Universal $1/f$ noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome

Wentian Li*
The Robert S. Boas Center for Genomics and Human Genetics, North Shore LIJ Institute for Medical Research, 350 Community Drive, Manhasset, New York 10030, USA

Dirk Holste†
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
(Received 28 January 2004; revised manuscript received 28 October 2004; published 20 April 2005)

Spatial fluctuations of guanine and cytosine base content (GC%) are studied by spectral analysis for the complete set of human genomic DNA sequences. We find that (i) $1/f^{\alpha}$ decay is universally observed in the power spectra of all 24 chromosomes, and (ii) the exponent $\alpha \approx 1$ extends to about 10^7 bases, one order of magnitude longer than has previously been observed. We further find that (iii) almost all human chromosomes exhibit a crossover from $\alpha_1 \approx 1$ $(1/f^{\alpha_1})$ at lower frequency to $\alpha_2 < 1$ $(1/f^{\alpha_2})$ at higher frequency, typically occurring at around 30 000–100 000 bases, while (iv) the crossover in this frequency range is virtually absent in human chromosome 22. In addition to the universal $1/f^{\alpha}$ noise in power spectra, we find (v) several lines of evidence for chromosome-specific correlation structures, including a 500 000 base long oscillation in human chromosome 21. The universal $1/f^{\alpha}$ spectrum in the human genome is further substantiated by a resistance to reduction in variance of guanine and cytosine content when the window size is increased.

DOI: 10.1103/PhysRevE.71.041910 PACS number(s): 87.10.+e, 87.14.Gg, 87.15.Cc, 02.50.−r

I. INTRODUCTION

By measuring the proportion of a signal's power $S(f)$ falling into a range of frequency components f, a power spectrum of the form $S(f) \sim 1/f^{\alpha}$ distinguishes between two prototypes of noise: white noise ($\alpha = 0$) and Brownian noise ($\alpha = 2$). The intermittent range, termed "$1/f$ noise," can practically be defined as $1/f^{\alpha}$ $(0.5 \lesssim \alpha \lesssim 1.5)$. $1/f$ noise was experimentally observed first in electric current fluctuations of the thermionic tube at the beginning of the 19th century [1]. Since then, $1/f$ noise has been found repeatedly in many other conducting materials [2]. More generally, it has also been observed in wide ranges of natural as well as human-related phenomena, including traffic flow, starlight, speech, music, and human coordination [3,4]. For biological sequences, such as DNA, the concept of slow-varying, multiple-length variations in the power of frequency components can be translated to long-ranging correlations in the spatial arrangement of the four bases adenine (A), cytosine (C), guanine (G), and thymine (T) along the sequence. One can categorize chemically A, C, G, and T as strong (G or C) or weak (A or T) bonding. It has been shown that fluctuations of the GC base content along a DNA sequence are typically more strongly correlated when compared to other possible binary classifications [5,6]. Initial studies of $1/f$ noise in DNA sequences were motivated by a model of spatial $1/f$ noise of symbolic sequence evolution [7]. Subsequently, empirical $1/f$ spectra were indeed observed in non-protein-coding DNA sequences [8], and their generality in DNA sequences was further illustrated in [9].

$1/f$ noise has been detected in a variety of different species and taxonomic classes, including bacteria [10], yeasts [11], insects [12], and other higher eukaryotic genomes. Integrating this and several other lines of evidence, a consensus on $1/f$ noise in DNA sequences has emerged: (1) for DNA sequences of the order of 10^6 bases (1 Mb), $1/f^{\alpha}$ spectrum $(\alpha \approx 1)$ is consistently observed; (2) for isochores, which are DNA sequences of relatively homogeneous base concentration at least 300×10^3 bases (300 kb) long [13–15], a $1/f^{\alpha}$ spectrum is also observed, but typically shows a smaller exponent $\alpha < 0.7$ [14,16,17]; (3) for DNA sequences of the order of several kb, the decay of $S(f)$ is nontrivial and may depend on whether the sequence is protein coding [8]. The viral DNA sequence of the λ phage, e.g., shows a single step in its GC base concentration and its spectrum is $S(f) \sim 1/f^2$, which is characteristic of random block sequences [18]. We note that the universal scaling of $S(f) \sim 1/f^{\alpha}$ $(\alpha \approx 1)$ across all species discussed in [9] has apparently been restricted to a length scale of 1 kb, by averaging the spectrum over many $N = 2$ kb DNA segments.

With the availability of the first completed version of the DNA sequence of the human genome [19], several studies have been able to demonstrate that the base-base correlation function $\Gamma(d)$ (d is the distance between bases) of several DNA sequences follows a power-law decay, $\Gamma(d) \sim 1/d^{\gamma}$. For instance, the DNA sequence of human chromosome 22 shows statistically significant power-law correlations up to $d = 1$ Mb, and correlations in the DNA sequence of chromosomes 21 are statistically significant up to several Mb (with the scaling exponent γ changing beyond a few kb) [6,20]. While the DNA sequences of human chromosomes 21 and 22 are about 34 Mb long, in order to estimate the limit of the range of the $1/f^{\alpha}$ spectrum, longer sequences are necessary.

After the release of the draft of the human genome sequence in February 2001, about three years later in 2004, a

*Electronic address: wli@nslij-genetics.org
†Electronic address: holste@mit.edu

dozen (out of 24) human chromosomes have been completed with a sequence accuracy following the standard of less than one error per 10 000 DNA bases (99.99% accuracy) [21]. Building upon the release of updated, high-quality sequence data, in the era of genomics we can now conduct a systematic analysis of several issues of $1/f$ noise in the DNA sequences of our own species *Homo sapiens*, which have been pursued over the last decade in a fragmentary manner.

In this paper, we use the DNA sequences of the complete set of 22 autosomes and two sex chromosomes to address the following issues. Is $1/f$ noise universally present across the entire set of human genome sequences? Does $1/f$ noise extend to lower-frequency ranges in longer DNA sequences? Is the decay of $S(f)$ characterized by a single exponent α, or does it exhibit crossovers (multiple scaling exponents)? Given the presence of universal variations at multiple scales, do these coexist with variations at chromosome-specific scales?

