of so-delineated TSS [129]. For a given proximity threshold $|\Delta n|$, increasing $\Delta S^*$ results in a decrease of the number of delineated TSS, characteristic of the right tail of the gene bias pdf. Concomitant to this decrease, we observe an increase of the efficiency up to a maximal value corresponding to some optimal value for $\Delta S^*$. For $|\Delta n| < 2$ kbp, we reach a maximal efficiency of 60% for $\Delta S^* = 0.225$; 1403 out of 2342 upward jumps delineate a TSS. Given the fact that the actual number of human genes is estimated to be significantly larger ($\sim 30000$) than the number provided by refGene, a large part of the the 40% (939) of upward jumps that have not been

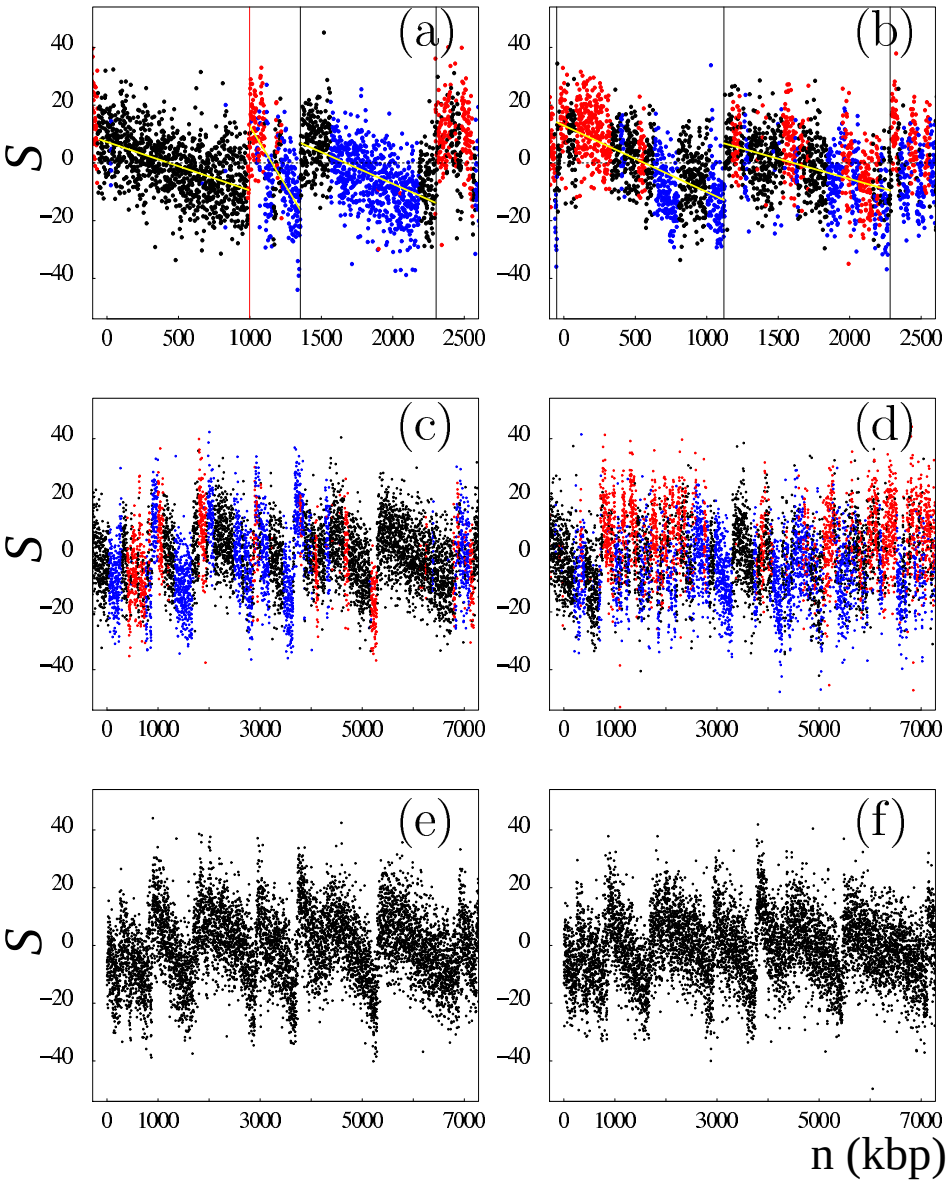
associated to a refGene could be explained by this limited coverage. In other words, jumps with sufficiently high amplitude are very good candidates for the location of highly-biased gene promoters. Let us point that out of the above 1403 (resp. 2342) upward jumps, 496 (resp. 624) jumps are still observed at scale $a^* = 200$ kbp. We will see in the next section that these jumps are likely to also correspond to replication origins underlying the fact that large upward jumps actually result from the cooperative contributions of both transcription- and replication- associated biases [98–101]. The observation that 80% (496/624) of the predicted replication origins are co-located with TSS enlightens the existence of a remarkable gene organisation at replication origins [101].

To summarize, we have demonstrated the bifractal character of skew DNA walks in the human genome. When using the WT microscope to explore (repeat-masked) scales ranging from 10 to 40 kbp, we have identified two competing homogeneous scale-invariant components characterized by Hölder exponents $h_1 = 0.78$ and $h_2 = 1$ that respectively correspond to LRC colored noise and sharp jumps in the original DNA compositional asymmetry profiles. Remarkably, the so-identified upward (resp. downward) jumps are mainly found at the TSS (resp. TES) of human genes with high transcription bias and thus very likely higly expressed. As illustrated in Fig. 6(a), similar bifractal properties are also observed when investigating the 19 mouse autosomes. This suggests that the results reported in this section are general features of mammalian genomes [129].

## V. FROM THE DETECTION OF RELICATION ORIGINS USING THE WAVELET TRANSFORM MICROSCOPE TO THE MODELING OF REPLICATION 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 [131], 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 initiate 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* [132]. 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 [133–137]. 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 [133–135, 137]. Actually, sequence requirements for a replication origin vary significantly between different eukaryotic organisms. In the unicellular eukaryote *Saccharomyces cerevisiae*, the replication origins spread over 100-150 bp and present some highly conserved motifs [132]. However, among eukaryotes, *S. cerevisiae* seems to be the exception that remains faithful to the replicon model. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp [132]. 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 [138]. 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 [136, 139, 140]. A developmental change occurs around midblastula transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites [136, 140]. 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 [141, 142]. As recently proposed by many authors [143–145], 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 only very few replication origins have been identified so far in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human [146]. Along the line of this epigenetic interpretation, one might wonder what can be learned about eukaryotic DNA replication from DNA sequence analysis.




**Fractals and Wavelets. Figure 10**: $S$ profiles along mammalian genome fragments [100, 146]. (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 (see Section V B); 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 syntenic to the human fragment shown in (c). In (e) and (f), genes are not represented.

### A. Replication induced factory-roof skew profiles in mammalian genomes

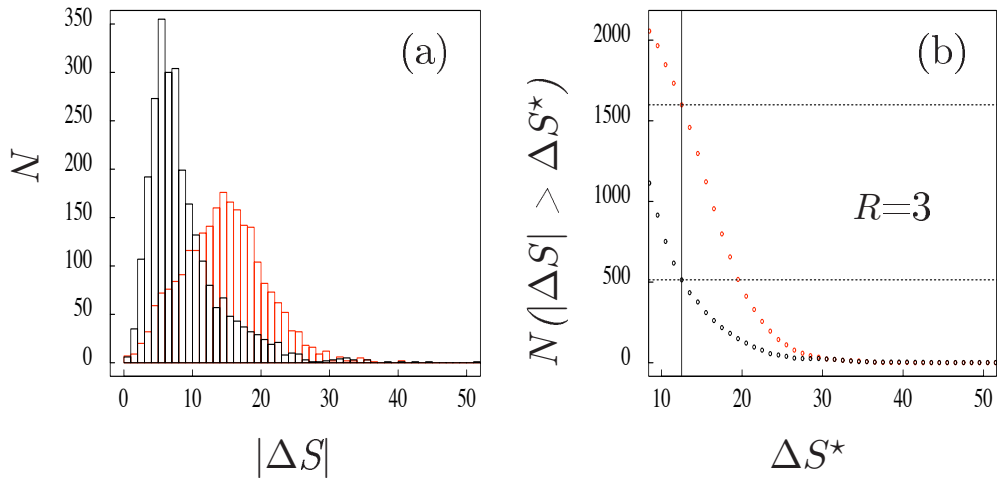
The existence of replication associated strand asymmetries has been mainly established in bacterial genomes [87, 90, 92–94]. $S_{\rm GC}$ and $S_{\rm TA}$ skews abruptly switch sign (over few kbp) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This step-like profile is characteristic of the replicon model [131] (see Fig. 13, left panel). 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 [92, 147, 148]. Other studies have led to rather opposite results. For instance, strand asymmetries associated with replication have been observed in the subtelomeric regions of *Saccharomyces cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism [149].