II. DATA AND METHODS

In this section, we introduce the data for human genome sequences, as well as the notation and definitions used throughout this study. Twenty-four chromosomes are assembled in build 34 by the number scheme of the National Center for Biotechnology Information (NCBI) (human genome hg16 release). Sequence data were downloaded from the UCSC human genome repository [45]. Unsequenced bases are kept to preserve the estimated spacing between sequenced bases. Human chromosomes (Chr) 13, 14, 15, 21, and 22 contain large amount of unsequenced bases in the left end of their DNA sequences, consisting of about 15%, 17%, 18%, 21%, and 29% of the individual chromosome size, respectively; 51% of chromosome Y are unsequenced.

Our analysis of human DNA sequences is conducted using coarse-grained data. Each original sequence was transformed into a spatial series of GC content (GC%) values. To this end, we evenly partition a DNA sequence into N nonoverlapping windows of length w bases, compute $\rho_i(w)$ = GC% for each window i, to obtain a spatial GC% series:

$$\{\rho_i\} \equiv \{\rho_i(w)\}, \quad i = 1,2, \ldots, N. \tag{1}$$

Table I lists the corresponding window sizes for each human chromosome. Since different human chromosomes have different sizes, whereas the number of partitions (N) is the same, the window lengths vary.

Human DNA sequences contain a large fraction of interspersed repeats, i.e., copies of ancestral sequence fragments that possess a high similarity between the duplicated and the ancestral sequence. One can detect interspersed repeats by using the program REPEATMASKER [22]. "Soft-masked" annotations of interspersed repeats are taken from the DNA sequences of the UCSC human genome repository [45], where repetitive (nonrepetitive) bases are annotated in small (capital) letters. Figure 1 shows the length distribution of the two sequence classes of uninterrupted nonrepetitive and interspersed repeat sequences for both the human and the mouse genome. On the length scale 10–1000 bases, the repeat distributions between human and mouse sequences differ most at the length of about 300 bases, where human repetitive sequences exhibit a clear peak that corresponds to short interspersed nucleic Alu elements. In addition, either genome harbors short repetitive sequences that characterize the distribution up to 200 bases, such as a peak at about 150 bases for mouse repetitive sequences.

TABLE I. Average GC content ($\overline{\rho}$) and the window size (w) for partitions using $N=2^{17}$ nonoverlapping windows for 24 human chromosomes. Low-frequency scaling exponents α_1 are estimated from $S(f;s=3)\sim 1/f^{\alpha_1}$ in the range of $10^{-7}<f<10^{-5}$ base^{-1}, and high-frequency scaling exponents α_2 are estimated in the range of $10^{-5}<f<2\times10^{-4}$ base^{-1}. The differences between the two scaling exponents, $\Delta\alpha\equiv\alpha_1-\alpha_2$, are listed in the fifth column. Low- and high-frequency exponents for $S(f)$ with substituted interspersed repeats are indicated by α_1' and α_2', and their difference by $\Delta\alpha' \equiv \alpha_1'-\alpha_2'$.

Chr	$\overline{\rho}$	w (kb)	α_1	α_2	$\Delta\alpha$	α_1'	α_2'	$\Delta\alpha'$
1	41.7	1.88	0.88	0.46	0.42	0.80	0.29	0.51
2	40.2	1.86	0.99	0.51	0.48	0.96	0.30	0.66
3	39.7	1.52	0.95	0.43	0.53	0.88	0.27	0.61
4	38.2	1.46	0.87	0.34	0.53	0.75	0.19	0.57
5	39.5	1.38	0.89	0.39	0.51	0.88	0.23	0.65
6	39.6	1.30	0.99	0.36	0.63	0.86	0.24	0.63
7	40.7	1.21	0.97	0.46	0.51	0.87	0.33	0.55
8	40.1	1.12	0.97	0.42	0.55	0.91	0.26	0.66
9	41.3	1.04	0.96	0.39	0.57	0.90	0.28	0.62
10	41.6	1.03	0.97	0.52	0.46	0.95	0.34	0.61
11	41.6	1.03	1.05	0.50	0.55	0.97	0.35	0.62
12	40.8	1.01	0.97	0.39	0.59	0.89	0.28	0.61
13	38.5	0.86	0.83	0.33	0.50	0.73	0.24	0.49
14	40.9	0.80	1.03	0.36	0.66	0.95	0.27	0.68
15	42.2	0.76	0.90	0.50	0.40	0.83	0.39	0.44
16	44.8	0.69	0.91	0.51	0.40	0.81	0.36	0.45
17	45.5	0.62	0.98	0.57	0.42	0.89	0.44	0.46
18	39.8	0.58	1.12	0.40	0.72	1.12	0.28	0.83
19	48.4	0.49	1.00	0.56	0.44	0.81	0.37	0.45
20	44.1	0.49	0.87	0.51	0.36	0.83	0.30	0.53
21	40.9	0.36	0.91	0.33	0.58	0.86	0.22	0.64
22	47.9	0.38	0.90	0.62	0.28	0.86	0.40	0.45
X	39.4	1.17	0.93	0.38	0.54	0.73	0.18	0.55
Y	39.1	0.38	0.83	0.38	0.45	0.70	0.21	0.49

To investigate the effect of interspersed repeats, we substitute each repeat by random bases according to the chromosomal level of GC%. Transformed, repeat-substituted DNA sequences of original human chromosomes are distinguished from original sequences. On the coarse-grained level, it is equivalent to the replacement in the $\{\rho_i\}$ ($i=1,2,\ldots,N$) series of any values calculated from the interspersed repeats by a random value which is sampled from a Gaussian distribution; the mean and variance of this Gaussian distribution are the same as those of GC% in the original sequence. Another possibility consists in substituting repetitive sequences by a constant value (e.g., the averaged GC%

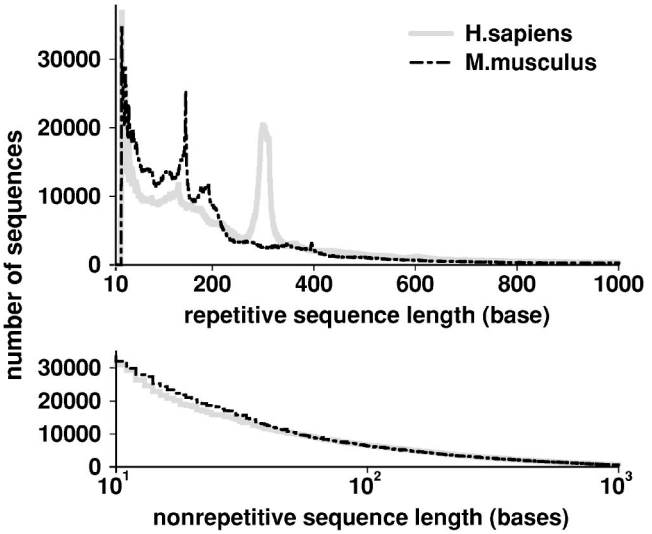

FIG. 1. Human and mouse genome-wide length distributions of interspersed repeat sequences (top, in linear scale) and nonrepetitive sequences (bottom, in log-linear scale). The length distribution of repetitive sequences exhibits a peak around 300 bases for the human genome, that correspond to Alu repeats, and a peak around 150 bases for the mouse genome.

value of the original sequence). This method introduces additional correlations (and less variance) in the $\{\rho_i\}$ series, and is not adopted in this paper.

Three different, albeit functionally related, measures are applied to the $\{\rho_i\}$ series: the power spectrum as a function of the frequency $S(f)$, the correlation function $\Gamma(d)$ as a function of the distance d between windows, and the variance $\sigma^2(w)$ of the GC% series as a function of the window size w.

First, we conduct spectral analyses by calculating the power spectrum, the absolute squared average of the Fourier transform, defined as

$$S(f) \equiv \frac{1}{N}\left|\sum_{k=1}^{N} \rho_k e^{-i2\pi kf/N}\right|^2, \tag{2}$$

where N is the total number of windows, and f is measured in units of cycles/window, which can be converted to units of cycles/base by using the window size (cf. Table I).

Coarse graining "hides" patterns at scales smaller than w bases. The choice of $N=2^{17}$ windows was made such that it is (i) sufficiently large to cover small-scale fluctuations, while (ii) at the same time sufficiently small so that the spectral analysis is computationally feasible. As different chromosomes have difference lengths, equal number of partitions leads to different window sizes w.

The original $S(f)$, or periodogram, contains $N/2$ independent spectral components. One can filter periodograms to obtain a "smoothed" spectrum $S(f;s)$, where s is the span-size parameter. Since filtering with a relatively large s value possibly distorts the shape of $S(f;s)$ at lower-frequency components, different span sizes are applied for different frequency ranges.

The second measure applied to the $\{\rho_i\}$ series is the correlation function $\Gamma(d)$, which is computed from two truncated series of $\{\rho_i\}$, $\rho'=\{\rho_k\}$ $(k=1,2,\ldots,N-d)$ and $\rho''=\{\rho_k\}$ $(k=d+1,d+2,\ldots,N)$, and defined as

$$\Gamma(d) \equiv \frac{\mathrm{cov}(\rho',\rho'')}{\sqrt{\mathrm{var}(\rho')}\sqrt{\mathrm{var}(\rho'')}}, \tag{3}$$

where $\mathrm{cov}(\rho',\rho'')=\langle\rho'\rho''\rangle-\langle\rho'\rangle\langle\rho''\rangle$ and $\mathrm{var}(\rho')=\langle\rho'^2\rangle-\langle\rho'\rangle^2$ are the covariance and variance, respectively, andρ' is the average taken over k. Note that $\Gamma(d)$ defined in Eq. (3) is slightly different from that defined using a periodic boundary condition.

The third and final measure applied to the $\{\rho_i\}$ series is the variance $\sigma^2(w)$, defined as

$$\sigma^2(w) \equiv \langle\rho(w)^2\rangle - \langle\rho(w)\rangle^2 \tag{4}$$

as a function of the window size w. The power spectrum, the correlation function, and the window-size-dependent variance are interrelated quantities [16]. If $S(f)\sim 1/f^{\alpha}$, $\Gamma(d)\sim 1/d^{\gamma}$, and $\sigma^2(w)\sim 1/w^{\beta}$ are power-law functions, then their scaling exponents are related by $\alpha=1-\gamma$ and $\gamma=\beta$ [16]. Moreover, $\sigma^2(w)$ can be obtained from $\Gamma(d)$as:

$$\sigma^2(w) \sim \frac{\Gamma(0)}{w}\left\{1+\frac{2}{w}\sum_{d=1}^{w-1}(w-d)\Gamma(d)\right\}. \tag{5}$$

The calculation of $S(f)$ and $\Gamma(d)$ was carried out using the statistical package S-PLUS (version 3.4, MathSoft, Inc.), and the type of filter implemented for $S(f)$ is the Daniell filter [23].

III. $1/f$ NOISE IS A UNIVERSAL FEATURE OF HUMAN DNA SEQUENCES

In this section, we use the power spectrum $S(f)$ to study the GC% of human genome sequences, with the goals of testing the universality of $1/f$ noise, quantifying different decay ranges for $S(f)\sim 1/f^{\alpha}$, and comparing $S(f)$ across DNA sequences of different human chromosomes.

Figure 2 shows for $N=2^{17}$ GC% values the power spectra $S(f)$ across all human chromosomes. We find that $S(f)$ exhibits no clear plateau at low frequency ($<10^{-6}$ cycles/base) and increases steadily with decreasing frequency. The decay can be mathematically approximated by a power law of the form $S(f)\sim 1/f^{\alpha}$ with $\alpha\approx 1$. Table I lists for the frequency range $f=10\ \mathrm{Mb}^{-1}-100\ \mathrm{kb}^{-1}$ the estimated scaling exponent α_1 for all chromosomes, using a best-fit regression of $\log_{10}S(f;s=3)=a+\alpha_1\log_{10}(f)$. We find that α_1 is typically close to $\alpha_1\approx 1$ with very little variation across chromosomes.