As shown in Fig. 10(a) for TOP1 replication origin [146], most of the known replication origins in the human genome correspond to rather sharp (over several kbp) transitions from negative to positive $S$ ($S_{\rm TA}$ as well as $S_{\rm GC}$) skew values that clearly emerge from the noisy background. But when examining the behavior of the skews at larger distances from the origin, one does not observe a step-like pattern with upward and downward jumps at the origin and termination positions respectively as expected for the bacterial replicon model (Fig. 13, left panel). Surprisingly, on both sides of the upward jump, the noisy S profile decreases steadily in the 5' to 3' direction without clear evidence of pronounced downward jumps. As shown in Fig. 10(b–d), sharp upward jumps of amplitude $\Delta S \gtrsim 15\%$, similar to the ones observed for the known replication origins (Fig. 10(a)), seem to exist also at many other locations along the human chromosomes. But 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" [98, 100, 146]. Note that the jagged S profiles shown in Fig. 10(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) mainly in GC rich regions (Fig. 10(d)), up to 1–2 Mbp ($\sim$ 2–3 Mbp for native sequences) (Fig. 10(c)) in agreement with recent experimental studies [150] that have shown that mammalian replicons are heterogeneous in size with an average size $\sim$ 500 kbp, the largest ones being as large as a few Mbp. But what is important to notice is that some of these segments between two successive skew upward jumps are entirely intergenic (Fig. 10(a,c)), clearly illustrating the particular profile of a strand bias resulting solely from replication [98, 100, 146]. In most other cases, we observe the superimposition of this replication profile and of the step-like profiles of (+) and (−) genes (Fig. 7), appearing as upward and downward blocks standing out from the replication pattern (Fig. 10(c)). Importantly, as illustrated in Fig. 10(e,f), the factory-roof pattern is not specific to human sequences but is also observed in numerous regions of the mouse and dog genomes [100]. Hence, the presence of strand asymmetry in regions that have strongly diverged during evolution further supports the existence of compostional bias associated with replication in mammalian germ-line cells [98, 100, 146].

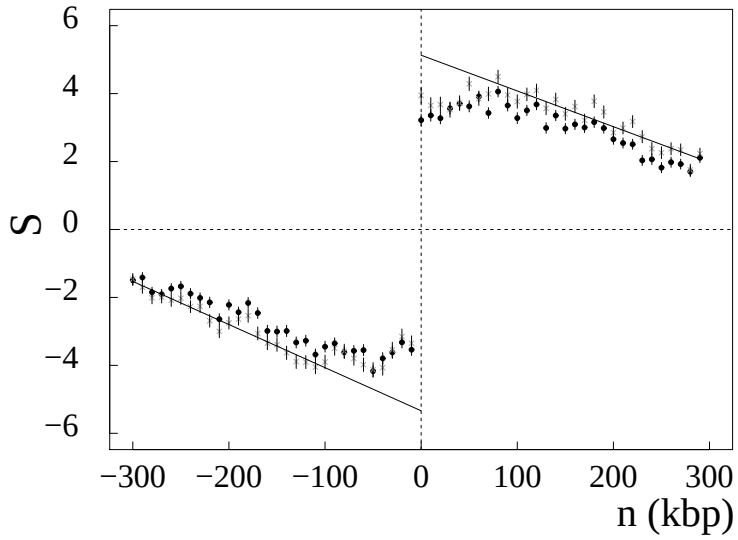
### B. Detecting replication origins from the skew WT skeleton

We have shown in Fig. 10(a) that experimentally determined human replication origins coincide with large-amplitude upward transitions in noisy skew profiles. The corresponding $\Delta S$ ranges between 14% and 38%, owing to possible different replication initiation efficiencies and/or different contributions of transcriptional biases (Sect. IV). Along the line of the jump detection methodology described in Sect. IV, we have checked that upward jumps observed in the skew S at these known replication origins correspond to maxima lines in the WT skeleton that extend to rather large scales $a > a^* = 200$ kbp. This observation has led us to select the maxima lines that exist above $a^* = 200$ kbp, i.e. a scale which is smaller than the typical replicon size and larger than the typical gene size [98, 100]. In this way, we not only reduce the effect of the noise but we also reduce the contribution of the upward (5' extremity) and backward (3' extremity) jumps associated to the step-like skew pattern induced by transcription only (Sect. IV), to the benefit of maintaining a good sensitivity to replication induced jumps. The detected jump locations are estimated as the positions at scale 20 kbp of the so-selected maxima lines. According to Eq. (11), upward




**Fractals and Wavelets. Figure 11**: Statistical analysis of the sharp jumps detected in the $S$ profiles of the 22 human autosomal chromosomes by the WT microscope at scale $a^* = 200$ kbp for repeat-masked sequences [98, 100]. $|\Delta S| = |\bar{S}(3') - \bar{S}(5')|$, where the averages were computed over the two adjacent 20 kbp windows, respectively, in the 3' and 5' direction from the detected jump location. (a) Histograms $N(|\Delta S|)$ of $|\Delta S|$ values. (b) $N(|\Delta S| > \Delta S^*)$ vs $\Delta S^*$. In (a) and (b), the black (resp. red) line corresponds to downward $\Delta S < 0$ (resp. upward $\Delta S > 0$) jumps. $R = 3$ corresponds to the ratio of upward over downward jumps presenting an amplitude $|\Delta S| \geq 12.5\%$ (see text).

(resp. downward) jumps are identified by the maxima lines corresponding to positive (resp. negative) values of the WT as illustrated in Fig. 5(c) by the green solid (resp. dashed) maxima lines. When applying this methodology to the total skew $S$ along the repeat-masked DNA sequences of the 22 human autosomal chromosomes, 2415 upward jumps are detected and, as expected, a similar number (namely 2686) of downward jumps. In Fig. 11(a) are reported the histograms of the amplitude $|\Delta S|$ of the so-identified upward ($\Delta S > 0$) and downward ($\Delta S < 0$) jumps respectively. These histograms no longer superimpose as previously observed at smaller scales in Fig. 8(a), the former being significantly shifted to larger $|\Delta S|$ values. When plotting $N(|\Delta S| > \Delta S^*)$ versus $\Delta S^*$ in Fig. 11(b), we can see that the number of large amplitude upward jumps overexceeds the number of large amplitude downward jumps. These results confirm that most of the sharp upward transitions in the S profiles in Fig. 10 have no sharp downward transition counterpart [98, 100]. This excess likely results from the fact that, contrasting with the prokaryote replicon model (Fig. 13, left panel) where downward jumps result from precisely positioned replication terminations, in mammals termination appears not to occur at specific positions but to be randomly distributed. Accordingly the small number of downward jumps with large $|\Delta S|$ is likely to




**Fractals and Wavelets. Figure 12**: Mean skew profile of intergenic regions around putative replication origins [100]. The skew $S$ was calculated in 1 kbp windows (Watson strand) around the position ($\pm$300 kbp without repeats) of the 1012 detected upward jumps; 5' and 3' transcript extremities were extended by 0.5 and 2 kbp, respectively ($\bullet$), or by 10 kbp at both ends ($*$). The abscissa represents the distance (in kbp) to the corresponding origin; the ordinate represents the skews calculated for the windows situated in intergenic regions (mean values for all discontinuities and for 10 consecutive 1 kbp window positions). The skews are given in percent (vertical bars, SEM). The lines correspond to linear fits of the values of the skew ($*$) for $n < -100$ kbp and $n > 100$ kbp.

result from transcription (Fig. 5) and not from replication. These jumps are probably due to highly biased genes that also generate a small number of large-amplitude upward jumps, giving rise to false-positive candidate replication origins. In that respect, the number of large downward jumps can be taken as an estimation of the number of false positives. In a first step, we have retained as acceptable a proportion of 33% of false positives. As shown in Fig. 11(b), this value results from the selection of upward and downward jumps of amplitude $|\Delta S| \geq 12.5\%$, corresponding to a ratio of upward over downward jumps $R = 3$. Let us notice that the value of this ratio is highly variable along the chromosome [146] and significantly larger than 1 for $G + C \lesssim 42\%$.