A closer inspection of Fig. 2 shows that the majority of $1/f$ spectra undergo a crossover from $\alpha_1\approx 1$ to $\alpha_2<1$ at high frequency. The deviation from $\alpha_1\approx 1$ starts about 30–100 kb and continues at smaller distances. Figure 3 illustrates this feature for $S(f;s=31)$ of the DNA sequences of Chr15, Chr21, and Chr22 in more detail. We find that chromosomes 15 and 21 exhibit clear crossovers at about 100 kb, while chromosome 22 exhibits no apparent break point. Table I contains for the frequency range of $f=100\ \mathrm{kb}^{-1}-5\ \mathrm{kb}^{-1}$ the

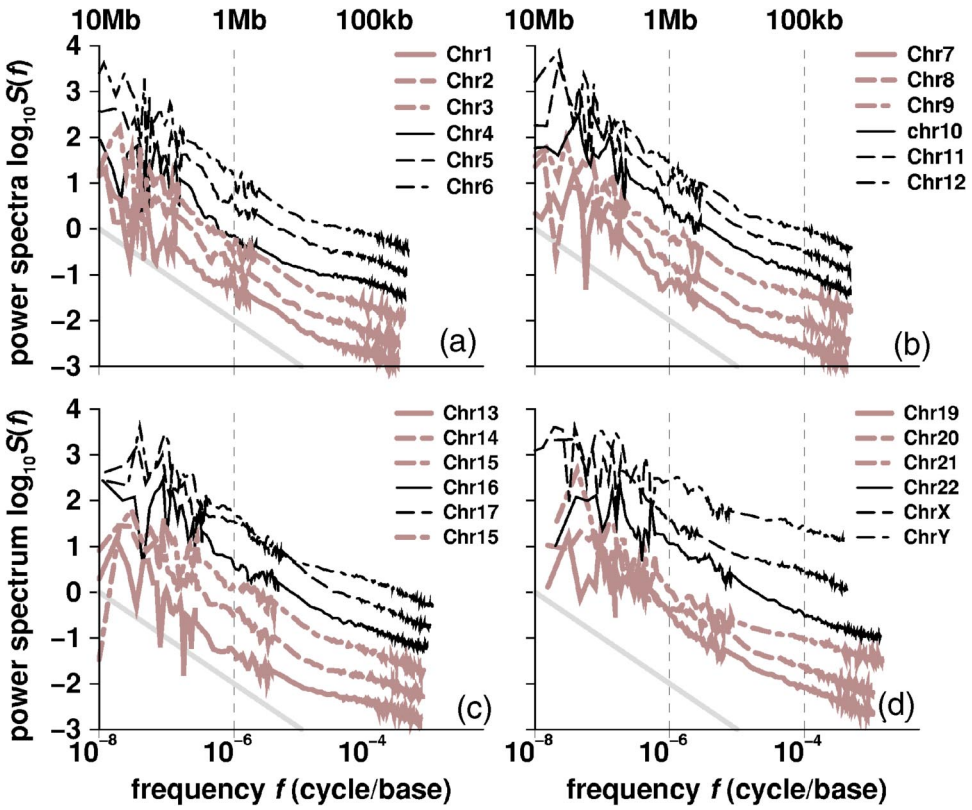

FIG. 2. Double-logarithmic representation of power spectra $S(f)$ of GC% of all 24 human chromosomes. Each plot shows $S(f)$ of six chromosomes (shifted on the y axis for clearer representation): chromosomes (a) 1–6; (b) 7–12; (c) 13–18; (d) 19–22, X, and Y. The x axis (in logarithmic scale) is converted from cycles/window to cycles/base by using the window sizes listed in Table I. $S(f)$ is filtered at different levels for different frequency ranges: $S(f;s=1)$ for the first ten spectral components, $S(f;s=3)$ for the components 11–30, $S(f;s=31)$ for the components 31–400, and $S(f;s=501)$ for the components 400–65 536 ($=2^{16}$). The decay according to $S(f)\sim 1/f$ is drawn for comparison in each plot.

corresponding scaling exponents α_2, obtained from the regression $\log_{10} S(f;s=3)=a+\alpha_2 \log_{10}(f)$. We find a pronounced difference in absolute values between $\alpha_1 \approx 1$ and $\alpha_2 < 1$, indicating a transition from the universal $1/f^{\alpha_1}$ ($\alpha_1 \approx 1$) spectrum at low frequency to a more flattened $1/f^{\alpha_2}$ ($\alpha_2 < 1$) spectrum at high frequency.

Figure 4(a) shows for all human chromosomes α_1 and α_2 as a function of chromosome-specific GC%. The majority of human chromosomes have a specific GC content ranging between 38% and 43%, whereas chromosomes 16, 17, 19, 20, and 22 have higher GC% up to 49%. While the low-frequency scaling exponent α_1 remains approximately independent of GC%, Fig. 4(a) shows that α_2 increases with increasing GC% and gives rise to a positive correlation between α_2 and GC%. Note that the x axis in Fig. 4(b) is slightly shifted from that of Fig. 4(a) because in (a) GC% represents the content of the whole sequence, whereas in (b) it represents only the content of nonrepetitive sequences. It is known that the GC% of repeats in the human genome is higher than the GC% in nonrepetitive sequences (data not shown).

The three chromosomes illustrated in Fig. 3 exhibit different degrees of transition from the $1/f^{\alpha_1}$ ($\alpha_1 \approx 1$) to the flattened $1/f^{\alpha_2}$ ($\alpha_2 < 1$) spectrum, with chromosome 21 (22) undergoing the sharpest (smoothest) transition. This observa-

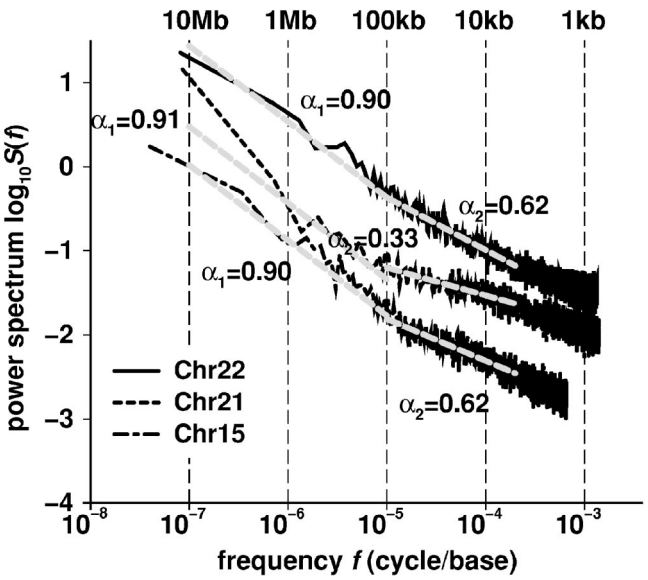

FIG. 3. Crossover from $S(f)\sim 1/f^{\alpha_1}$ to $S(f)\sim 1/f^{\alpha_2}$ illustrated for human chromosomes 15, 21, and 22 (smoothed with a span size of 31, and shown in double-logarithmic scale). The scaling exponents α_1 and α_2 for the frequency ranges 10 Mb^{-1}–100 kb^{-1} and 100 kb^{-1}–5 kb^{-1} are 0.90, 0.50 for chromosome 15, 0.91, 0.33 for chromosome 21, and 0.90, 0.62 for chromosome 22.