In a final step, we have decided [98, 100, 146] to retain as putative replication origins upward jumps with $|\Delta S| \geq 12.5\%$ detected in regions with G+C $\leq 42\%$. This selection leads to a set of 1012 candidates among which our estimate of the proportion of true replication origins is 79% ($R = 4.76$). In Fig. 12 is shown the mean skew profile calculated in intergenic

windows on both sides of the 1012 putative replication origins [100]. This mean skew profile presents a rather sharp transition from negative to positive values when crossing the origin position. To avoid any bias in the skew values that could result from incompletely annotated gene extremities (e.g. 5' and 3' UTRs), we have removed 10-kbp sequences at both ends of all annotated transcripts. As shown in Fig. 12, the removal of these intergenic sequences does not significantly modifies the mean skew profile, indicating that the observed values do not result from transcription. On both sides of the jump, we observe a linear decrease of the bias with some flattening of the profile close to the transition point. Note that, due to (i) the potential presence of signals implicated in replication initiation and (ii) the possible existence of dispersed origins [151], one might question the meaningfulness of this flattening that leads to a significant underestimate of the jump amplitude. Furthermore, according to our detection methodology, the numerical uncertainty on the putative origin position estimate may also contribute to this flattening. As illustrated in Fig. 12, when extrapolating the linear behavior observed at distances $> 100$ kbp from the jump, one gets a skew of 5.3%, i.e. a value consistent with the skew measured in intergenic regions around the six experimentally known replication origins namely $7.0 \pm 0.5\%$. Overall, the detection of sharp upward jumps in the skew profiles with characteristics similar to those of experimentally determined replication origins and with no downward counterpart further supports the existence, in human chromosomes, of replication-associated strand asymmetries, leading to the identification of numerous putative replication origins active in germ-line cells.

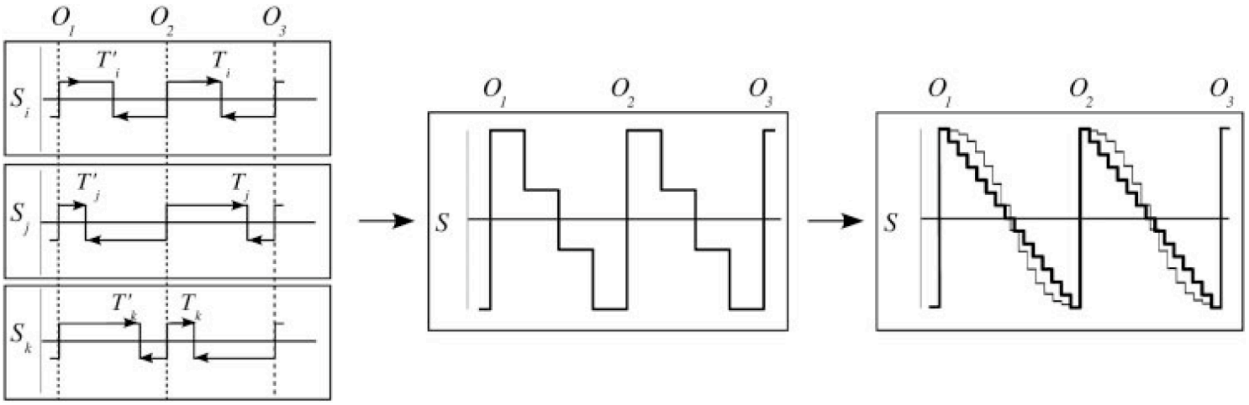
### C. A model of replication in mammalian genomes

Following the observation of jagged skew profiles similar to factory roofs in Sect. V A, and the quantitative confirmation of the existence of such (piecewise linear) profiles in the neighborhood of 1012 putative origins in Fig. 12, we have proposed, in Touchon *et al.* [100] and Brodie of Brodie *et al.* [98], a rather crude model for replication in the human genome that relies on the hypothesis that the replication origins are quite well positioned while the terminations are randomly distributed. Although some replication terminations have been found at specific sites in *S. cerevisiae* and to some extent in *Schizosaccharomyces pombe* [152], they occur randomly between active origins in *Xenopus* egg extracts [153, 154]. Our results indicate that this property can be extended to replication in human germ-line cells. As

illustrated in Fig. 13, replication termination is likely to rely on the existence of numerous potential termination sites distributed along the sequence. For each termination site (used in a small proportion of cell cycles), strand asymmetries associated with replication will generate a step-like skew profile with a downward jump at the position of termination and upward jumps at the positions of the adjacent origins (as in bacteria). Various termination positions will thus correspond to classical replicon-like skew profiles (Fig. 13, left panel). Addition of these profiles will generate the intermediate profile (Fig. 13, central panel). In a simple picture, we can reasonably suppose that termination occurs with constant probability at any position on the sequence. This behavior can, for example, result from the binding of some termination factor at any position between successive origins, leading to a homogeneous distribution of termination sites during successive cell cycles. The final skew profile is then a linear segment decreasing between successive origins (Fig. 13, right panel). Let us point out that firing of replication origins during time interval of the S phase [155] might result in some flattening of the skew profile at the origins as sketched in Fig. 13 (right panel, gray curve). In the present state, our results [98, 100, 146] support the hypothesis of random replication termination in human, and more generally in mammalian cells (Fig. 10), but further analyses will be necessary to determine what scenario is precisely at work.

## VI. A WAVELET-BASED METHODOLOGY TO DISENTANGLE TRANSCRIPTION- AND REPLICATION-ASSOCIATED STRAND ASYMMETRIES REVEALS A REMARKABLE GENE ORGANIZATION IN THE HUMAN GENOME

During the duplication of eukaryotic genomes that occurs during the S phase of the cell cycle, the different replication origins are not all activated simultaneously [132, 135, 138, 150, 155, 156]. Recent technical developments in genomic clone microarrays have led to a novel way of detecting the temporal order of DNA replication [155, 156]. The arrays are used to estimate *replication timing ratios i.e.* ratios between the average amount of DNA in the S phase at a locus along the genome and the usual amount of DNA present in the G1 phase for that locus. These ratios should vary between 2 (throughout the S phase, the amount of DNA for the earliest replicating regions is twice the amount during G1 phase) and 1 (the latest replicating regions are not duplicated until the end of S phase). This approach has been successfully used to generate genome-wide maps of replication timing for *S. cerevisiae* [157],

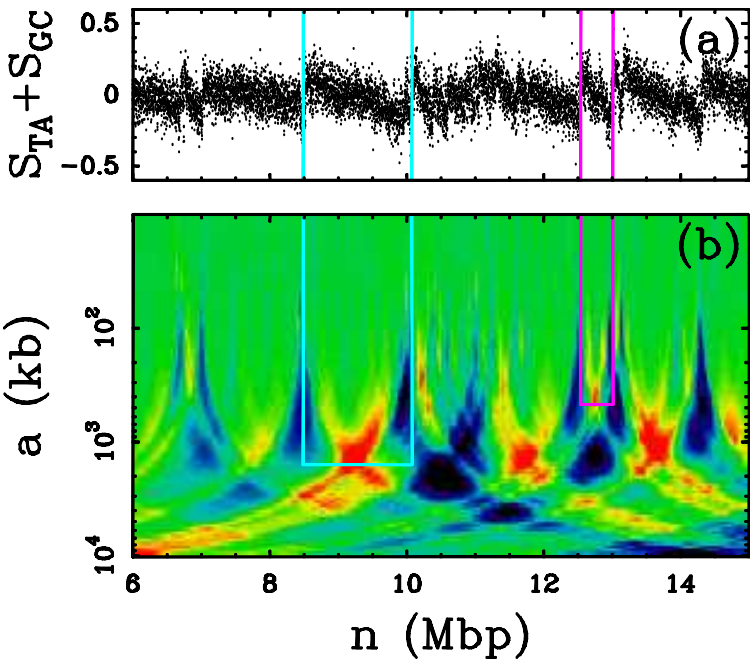



**Fractals and Wavelets. Figure 13**: Model of replication termination [98, 100]. Schematic representation of the skew profiles associated with three replication origins $O_1$, $O_2$, and $O_3$; we suppose that these replication origins are adjacent, bidirectional origins with similar replication efficiency. The abscissa represents the sequence position; the ordinate represents the $S$ value (arbitrary units). Upward (or downward) steps correspond to origin (or termination) positions. For convenience, the termination sites are symmetric relative to $O_2$. (Left) Three different termination positions $T_i$, $T_j$, and $T_k$, leading to elementary skew profiles $S_i$, $S_j$, and $S_k$ as predicted by the replicon model [146]. (Center) Superposition of these three profiles. (Right) Superposition of a large number of elementary profiles leading to the final factory-roof pattern. In the simple model, termination occurs with equal probability on both sides of the origins, leading to the linear profile (thick line). In the alternative model, replication termination is more likely to occur at lower rates close to the origins, leading to a flattening of the profile (gray line).

*Drosophila melanogaster* [137] and human [158]. Very recently, two new analyses of human chromosome 6 [156] and 22 [155] have improved replication timing resolution from 1 Mbp down to ∼100 kbp using arrays of overlapping tile path clones. In this section, we report on a very promising first step towards the experimental confirmation of the thousand putative replication origins described in Sect. V. The strategy will consist in mapping them on the recent high-resolution timing data [156] and in checking that these regions replicate earlier than their surrounding [114]. But to provide a convincing experimental test, we need as a prerequisite to extract the contribution of the compositional skew specific to replication.