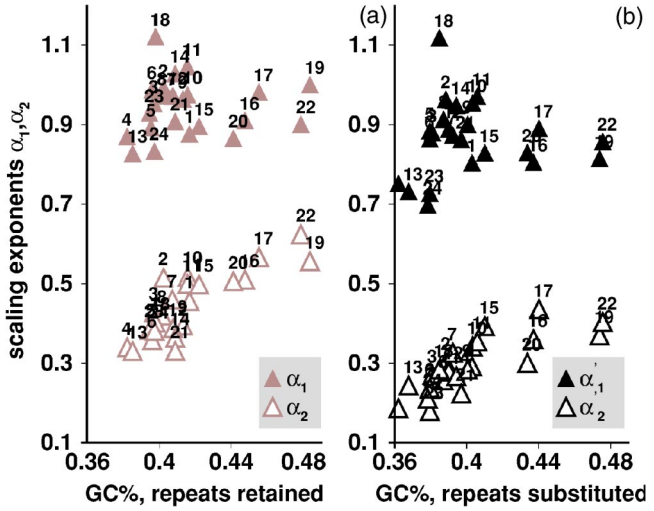

FIG. 4. (a) Scaling exponents α_1 and α_2 for fitting the power spectrum $S(f)\sim 1/f^{\alpha_i}$ $(i=1,2)$ at the frequency ranges of 10 Mb^{-1}–100 kb^{-1} and 100 kb^{-1}–5 kb^{-1}, respectively, versus the chromosome-specific GC content of all 24 human chromosomes (ChrX and ChrY are labeled as 23 and 24, respectively). (b) Scaling exponents α'_1 and α'_2 for $S(f)$ with substituted interspersed repeats, versus the GC content of nonrepetitive sequences.

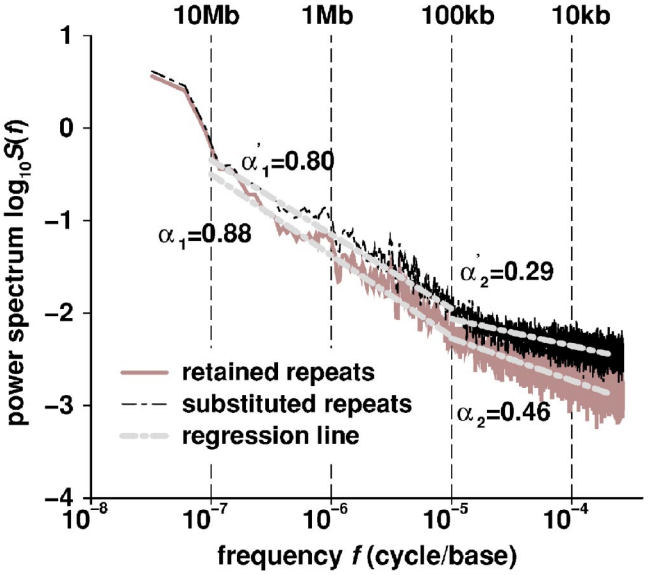

FIG. 5. Power spectra $S(f)$ of GC% for the original and the transformed (interspersed repeats-substituted) DNA sequence of human chromosome 1. The scaling exponents for low-frequency (10 Mb–100 kb) and high-frequency (100–5 kb) ranges are obtained by a best-fit regression of $\log_{10} S(f)$ over $\log_{10} f$.

tion can be further quantified by the change in scaling exponents α_1 and α_2. Table I lists for all chromosomes $\Delta\alpha=\alpha_1-\alpha_2$. Chromosome 22 is distinct from all other human chromosomes as the most scale-invariant one (the same or similar scaling exponent at different length scales). The same observation that human chromosome 22 was perhaps different from the remaining human chromosomes was made using limited sequence data in [14,20].

IV. INTERSPERSED REPEATS ARE NOT RESPONSIBLE FOR 1/f SPECTRUM

About 45% of human genomic DNA sequences are interspersed repeats [19]. Interspersed repeats consist of copies of the same sequence segment that are inserted in the human genome, possess a high similarity between the duplicated and ancestral sequence, and have been implicated in a variety of biological functions, including genome organization, human chromosome segregation, or regulation of gene expression [24]. Large copy numbers increase the sequence redundancy and it has been shown, e.g., that about 10% interspersed Alu repeats significantly increase base-base correlations in the range up to 300 bases [6].

Figure 5 shows the power spectrum $S(f)$ for the original human chromosome 1 and for the transformed sequence in which interspersed repeats are substituted. We find in the low-frequency range of $10^{-7}<f<10^{-5}$ cycles/base that $S(f)$ decays in the original sequence with $\alpha_1\approx 0.88$ and in the transformed sequence with $\alpha'_1\approx 0.80$, indicating only marginal differences in the decay properties of $S(f)$ due to repetitive sequences. In contrast, in the high-frequency range of $10^{-5}<f<2\times 10^{-4}$ we find $\alpha_2\approx 0.46$ and $\alpha'_2\approx 0.29$, and thus interspersed repeats contribute to the decay properties of $S(f)$ for high-frequency components by unflattening the power spectrum.

The scaling exponents α'_1 and α'_2 for repeat-substituted DNA sequences of all 24 human chromosomes are shown in Table I. The difference between low- and high-frequency ranges for DNA sequences of the original chromosomes, $\Delta\alpha=\alpha_1-\alpha_2$, is smaller than the difference between low- and high-frequency ranges for transformed sequences, $\Delta\alpha'=\alpha'_2-\alpha'_1$. When we compare α_1 and α'_1, as well as α_2 and α'_2, we find that the magnitude of α'_1 (α'_2) is always smaller than that of α_1 (α_2), which means a flattened spectrum due to the substitution of interspersed repeats. The average change of low-frequency scaling exponents, $\alpha_1-\alpha'_1$, is about 0.07, whereas the average change of high-frequency scaling exponents, $\alpha_2-\alpha'_2$, is about 0.14. This confirms that the universal presence of $1/f$ spectra at low frequency is not caused by interspersed repeats, but that interspersed repeats affect $S(f)$ predominantly at high frequency. A similar conclusion that the decay rate of base-base correlations in DNA sequences of human chromosomes 20, 21, and 22 is not markedly affected by the substitution of interspersed repeats was reached in [6].