### A. Disentangling transcription- and replication- associated strand asymmetries

The first step to detect putative replication domains consists in developing a multi-scale pattern recognition methodology based on the WT of the strand compositional asymmetry




**Fractals and Wavelets. Figure 14**: Wavelet-based analysis of genomic sequences. (a) Skew profile $S$ of a 9 Mbp repeat-masked fragment of human chromosome 21. (b) WT of $S$ using $\phi_R$ (Fig. 1(c)); $T_{\phi_R}[S](n,a)$ 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 that 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 replication domains match the distance between the putative replication origins (1.6 Mbp and 470 kbp respectively).

$S$ using as analyzing wavelet $\phi_R(x)$ (Eq. (12)) that is adapted to perform an objective segmentation of factory-roof skew profiles (Fig. 1(c)). As illustrated in Fig. 14, the space-scale location of significant maxima values in the 2d WT decomposition (red areas in Fig. 14(b)) indicates the middle position (spatial location) of candidate replication domains whose size is given by the scale location. In order to avoid false positives, we then check that there does exist a well-defined upward jump at each domain extremity. These jumps appear in Fig. 14(b) as blue cone-shape areas pointing at small scale to the jumps positions where are located the putative replication origins. Note that because the analyzing wavelet is of zero mean (Eq. (2)), the WT decomposition is insensitive to (global) asymmetry offset.

But as discussed in Sect. IV, the overall observed skew $S$ also contains some contribution induced by transcription that generates step-like blocks corresponding to (+) and (−) genes [96, 97, 129]. Hence, when superimposing the replication serrated and transcrip-

tion step-like skew profiles, we get the following theoretical skew profile in a replication domain [114]:

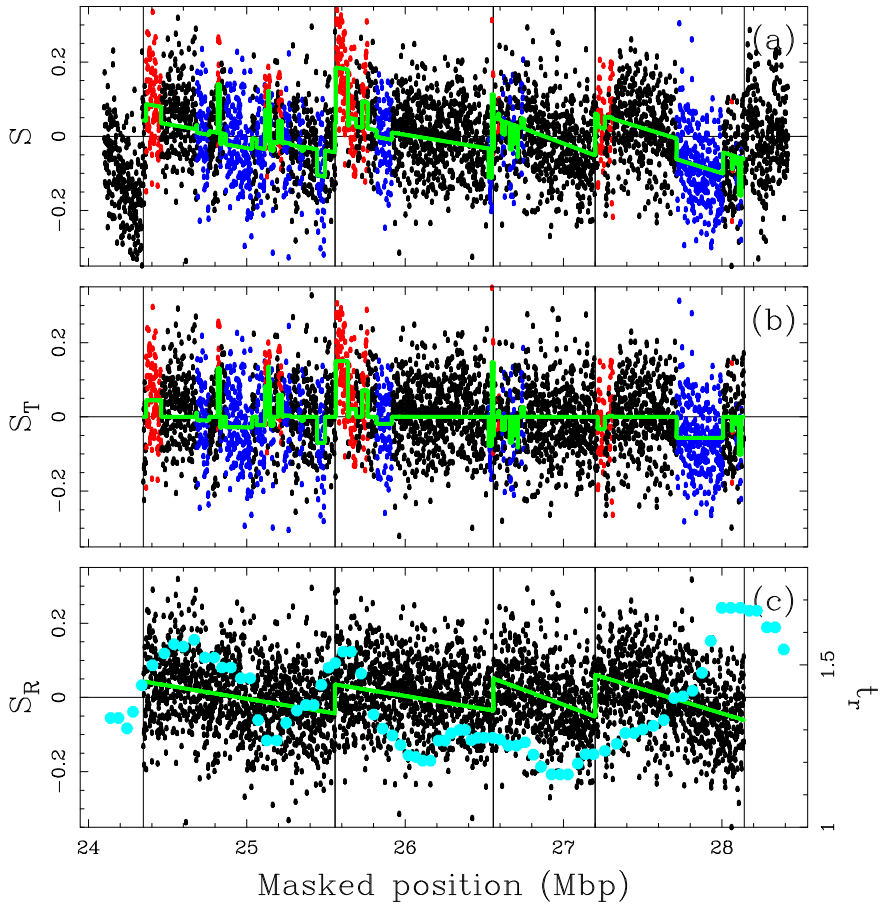
$$S(x') = S_R(x') + S_T(x') = -2\delta \times (x' - 1/2) + \sum_{\text{gene}} c_g \chi_g(x'), \tag{23}$$

where position $x'$ within the domain has been rescaled between 0 and 1, $\delta > 0$ is the replication bias, $\chi_g$ is the characteristic function for the $g^{\text{th}}$ gene (1 when $x'$ points within the gene and 0 elsewhere) and $c_g$ is its transcriptional bias calculated on the Watson strand (likely to be positive for (+) genes and negative for (−) genes). The objective is thus to detect human replication domains by delineating, in the noisy $S$ profile obtained at 1 kbp resolution (Fig. 15(a)), all chromosomal loci where $S$ is well fitted by the theoretical skew profile Eq. (23).

In order to enforce strong compatibility with the mammalian replicon model (Sect. V C), we will only retain the domains the most likely to be bordered by putative replication origins, namely those that are delimited by upward jumps corresponding to a transition from a negative $S$ value $< -3\%$ to a positive $S$ value $> +3\%$. Also, for each domain so-identified, we will use a least-square fitting procedure to estimate the replication bias $\delta$, and each of the gene transcription bias $c_g$. The resulting $\chi^2$ value will then be used to select the candidate domains where the noisy $S$ profile is well described by Eq. (23). As illustrated in Fig. 15 for a fragment of human chromosome 6 that contains 4 adjacent replication domains (Fig. 15(a)), this method provides a very efficient way of disentangling the step-like transcription skew component (Fig. 15(b)) from the serrated component induced by replication (Fig. 15(c)). Applying this procedure to the 22 human autosomes, we delineated 678 replication domains of mean length $\langle L \rangle = 1.2 \pm 0.6$ Mbp, spanning 28.3% of the genome and predicted 1060 replication origins.

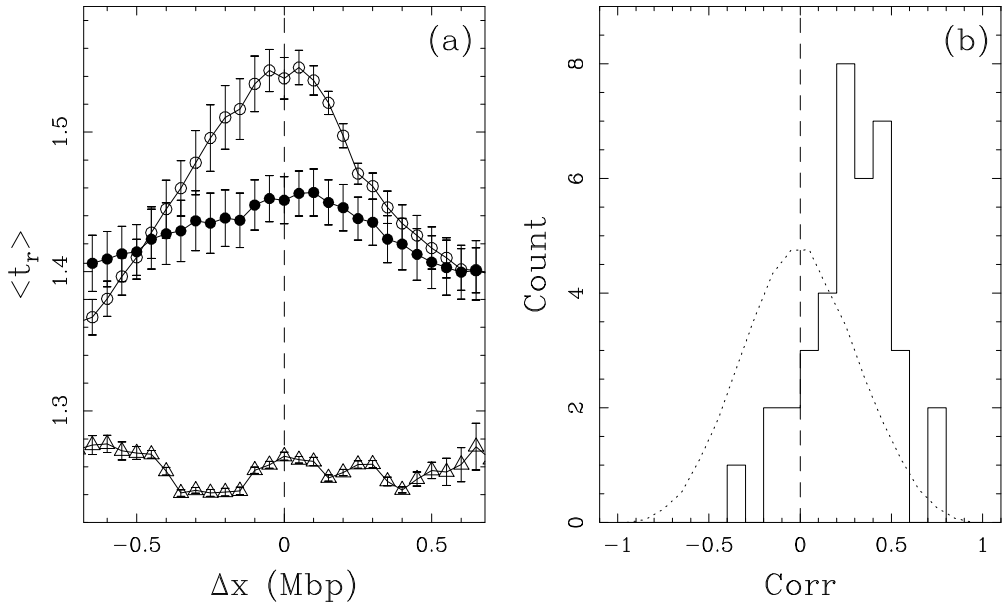
### B. DNA replication timing data corroborate *in silico* human replication origin predictions

Chromosome 22 being rather atypical in gene and GC contents, we mainly report here on the correlation analysis [114] between nucleotide compositional skew and timing data for chromosome 6 which is more representative of the whole human genome. Note that timing data for clones completely included in another clone have been removed after checking for




**Fractals and Wavelets. Figure 15**: (a) Skew profile $S$ of a 4.3 Mbp repeat-masked fragment of human chromosome 6 [114]; each point corresponds to a 1 kbp window: red, (+) genes; blue, (−) genes; black, intergenic regions (the color was defined by majority rule); the estimated skew profile (Eq. (23)) is shown in green; vertical lines correspond to the locations of 5 putative replication origins that delimit 4 adjacent domains identified by the wavelet-based methodology. (b) Transcription-associated skew $S_T$ obtained by subtracting the estimated replication-associated profile (green lines in (c)) from the original $S$ profile in (a); the estimated transcription step-like profile (second term on the *rhs* of Eq. (23)) is shown in green. (c) Replication-associated skew $S_R$ obtained by subtracting the estimated transcription step-like profile (green lines in (b)) from the original $S$ profile in (a); the estimated replication serrated profile (first term in the *rhs* of Eq. (23)) is shown in green; the light-blue dots correspond to high-resolution $t_r$ data.

timing ratio value consistency leaving 1648 data points. The timing ratio value at each point has been chosen as the median over the 4 closest data points to remove noisy fluctuations resulting from clone heterogeneity (clone length 100±51 kbp and distance between successive clone mid-points 104 ± 89 kbp), so that the spatial resolution is rather inhomogeneous ∼300 kbp. Note that using asynchronous cells also results in some smoothing of the data,




**Fractals and Wavelets. Figure 16**: (a) Average replication timing ratio (±SEM) determined around the 83 putative replication origins (•), 20 origins with well-defined local maxima (○) and 10 late replicating origins (△). $\Delta x$ is the native distance to the origins in Mbp units [114]. (b) Histogram of Pearson's correlation coefficient values between $t_r$ and the absolute value of $S_R$ over the 38 predicted domains of length $L \geq 1$ Mbp. The dotted line corresponds to the expected histogram computed with the correlation coefficients between $t_r$ and $|S|$ profiles over independent windows randomly positioned along chromosome 6 and with the same length distribution as the 38 detected domains.

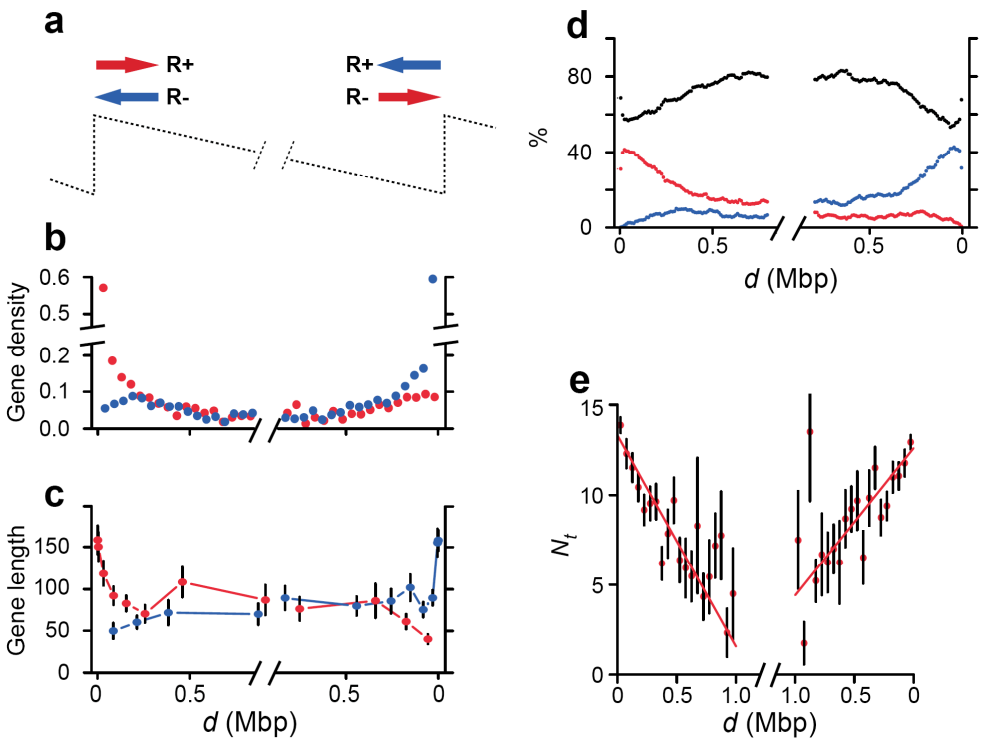
possibly masking local maxima.

Our wavelet-based methodology has identified 54 replication domains in human chromosome 6 [114]; these domains are bordered by 83 putative replication origins among which 25 are common to two adjacent domains. Four of these contiguous domains are shown in Fig. 15. In Fig. 15(c), on top of the replication skew profile $S_R$, are reported for comparison the high-resolution timing ratio $t_r$ data from Ref. [156]. The histogram of $t_r$ values obtained at the 83 putative origin locations displays a maximum at $t_r \simeq \langle t_r \rangle \simeq 1.5$ (data not shown) and confirms what is observed in Fig. 15(c), namely that a majority of the predicted origins are rather early replicating with $t_r \gtrsim 1.4$. This contrasts with the rather low $t_r$ ($\simeq 1.2$) values observed in domain central regions (Fig. 15(c)). But there is an even more striking feature in the replication timing profile in Fig. 15(c): 4 among the 5 predicted origins correspond, relatively to the experimental resolution, to local maxima of the $t_r$ profile. As shown in Fig. 16(a), the average $t_r$ profile around the 83 putative replication origins decreases regu-

larly on both sides of the origins over a few (4–6) hundreds kbp confirming statistically that domain borders replicate earlier than their left and right surroundings which is consistent with these regions being true replication origins mostly active early in $S$ phase. In fact, when averaging over the top 20 origins with a well-defined local maximum in the $t_r$ profile, $\langle t_r \rangle$ displays a faster decrease on both sides of the origin and a higher maximum value $\sim 1.55$ corresponding to the earliest replicating origins. On the opposite, when averaging $t_r$ profiles over the top 10 late replicating origins, we get, as expected, a rather flat mean profile ($t_r \sim 1.2$) (Fig. 16(a)). Interestingly, these origins are located in rather wide regions of very low GC content ($\lesssim 34\%$, not shown) correlating with chromosomal G banding patterns predominantly composed of GC-poor isochores [159, 160]. This illustrates how the statistical contribution of rather flat profiles observed around late replicating origins may significantly affect the overall mean $t_r$ profile. Individual inspection of the 38 replication domains with $L \geq 1$ Mbp shows that, in those domains that are bordered by early replicating origins ($t_r \gtrsim 1.4 - 1.5$), the replication timing ratio $t_r$ and the absolute value of the replication skew $|S_R|$ turn out to be strongly correlated. This is quantified in Fig. 16(b) by the histogram of the Pearson's correlation coefficient values that is clearly shifted towards positive values with a maximum at $\sim 0.4$. Altogether the results of this comparative analysis provide the first experimental verification of *in silico* replication origins predictions: the detected putative replication domains are bordered by replication origins mostly active in the early $S$ phase, whereas the central regions replicate more likely in late $S$ phase.

### C. Gene organization in the detected replication domains

Most of the 1060 putative replication origins that border the detected replication domains are intergenic (77%) and are located near to a gene promoter more often than would be expected by chance (data not shown) [101]. The replication domains contain approximately equal numbers of genes oriented in each direction (1511 (+) genes and 1507 (−) genes). Gene distributions in the 5' halves of domains contain more (+) genes than (−) genes, regardless of the total number of genes located in the half-domains (Fig. 17(b)). Symmetrically, the 3' halves contain more (−) genes than (+) genes (Fig. 17(b)). 32.7% of half-domains contain one gene, and 50.9% contain more than one gene. For convenience, (+) genes in the 5' halves and (−) genes in the 3' halves are defined as R+ genes (Fig. 17(a)): their transcription is,




**Fractals and Wavelets. Figure 17**: Analysis of the genes located in the identified replication domains [101]. (a) Arrows indicate the R+ orientation, i.e. the same orientation as the most frequent direction of putative replication fork progression; R− orientation (opposed direction); red, (+) genes; blue, (−) genes. (b) Gene density. The density is defined as the number of 5' ends (for (+) genes) or of 3' ends (for (−) genes) in 50-kbp adjacent windows, divided by the number of corresponding domains. In abscissa, the distance, $d$, in Mbp, to the closest domain extremity. (c) Mean gene length. Genes are ranked by their distance, $d$, from the closest domain extremity, grouped by sets of 150 genes, and the mean length (kbp) is computed for each set. (d) Relative number of base pairs transcribed in the + direction (red), − direction (blue) and non-transcribed (black) determined in 10-kbp adjacent sequence windows. (e) Mean expression breadth using EST data [101].

in most cases, oriented in the same direction as the putative replication fork progression (genes transcribed in the opposite direction are defined as R− genes). The 678 replication domains contain significantly more R+ genes (2041) than R− genes (977). Within 50 kbp of putative replication origins, the mean density of R+ genes is 8.2 times greater than that of R− genes. This asymmetry weakens progressively with the distance from the putative origins, up to $\sim$ 250 kbp (Fig. 17(b)). A similar asymmetric pattern is observed when the domains containing duplicated genes are eliminated from the analysis, whereas control domains obtained after randomization of domain positions present similar R+ and R− gene density distributions (Supplementary in Ref. [101]). The mean length of the R+ genes near

the putative origins is significantly greater ($\sim$ 160 kbp) than that of the R$-$ genes ($\sim$ 50 kbp), however both tend towards similar values ($\sim$ 70 kbp) at the center of the domain (Fig. 17(c)). Within 50 kbp of the putative origins, the ratio between the numbers of base pairs transcribed in the R+ and R$-$ directions is 23.7; this ratio falls to $\sim 1$ at the domain centers (Fig. 17(d)). In Fig. 17(e) are reported the results of the analysis of the breadth of expression, $N_t$ (number of tissues in which a gene is expressed) of genes located within the detected domains [101]. As measured by EST data (similar results are obtained by SAGE or microarray data [101]), $N_t$ is found to decrease significantly from the extremities to the center in a symmetrical mannner in the 5' and 3' half-domains (Fig. 17(e)). Thus, genes located near the putative replications origins tend to be widely expressed whereas those located far from them are mostly tissue-specific.