We note that the extent of the deviation $|\alpha'-\alpha|$ depends on how the replacement of interspersed repeats is conducted. Possible substitutions of interspersed repeats include the substitution by a constant value or a randomly sampled value. In general, the substitution of GC% values calculated from the repetitive sequences by random values enhances the deviation and flattens the spectrum $S(f)$ more than the substitution by a constant value (e.g., average GC%).

V. RESISTANCE TO VARIANCE REDUCTION AT LARGER WINDOW SIZES

In this section, we study the decay properties of the variance (σ^2) of the spatial GC% series as a function of differ-

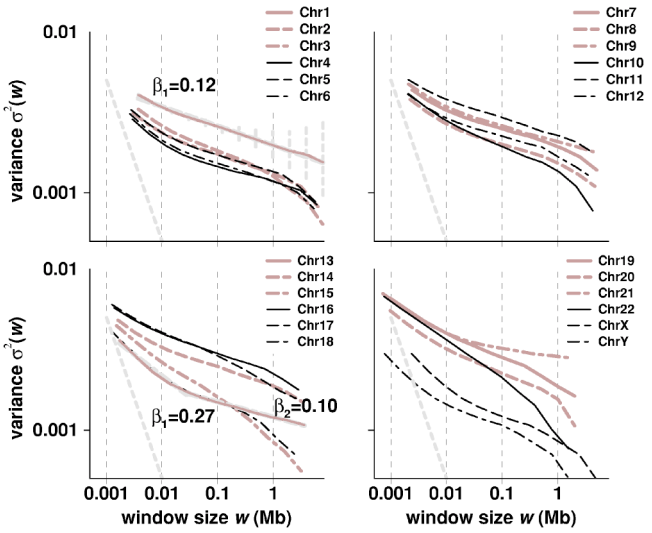

FIG. 6. Double-logarithmic representation of the variance $\sigma^2(w)$ of the spatial GC% series for all human chromosomes (Chr) as a function of the window size w: (a) Chr 1–6; (b) Chr 7–12; (c) Chr 13–18; and (d) Chr 19–22, X, and Y. Straight lines indicate $\sigma^2(w) \sim 1/w$ (corresponding to white noise). One regression line for Chr1 ($\beta \approx 0.12$) and a piecewise regression for Chr13 ($\beta \approx 0.27$ and 0.10) are drawn. The 95% confidence interval for the $\sigma^2(w)$ estimation of Chr1 at each point of w is marked by a vertical dashed line.

ence window sizes w, and we compare the scaling of σ^2 with the scaling of the power spectrum $S(f)$.

Early experimental measurement of the GC% distribution by using cesium chloride (CsCl) profiles [25] showed for mouse *Mus musculus* genomic DNA sequences that the variance of GC% values does not markedly decrease with the DNA segment size [26]. This experimental observation is directly related to the presence of $1/f$ spectra in DNA sequences [14,27]. If the variance of the spatial GC% series calculated at the window size w is $\sigma^2(w)$, then a scaling of $\sigma^2(w) \sim 1/w^\beta$ implies a corresponding scaling in the power spectrum $S(f) \sim 1/f^{1-\beta}$ [14,28]. If GC% is obtained from w uncorrelated bases, it follows a binomial distribution. Consequently, $\sigma^2(w) \sim \langle\rho\rangle(1-\langle\rho\rangle)/w \sim 1/w$ with $\beta=1$. The corresponding scaling exponent of the power spectrum is $\alpha=1-\beta=0$, and thus $S(f) \sim$ const is equivalent to white noise.

Figure 6 shows $\sigma^2(w)$ as a function of window size w for all human chromosomes. In a double-logarithmic representation, we find that $\log \sigma^2(w)$ decays approximately linearly with $\log(w)$. A decay according to $\sigma^2(w) \sim 1/w^\beta$ with $\beta=1$ leads to white noise. This situation is indicated in Fig. 6 by a straight line. Inspection of Fig. 6 shows, however, that the variance decays at a much slower rate than would be expected for white noise. The variance of the DNA sequence of human chromosome 1, e.g., gives rise to $\beta \approx 0.12$, and the corresponding scaling exponent $\alpha_1 \approx 1-\beta=0.88$ is indeed close to the estimated exponent listed in Table I. The scaling of the variance with the exponent $\beta \ll 1$ is in accord with low-frequency $1/f$ noise.

The scaling of $\sigma^2(w)$ shown in Fig. 6 differs from one human chromosome to another. For instance, in the range of $w=1$ kb–5 Mb, human chromosome 13 exhibits a clear transition from $\beta_2 \approx 0.27$ ($w<50$ kb) to $\beta_1 \approx 0.10$ ($w>50$ kb), corresponding to $S(f) \sim 1/f^{0.63}$ and $S(f) \sim 1/f^{0.9}$, respectively, at high and low frequencies. Other human chromosomes, although generally exhibiting a power-law scaling form of $\sigma^2(w)$, show deviations from the $\sigma^2(l) \sim 1/l^\beta$ line for the largest window sizes tested.

The investigation of $\sigma^2(w)$ as a function of different window sizes w requires careful examination [29,30]. First, since we partition each human chromosome in 2^k ($k=17,16,\ldots$) windows, the variance of the GC% series $\{\rho_i\}$ could be large when windows reside on the isochore borders, and small by chance if they start and/or end within an isochore.

Second, when the number of windows is small [e.g., the last data point of $\sigma^2(w)$ for each chromosome in Fig. 6 is calculated with the largest window size that gives rise to 32 windows], the standard error of the sample variance is large. The 95% confidence interval for $\sigma^2(w)$ of Chr1 is shown in Fig. 6(a), using the interval $[(w-1)\sigma^2/t_{0.025}, (w-1)\sigma^2/t_{0.975}]$, where t_x is defined by $\int_{-\infty}^{t_x}\chi^2(\mathcal{F}=w-1)dt = x$ where $\chi^2(\mathcal{F})$ is the chi-squared distribution with $\mathcal{F}$ the degrees of freedom [31]. Figure 6(a) shows that for fewer windows (and larger window sizes), the 95% confidence interval of $\sigma^2(w)$ could be sufficiently large such that the estimated value of β may change from sample to sample.

Finally, the relationship $\alpha+\beta=1$ [14,28], is based on the assumption that both $S(f)$ and $\sigma^2(w)$ are theoretical power-law functions. If $S(f)$ is a piecewise power-law function, as in the case of GC% fluctuation of human chromosomes, a correction term to the relationship $\alpha+\beta=1$ is expected.

VI. CHROMOSOME-SPECIFIC CORRELATION STRUCTURES

Apparently, $1/f$ noise in music and speech signals [32] does not prevent music and speech from sounding differently. Similarly, universal $1/f^\alpha$ spectra in GC% fluctuations across human chromosomes do not imply that all chromosomes exhibit the same detailed correlation structure. The generic trend of $S(f)$ spectra to increase at low frequency may "coexist" with small peaks at high frequency. Such chromosome-specific characteristic length scales can be more intuitively examined by correlation functions. In this section, we investigate the correlation function $\Gamma(d)$ of coarse-grained DNA sequences of human chromosomes with the aim of further examining chromosome-specific structures, such as characteristic length scales and oscillations detected by $\Gamma(d)$.

Figure 7 shows for all human chromosomes $\Gamma(d)$ of the GC% series $\{\rho_i\}$ calculated for the window sizes given in Table I, of all human chromosomes. For each chromosome, the minimum (maximum) distance is 80 (16 000) windows. Since each chromosome is partitioned into 2^{17} windows, the maximum distance d at which the correlation is examined is about $16\,000/2^{17} \approx 12\%$ of the total sequence length.

An inspection of Fig. 7 shows that the magnitude of correlation at the distance of $d=1$ Mb is clearly above the noise level. With the exceptions of Chr15, Chr22, and ChrY, the

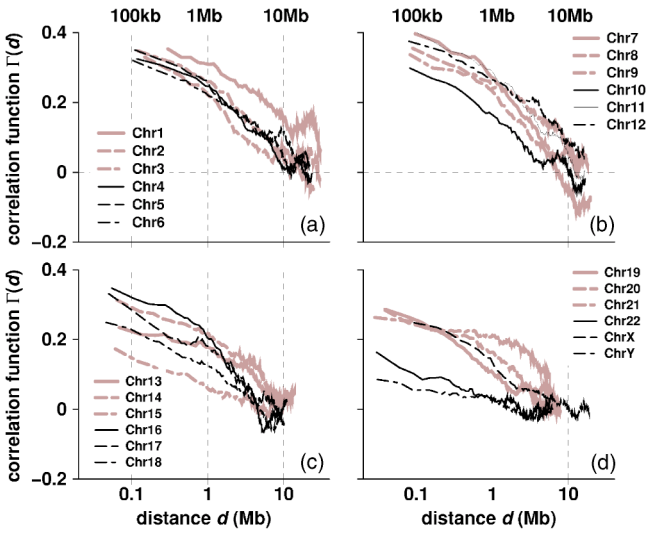

FIG. 7. Correlation function $\Gamma(d)$ for 24 human chromosomes (Chr) as a function of the window distance d (converted to bases by the window size listed in Table I). The distance is represented on a logarithmic scale. (a) Chr 1–6; (b) Chr 7–12; (c) Chr 13–18; and (d) Chr 19–22, ChrX, and ChrY.

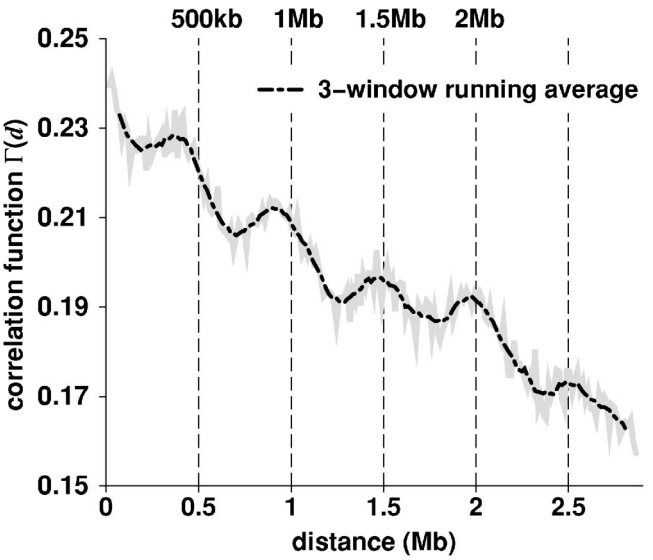

FIG. 8. Correlation function $\Gamma(d)$, as well as a three-window moving average, for human chromosome 21 as a function of the window distance d (converted to bases by the window size given in Table I). The oscillation in $\Gamma(d)$ is highlighted by vertical lines, indicating the distances of $d=500$ kb, 1 Mb, 1.5 Mb, and 2 Mb.

correlation function $\Gamma(d)>0.1$ at $d=1$ Mb for all other chromosomes. The low correlation in ChrY is due to the fact that about half of the bases are unsequenced, and the substitution of gaps by random values lowers correlation. At even longer distances such as $d=10$ Mb, correlations $\Gamma(d=10\ \text{Mb})$ for chromosomes 1 and 6 are still above the 0.1 level.

Given different windows (w) due to different chromosome sizes and provided that the covariance of GC% is approximately independent of w, a scaling of the variance according to $1/w^{\beta}$ implies that the correlation function $\Gamma(d)$ in Eq. (3) increases with the window size as $\sim w^{\beta}$. Test calculations of the covariance for 2^{15} and 2^{17} windows show that the covariance differs by less than 1% (and hence is fairly independent in this range of window sizes). Yet for a detailed comparison of correlation functions calculated for different chromosomes one might have to take into account different windows sizes.

Any deviation from the monotonic decrease of $\Gamma(d)$ might be indicative of correlations at characteristic length scales (visible as "bumps"). For example, Fig. 7 shows for chromosome 1 such a bump at $d\approx 21-23$ Mb. Bumps or sharper peaks in other chromosomes include $d\approx 9.3$ Mb (Chr2), 7.2 Mb (Chr10), 3.2–3.8 Mb (Chr12), and 2.4–3.1 Mb (Chr19). One plausible explanation is that for chromosomes 2, 10, 12, and 19 one or few alterations of GC-rich or GC-low isochores [13] with these length scales enhance the correlation.