To summarize, the results reported in this section provide the first demonstration of quantitative relationships in the human genome between gene expression, orientation and distance from putative replication origins [101]. A possible key to the understanding of this complex architecture is the coordination betwen replication and transcription [101]. The putative replication origins would mostly be active early in the $S$ phase in most tissues. Their activity could result from particular genomic context involving transcription factor binding sites and/or from the transcription of their neighboring housekeeping genes. This activity could also be associated with an open chromatin structure, permissive to early replication and gene expression in most tissues [161–164]. This open conformation could extend along the first gene, possibly promoting the expression of further genes. This effect would progressively weaken with the distance from the putative replication origin, leading to the observed decrease in expression breadth. This model is consistent with a number of data showing that in metazoans, ORC and RNA polymerase II colocalize at transcriptional promoter regions [165], and that replication origins are determined by epigenetic information such as transcription factor binding sites and/or transcription [166–169]. It is also consistent with studies in *Drosophila* and humans that report correlation between early replication timing and increased probability of expression [137, 155, 156, 165, 170]. Furthermore, near the putative origins bordering the replication domains, transcription is preferentially oriented in the same direction as replication fork progression. This co-orientation is likely to reduce head-on collisions between the replication and transcription machineries, which may induce deleterious recombination events either directly or via stalling of the replication fork [171,

172]. In bacteria, co-orientation of transcription and replication has been observed for essential genes, and has been associated with a reduction in head-on collisions between DNA and RNA polymerases [173]. It is noteworthy that in human replication domains such co-orientation usually occurs in widely-expressed genes located near putative replication origins. Near domain centers, head-on collisions may occur in 50% of replication cycles, regardless of the transcription orientation, since there is no preferential orientation of the replication fork progression in these regions. However, in most cell types, there should be few head-on collisions due to the low density and expression breadth of the corresponding genes. Selective pressure to reduce head-on collisions may thus have contributed to the simultaneous and coordinated organization of gene orientation and expression breadth along the detected replication domains [101].

## VII. FUTURE DIRECTIONS

From a statistical multifractal analysis of nucleotide strand asymmetries in mammalian genomes, we have revealed the existence of jumps in the noisy skew profiles resulting from asymmetries intrinsic to the transcription and replication processes [98, 100]. This discovery has led us to extend our 1D WTMM methodology to an adapted multi-scale pattern recognition strategy in order to detect putative replication domains bordered by replication origins [101, 114]. The results reported in this manuscript show that directly from the DNA sequence, we have been able to reveal the existence in the human genome (and very likely in all mammalian genomes), of regions bordered by early replicating origins in which gene position, orientation and expression breadth present a high level of organization, possibly mediated by the chromatin structure.

These results open new perspectives in DNA sequence analysis, chromatin modeling as well as in experiment. From a bioinformatic and modeling point of view, we plan to study the lexical and structural characteristics of our set of putative origins. In particular we will search for conserved sequence motifs in these replication initiation zones. Using a sequence-dependent model of DNA-histones interactions, we will develop physical studies of nucleosome formation and diffusion along the DNA fiber around the putative replication origins. These bioinformatic and physical studies, performed for the first time on a large number of replication origins, should shed light on the processes at work during the recog-

nition of the replication initiation zone by the replication machinery. From an experimental point of view, our study raises new opportunities for future experiments. The first one concerns the experimental validation of the predicted replication origins (e.g. by molecular combing of DNA molecules [174]), which will allow us to determine precisely the existence of replication origins in given genome regions. Large scale study of all candidate origins is in current progress in the laboratory of O. Hyrien (École Normale Supérieure, Paris). The second experimental project consists in using Atomic Force Microscopy (AFM) [175] and Surface Plasmon Resonance Microscopy (SPRM) [176] to visualize and study the structural and mechanical properties of the DNA double helix, the nucleosomal string and the 30nm chromatin fiber around the predicted replication origins. This work is in current progress in the experimental group of F. Argoul and C. Moskalenko at the Laboratoire Joliot-Curie (ENS, Lyon) [83]. Finally the third experimental perspective concerns *in situ* studies of replication origins. Using fluorescence techniques (FISH chromosome painting [177]), we plan to study the distributions and dynamics of origins in the cell nucleus, as well as chromosome domains potentially associated with territories and their possible relation to nuclear matrix attachment sites. This study is likely to provide evidence of chromatin rosette patterns as suggested in Ref. [146]. This study is under progress in the molecular biology experimental group of F. Mongelard at the Laboratoire Joliot-Curie.

**Acknowledgement**

We thank O. Hyrien, F. Mongelard and C. Moskalenko for interesting discussions. This work was supported by the Action Concertée Incitative Informatique, Mathématiques, Physique en Biologie Moléculaire 2004 under the project "ReplicOr", the Agence Nationale de la Recherche under the project "HUGOREP" and the program "Emergence" of the Conseil Régional Rhônes-Alpes and by the Programme d'Actions Intégrées Tournesol.

**Bibliography**

**A. Primary Literature**

[1] Goupillaud, P., Grossmann, A. & Morlet, J. (1984). Cycle-octave and related transforms in

seismic signal analysis. *Geoexploration* **23**, 85–102.

[2] Grossmann, A. & Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *S.I.A.M. J. of Math. Anal.* **15**, 723–736.

[3] Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J., Freysz, E., Grasseau, G., Muzy, J.-F. & Pouligny, B. (1992). Wavelet transform of fractals. In *Wavelets and Applications* (Y. Meyer, ed.), pp. 286 – 352. Springer, Berlin.

[4] Arneodo, A., Argoul, F., Elezgaray, J. & Grasseau, G. (1989). Wavelet transform analysis of fractals: application to nonequilibrium phase transitions. In *Nonlinear Dynamics* (G. Turchetti, ed.), pp. 130–180. World Scientific, Singapore.

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

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

[7] Holschneider, M. & Tchamitchian, P. (1990). Régularité locale de la fonction non-différentiable de Riemann. In *Les Ondelettes en 1989* (P. G. Lemarié, ed.), pp. 102–124. Springer, Berlin.

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

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

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

[11] Mallat, S. & Zhong, S. (1992). Characterization of signals from multiscale edges. *IEEE Trans. Patt. Recog. Mach. Intell.* **14**, 710–732.

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

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

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

[15] Muzy, J.-F., Bacry, E. & Arneodo, A. (1993). Multifractal formalism for fractal signals: The

structure-function approach versus the wavelet-transform modulus-maxima method. *Phys. Rev. E* **47**, 875–884.

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

[17] Jaffard, S. (1997). Multifractal formalism for functions part I: results valid for all functions. *SIAM J. Math. Anal.* **28**, 944–970.

[18] Jaffard, S. (1997). Multifractal formalism for functions part II: self-similar functions. *SIAM J. Math. Anal.* **28**, 971–998.

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

[20] Arneodo, A., Audit, B., Decoster, N., Muzy, J.-F. & Vaillant, C. (2002). 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* (A. Bunde, J. Kropp & H. J. Schellnhuber, eds.), pp. 26–102. Springer Verlag, Berlin.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[44] Khalil, A., Joncas, G., Nekka, F., Kestener, P. & Arneodo, A. (2006). Morphological analysis of $H_I$ features. II. Wavelet-based multifractal formalism. *Astrophysical Journal Supplement Series* **165**, 512–550.

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

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

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

[48] Meneveau, C. & Sreenivasan, K. R. (1991). The multifractal nature of turbulent energy-dissipation. *J. Fluid Mech.* **224**, 429–484.

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

[50] Kestener, P. & Arneodo, A. (2007). A multifractal formalism for vector-valued random fields based on wavelet analysis: application to turbulent velocity and vorticity 3D numerical data. *Stoch. Environ. Res. Risk Assess. DOI: 10.1007/s00477-007-0121-6* .

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

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

[53] Li, W. (1990). Mutual information functions versus correlation-functions. *J. Stat. Phys* **60**, 823–837.

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

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

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

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

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

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

[60] Havlin, S., Buldyrev, S. V., Goldberger, A. L., Mantegna, R. N., Peng, C.-K., Simons, M. & Stanley, H. E. (1995). Statistical and linguistic features of DNA sequences. *Fractals* **3**, 269–284.

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

[62] Herzel, H., Ebeling, W. & Schmitt, A. (1994). Entropies of biosequences: The role of repeats. *Phys. Rev. E* **50**, 5061–5071.

[63] Li, W. (1997). The measure of compositional heterogeneity in DNA sequences is related to measures of complexity. *Complexity* **3**, 33–37.

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

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

[66] Nee, S. (1992). Uncorrelated DNA walks. *Nature* **357**, 450.

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

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

[69] Berthelsen, C. L., Glazier, J. A. & Raghavachari, S. (1994). Effective multifractal spectrum of a random walk. *Phys. Rev. E* **49**, 1860–1864.

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

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

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

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

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

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

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

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

[78] Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.-F. & Thermes, C. (1998). Nucleotide composition effects on the long-range correlations in human genes. *Eur.*

*Phys. J. B* **1**, 259–263.

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

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

[81] Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J. & Rando, O. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626–630.

[82] Vaillant, C., Audit, B. & Arneodo, A. (2007). Experiments confirm the influence of genome long-range correlations on nucleosome positioning. Submitted to *Phys. Rev. Lett.* .

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

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

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

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

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

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

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

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

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

[92] Mrázek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral

genomes. *Proc. Natl. Acad. Sci. USA* **95**, 3720–3725.

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

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

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

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

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

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

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

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

[101] Huvet, M., Nicolay, S., Touchon, M., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A. & Thermes, C. (2007). Human gene organization driven by the coordination of replication and transcription. To appear in *Genome Res.* .

[102] Arneodo, A., Bacry, E., Jaffard, S. & Muzy, J.-F. (1997). Oscillating singularities on Cantor sets: A grand-canonical multifractal formalism. *J. Stat. Phys.* **87**, 179–209.

[103] Arneodo, A., Bacry, E., Jaffard, S. & Muzy, J.-F. (1998). Singularity spectrum of multifractal functions involving oscillating singularities. *J. of Fourier Anal. & Appl.* **4**, 159–174.

[104] Parisi, G. & Frisch, U. (1985). Fully developed turbulence and intermittency. In *Turbulence*

*and Predictability in Geophysical Fluid Dynamics and Climate Dynamics* (M. Ghil, R. Benzi & G. Parisi, eds.), Proc. of Int. School, pp. 84–88. North-Holland, Amsterdam.

[105] Collet, P., Lebowitz, J. & Porzio, A. (1987). The dimension spectrum of some dynamical systems. *J. Stat. Phys.* **47**, 609–644.

[106] Grassberger, P., Badii, R. & Politi, A. (1988). Scaling laws for invariant measures on hyperbolic and non hyperbolic attractors. *J. Stat. Phys.* **51**, 135 –178.

[107] Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia, I. & Shraiman, B. I. (1986). Fractal measures and their singularities: The characterization of strange sets. *Phys. Rev. A* **33**, 1141–1151.

[108] Paladin, G. & Vulpiani, A. (1987). Anomalous scaling laws in multifractal objects. *Phys. Rep.* **156**, 147–225.

[109] Rand, D. (1989). The singularity spectrum for hyperbolic Cantor sets and attractors. *Ergod. Th. Dyn. Sys.* **9**, 527–541.

[110] Argoul, F., Arneodo, A., Elezgaray, J. & Grasseau, G. (1990). Wavelet analysis of the self-similarity of diffusion-limited aggregates and electrodeposition clusters. *Phys. Rev. A* **41**, 5537–5560.

[111] Farmer, J. D., Ott, E. & Yorke, J. A. (1983). The dimension of chaotic attractors. *Physica D* **7**, 153–180.

[112] Grassberger, P. & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D* **9**, 189–208.

[113] Bohr, T. & Tèl, T. (1988). The thermodynamics of fractals. In *Direction in Chaos, Vol. 2* (B. L. Hao, ed.), pp. 194–237. World Scientific, Singapore.

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

[115] Mandelbrot, B. B. & Van Ness, J. W. (1968). Fractional brownian motions, fractal noises and applications. *S.I.A.M. Rev.* **10**, 422–437.

[116] Arneodo, A., Bacry, E. & Muzy, J. F. (1998). Random cascades on wavelet dyadic trees. *J. Math. Phys.* **39**, 4142–4164.

[117] Benzi, R., Biferale, L., Crisanti, A., Paladin, G., Vergassola, M. & Vulpiani, A. (1993). A random process for the construction of multiaffine fields. *Physica D* **65**, 352–358.

[118] Mandelbrot, B. B. (1974). Intermitttent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *J. Fluid Mech.* **62**, 331–358.

[119] Arneodo, A., Bacry, E., Manneville, S. & Muzy, J. F. (1998). Analysis of random cascades using space-scale correlation functions. *Phys. Rev. Lett.* **80**, 708–711.

[120] Castaing, B. & Dubrulle, B. (1995). Fully-developed turbulence - a unifying point-of-view. *J. Phys. II France* **5**, 895–899.

[121] Novikov, E. A. (1994). Infinitely divisible distributions in turbulence. *Phys. Rev. E* **50**, 3303–3305.

[122] Gojobori, T., Li, W. H. & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360–369.

[123] Li, W. H., Wu, C. I. & Luo, C. C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71.

[124] Petrov, D. A. & Hartl, D. L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci. USA* **96**, 1475–1479.

[125] Zhang, Z. & Gerstein, M. (2003). Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338–5348.

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

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

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

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

[130] Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., ichi Isono, K., Koseki, H., Fuchikami, T., Abe, K., Murray, H. L., Zucker, J. P., Yuan, B., Bell, G. W., Herbolsheimer, E., Hannett, N. M., Sun, K., Odom, D. T., Otte, A. P., Volkert, T. L., Bartel, D. P., Melton, D. A., Gifford, D. K., Jaenisch, R. & Young, R. A. (2006). Control of developmental regulators by polycomb

in human embryonic stem cells. *Cell* **125**, 301–313.

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

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

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

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

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

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

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

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

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

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

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

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

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

[144] McNairn, A. J. & Gilbert, D. M. (2003). Epigenomic replication: linking epigenetics to DNA

replication. *Bioessays* **25**, 647–656.

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

[146] Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Brodie of Brodie, E.-B., Nicolay, S., St-Jean, P., Thermes, C., Touchon, M. & Vaillant, C. (2007). DNA in chromatin: from genome-wide sequence analysis to the modeling of replication in mammals. *Advances in Chemical Physics* **135**, 203–252.

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

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

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

[150] Berezney, R., Dubey, D. D. & Huberman, J. A. (2000). Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* **108**, 471–484.

[151] Vassilev, L. T., Burhans, W. C. & DePamphilis, M. L. (1990). Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells. *Mol. Cell. Biol.* **10**, 4685–4689.

[152] Codlin, S. & Dalgaard, J. Z. (2003). Complex mechanism of site-specific DNA replication termination in fission yeast. *EMBO J.* **22**, 3431–3440.

[153] Little, R. D., Platt, T. H. & Schildkraut, C. L. (1993). Initiation and termination of DNA replication in human rRNA genes. *Mol. Cell Biol.* **13**, 6600–6613.

[154] Santamaria, D., Viguera, E., Martinez-Robles, M. L., Hyrien, O., Hernandez, P., Krimer, D. B. & Schvartzman, J. B. (2000). Bi-directional replication and random termination. *Nucleic Acids Res.* **28**, 2099–2107.

[155] White, E. J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E. J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M. & Schübeler, D. (2004). DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proc. Natl. Acad. Sci. USA* **101**, 17771–17776.

[156] Woodfine, K., Beare, D. M., Ichimura, K., Debernardi, S., Mungall, A. J., Fiegler, H., Collins,

V. P., Carter, N. P. & Dunham, I. (2005). Replication timing of human chromosome 6. *Cell Cycle* **4**, 172–176.

[157] Raghuraman, M. K., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. J. & Fangman, W. L. (2001). Replication dynamics of the yeast genome. *Science* **294**, 115–121.

[158] Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y. & Ikemura, T. (2002). Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum. Mol. Genet.* **11**, 13–21.

[159] Costantini, M., Clay, O., Federico, C., Saccone, S., Auletta, F. & Bernardi, G. (2007). Human chromosomal bands: nested structure, high-definition map and molecular basis. *Chromosoma* **116**, 29–40.

[160] Schmegner, C., Hameister, H., Vogel, W. & Assum, G. (2007). Isochores and replication time zones: a perfect match. *Cytogenet. Genome Res.* **116**, 167–172.

[161] Chakalova, L., Debrand, E., Mitchell, J. A., Osborne, C. S. & Fraser, P. (2005). Replication and transcription: shaping the landscape of the genome. *Nat. Rev. Genet.* **6**, 669–677.

[162] Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P. & Bickmore, W. A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566.

[163] Hurst, L. D., Pál, C. & Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310.

[164] Sproul, D., Gilbert, N. & Bickmore, W. A. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6**, 775–781.

[165] MacAlpine, D. M., Rodriguez, H. K. & Bell, S. P. (2004). Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev.* **18**, 3094–3105.

[166] Danis, E., Brodolin, K., Menut, S., Maiorano, D., Girard-Reydet, C. & Méchali, M. (2004). Specification of a DNA replication origin by a transcription complex. *Nat. Cell Biol.* **6**, 721–730.

[167] DePamphilis, M. L. (2005). Cell cycle dependent regulation of the origin recognition complex. *Cell Cycle* **4**, 70–79.

[168] Ghosh, M., Liu, G., Randall, G., Bevington, J. & Leffak, M. (2004). Transcription factor binding and induced transcription alter chromosomal c-myc replicator activity. *Mol. Cell*

*Biol.* **24**, 10193–10207.

[169] Lin, C. M., Fu, H., Martinovsky, M., Bouhassira, E. & Aladjem, M. I. (2003). Dynamic alterations of replication timing in mammalian cells. *Curr. Biol.* **13**, 1019–1028.

[170] Jeon, Y., Bekiranov, S., Karnani, N., Kapranov, P., Ghosh, S., MacAlpine, D., Lee, C., Hwang, D. S., Gingeras, T. R. & Dutta, A. (2005). Temporal profile of replication of human chromosomes. *Proc. Natl. Acad. Sci. USA* **102**, 6419–6424.

[171] Deshpande, A. M. & Newlon, C. S. (1996). DNA replication fork pause sites dependent on transcription. *Science* **272**, 1030–1033.

[172] Takeuchi, Y., Horiuchi, T. & Kobayashi, T. (2003). Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev.* **17**, 1497–1506.

[173] Rocha, E. P. C. & Danchin, A. (2003). Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet.* **34**, 377–378.

[174] Herrick, J., Stanislawski, P., Hyrien, O. & Bensimon, A. (2000). Replication fork density increases during DNA synthesis in X. laevis egg extracts. *J. Mol. Biol.* **300**, 1133–1142.

[175] Zlatanova, J. & Leuba, S. H. (2003). Chromatin fibers, one-at-a-time. *J. Mol. Biol.* **331**, 1–19.

[176] Tassius, C., Moskalenko, C., Minard, P., Desmadril, M., Elezgaray, J. & Argoul, F. (2004). Probing the dynamics of a confined enzyme by surface plasmon resonance. *Physica A* **342**, 402–409.

[177] Müller, W. G., Rieder, D., Kreth, G., Cremer, C., Trajanoski, Z. & McNally, J. G. (2004). Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol. Cell Biol.* **24**, 9359–9370.

## B. Books and Reviews

### *1. Fractals*

- Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature.* Freeman, San Francisco.

- Stanley, H. E. & Osbrowski, N., eds. (1986). *On Growth and Form: Fractal and Non-Fractal Patterns in Physics.* Martinus Nijhof, Dordrecht.

- Pietronero, L. & Tosatti, E., eds. (1986). *Fractals in Physics.* North-Holland,

Amsterdam.

- Peitgen, H. O. & Saupe, D., eds. (1987). *The Science of Fractal Images*. Springer, New-York.

- Stanley, H. E. & Ostrowski, N., eds. (1988). *Random Fluctuations and Pattern Growth*. Kluwer, Dordrecht.

- Feder, J. (1988). *Fractals*. Pergamon, New-York.

- Vicsek, T. (1989). *Fractal Growth Phenomena*. World Scientific, Singapore.

- Avnir, D., ed. (1988). *The Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers*. Wiley, New-York.

- Aharony, A. & Feder, J., eds. (1989). Fractals in Physics, Essays in Honour of BB Mandelbrot, *Physica D* **38**. North-Holland, Amsterdam.

- West, B. J. (1990). *Fractal Physiology and Chaos in Medecine*. World Scientific, Singapore.

- Family, F. & Vicsek, T. (1991). *Dynamics of Fractal Surfaces*. World Scientific, Singapore.

- Bunde, A. & Havlin, S., eds. (1991). *Fractals and Disordered Systems*. Springer, Berlin.

- Vicsek, T., Schlesinger, M. & Matsuchita, M., eds. (1994). *Fractals in Natural Science*. World Scientific, Singapore.

- Bunde, A. & Havlin, S., eds. (1994). *Fractals in Science*. Springer, Berlin.

- West, B. J. & Deering, W. (1994). Fractal Physiology for Physicists: Levy Statistics. *Phys. Rep.* **246**, 1–100.

- Frisch, U. (1995). *Turbulence*. Cambridge University Press, Cambridge.

- Wilkinson, G. G., Kanellopoulos, J. & Megier, J., eds. (1995). *Fractals in Geoscience and Remote Sensing, Image Understanding Research Senes*, vol. 1. ECSC-EC-EAEC, Brussels, Luxembourg.

- Barabàsi, A. L. & Stanley, H. E. (1995). *Fractals Concepts in Surface Growth.* Cambridge University Press, Cambridge.

- Family, F., Meakin, P., Sapoval, B. & Wood, R., eds. (1995). *Fractal Aspects of Materials. Material Research Society Symposium Proceedings*, vol. 367. MRS, Pittsburg.

- Bouchaud, J.-P. & Potters, M. (1997). *Théorie des Risques Financiers.* Cambridge University Press, Cambridge.

- Mantegna, R. N. & Stanley, H. E. (2000). *An Introduction to Econophysics.* Cambridge University Press, Cambridge.

- Bunde, A., Kropp, J. & Schellnhuber, H. J., eds. (2002). *The Science of Disasters: Climate Disruptions, Heart Attacks and Market Crashes.* Springer, Berlin.

*2. Wavelets*

- Combes, J.-M., Grossmann, A. & Tchamitchian, P., eds. (1989). *Wavelets.* Springer, Berlin.

- Meyer, Y. (1990). *Ondelettes.* Herman, Paris.

- Lemarie, P. G., ed. (1990). *Les Ondelettes en 1989.* Springer, Berlin.

- Meyer, Y., ed. (1992). *Wavelets and Applications.* Springer, Berlin.

- Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

- Ruskai, M. B., Beylkin, G., Coifman, R., Daubechies, I., Mallat, S., Meyer, Y. & Raphael, L., eds. (1992). *Wavelets and their Applications.* Jones and Barlett, Boston.

- Chui, C. K. (1992). *An Introduction to Wavelets.* Academic Press, Boston.

- Meyer, Y. & Roques, S., eds. (1993). *Progress in Wavelets Analysis and Applications.* Éditions Frontières, Gif-sur-Yvette.

- Flandrin, P. (1993). *Temps-Fréquence.* Hermès, Paris.

- Farge, M., Hunt, J. C. R. & Vassilicos, J. C., eds. (1993). *Wavelets, Fractals and Fourier.* Clarendon Press, Oxford.

- Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J. & Muzy, J.-F. (1995). *Ondelettes, Multifractales et Turbulences : de l'ADN aux Croissances Cristallines.* Diderot Éditeur, Art et Sciences, Paris.

- Erlebacher, G., Hussaini, M. Y. & Jameson, L. M., eds. (1996). *Wavelets: Theory and Applications.* Oxford University Press, Oxford.

- Holschneider, M. (1996). *Wavelets: An Analysis Tool.* Oxford University Press, Oxford.

- Abry, P. (1997). *Ondelettes et Turbulences.* Diderot Éditeur, Art et Sciences, Paris.

- Mallat, S. (1998). *A Wavelet Tour in Signal Processing.* Academic Press, New-York.

- Torresani, B. (1998). *Analyse Continue par Ondelettes.* Éditions de Physique, Les Ulis.

- Silverman, B. W. & Vassilicos, J. C., eds. (2000). *Wavelets: The Key to Intermittent Information?* Oxford University Press, Oxford.

- Jaffard, S., Meyer, Y. & Ryan, R. D., eds. (2001). *Wavelets: Tools for Science & Technology.* SIAM, Philadelphia.

*3. DNA and Chromatin*

- van Holde, K. E. (1988). *Chromatin.* Springer-Verlag, New-York.

- Watson, J. D., Gilman, M., Witkowski, J. & Zoller, M. (1992). *Recombinant DNA.* Freeman, New-York.

- Alberts, B. & Watson, J. (1994). *Molecular Biology of the Cell.* Garland Publishing, New-York, 3rd edn.

- Lewin, B. (1994). *Genes V.* Oxford University Press, Oxford.

- Kolchanov, N. A. & Lim, H. A. (1994). *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution.* World Scientific, Singapore.

- Sudbery, P. (1998). *Human Molecular Genetics.* Addison Wesley, Singapore.

- Wolfe, A. P. (1998). *Chromatin Structure and Function.* Academic Press, London, 3rd edn.

- Calladine, C. R. & Drew, H. R. (1999). *Understanding DNA.* Academic Press, San Diego.

- Graur, D. & Li, W. H. (1999). *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

- Hartl, D. L. & Jones, E. W. (2001). *Genetics: Analysis of Genes & Genomes.* Jones and Bartlett, Sudbury.

- Kornberg, A. & Baker, T. A. (1992). *DNA Replication.* WH Freeman and co., New-York.