Chromosome 21 stands out among all human chromosomes for having a comparatively higher correlation at distances of several Mb (despite having a smaller w^{β} factor than other chromosomes due to a smaller window size). A detailed inspection of Fig. 8 uncovers an oscillation of $\Gamma(d)$ of about 500 kb, ranging from $d=500$ kb to 2 Mb. It can be further shown that this oscillation is not due to the substitution of interspersed repeats [33], and it is localized to about one-eighth of the right distal end of chromosome 21 [33].

VII. DISCUSSION

We study spectral components and correlation structures in the set of human chromosomes, using power spectra, coarse-grained correlation functions, and the variance of different window sizes. All three measures are interrelated and highlight compositional structures at different feature levels. Our results firmly establish the presence of long-ranging correlations and $1/f^{\alpha}$ spectra in the DNA sequences of the set of 24 human chromosomes.

Using updated and completed human sequence data, we find the presence of $1/f$ noise in the DNA sequences of all human chromosomes. We further find that, with the exception of chromosome 22, all chromosomes exhibit a crossover from $1/f^{\alpha_1}$ at low frequency to $1/f^{\alpha_2}$ scaling at high frequency $(\alpha_1>\alpha_2)$, in the range of 30–100 kb. The result of two scaling ranges at low and high frequency is in accord with previous findings, obtained from sequence data of lower quality, and it refines break-point regions for each individual chromosome.

We also examined the effect of about 45% interspersed repeats in the human genome. Using a procedure that masks and subsequently substitutes interspersed repeats with random GC% values, we find that interspersed repeats (i) only marginally affect the scaling exponent α_1 in the low-frequency range, but (ii) lower α_2 in the high-frequency range [cf. Fig. 4(b)]. This supports the general understanding that interspersed repeats only contribute to short-ranging (high-frequency) correlations [6].

There are arguments both supporting and against the procedure to remove the interspersed repeats rather than substituting them. Removing interspersed repeats allows us to fo-

cus solely on nonrepetitive regions, without specifically treating arbitrary gap sizes introduced by interspersed repeats. However, if we consider the spacing introduced by interspersed repeats as part of the genome feature, removing repeats would distort this pattern. In either case, the main conclusion of this paper, that $1/f$ noise universally exists in DNA sequences of all human chromosomes, is not affected by the choice of either substituting or removing interspersed repeats.

We have shown elsewhere that $1/f^{\alpha}$ spectra of GC% fluctuation are also universally present in the mouse *Mus musculus* genomic DNA sequences [34]. It is known that human and mouse genomes are separated by approximately 65–75 million years of evolution. Besides the similarity (or homology) between these two genomes on a local scale [35], there is in fact a large amount of reshuffling of the chromosome segments at a global scale when two current-day copies of the two genomes are compared side by side [36]. Since reshuffling of a sequence at global scales could potentially destroy long-range correlations, it is still to be resolved under what conditions a reshuffling of the human genome into the mouse genome, or vice versa, conserves $1/f$ noise.

One possible hypothesis of why $1/f^{\alpha}$ spectra appear in both the human and the mouse genomes is that such long-range patterns were probably generated from ancestral DNA sequences by sequence evolutionary mechanisms. One sequence evolution model, termed the expansion-modification (EM) model, is known to generate $1/f^{\alpha}$ spectra [7]. The EM model incorporates duplications and mutations. Since the duplication process is an essential element in evolutionary genomics [37], whose role is perhaps as important as Darwin's natural selection [38], even a yet unsophisticated incorporation of duplications in the EM model may capture the essence of the evolutionary origin of long-range correlations in DNA sequences. In the EM model, only the duplication of segments with the same length scale is included, whereas in reality segments with a broad range of length scales are duplicated [19].

One frequently posed question concerns the "biological meaning" of $1/f^{\alpha}$ spectra or long-range correlations in DNA sequences. In order to address this question, one may ask two related questions beforehand. Does the compositional GC% have any biological effects? What biological functions of the DNA molecule are of relevance? From the *functional genomics* perspective, interesting biological processes related to DNA molecules include transcription, replication, and recombination, and their potential connection to GC% has been reviewed in [27,39,40]. Generally speaking, GC% has a statistical association with all three processes, though the cause-and-effect role has not yet been firmly established. Recent studies show that broadly expressed "housekeeping genes" tend to be located in GC-rich regions [41]. To understand the genome-wide organization of biological units that play a role in those processes (e.g., genes, origins and timing of replication, or recombination hotspots), at times it is more feasible to directly study the spatial distribution of functional units instead of using the GC% as a surrogate.

From the *biophysics and cellular biology* perspective, GC% is linked with bands from chromosome staining [42], and in addition, possibly with the matrix or scaffold attachment-associated regions located at the end of DNA loops [43]. It has also been suggested that GC-rich chromosomes (or regions) tend to be located in the interior of the nucleus during interphase and are more "open" in their tertiary structure, whereas GC-poor segments are more likely to be close to the surface of the nucleus and more condensed [44].

Further exploration of the relationship between GC% fluctuations, as well as their large-scale patterns, and the above biological processes is beyond the scope of this paper. An attempt for bacterial genomes has been made to relate the scale-invariance feature in sequence statistics to the genome organization of transcription activities [29]. It is clear that more integrated computational and experimental analyses need to be carried out along similar lines before one can give universal $1/f$ spectra in DNA sequences a satisfactory biological explanation.

ACKNOWLEDGMENTS

We thank S. Guharay for participating at an earlier stage of this project, as well as O. Clay, J. L. Oliver, and A. Fukushima for valuable discussions.

[1] J. B. Johnson, Phys. Rev. **26**, 71 (1925).
[2] A. van der Ziel, Adv. Electron. Electron Phys. **49**, 225 (1979); P. Dutta and P. M. Horn, Rev. Mod. Phys. **53**, 497 (1981); M. B. Weissman, *ibid.* **60**, 537 (1988); H. Wong, Microelectron. Reliab. **43**, 585 (2003).
[3] M. Gardner, Sci. Am. **238**(4), 16 (1978); W. Press, Comments. Astrophys. **7**,103 (1978); B. J. West and M. F. Shlesinger, Am. Sci. **78**, 40 (1990); E. Milotti, e-print physics/0204033.
[4] W. Li, http://www.nslij-genetics.org/wli/1fnoise/
[5] H. Herzel and I. Grosse, Physica A **216**, 518 (1995); S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[6] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001); D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, *ibid.* **67**, 061913 (2003).
[7] W. Li, Europhys. Lett. **10**, 395 (1989); Phys. Rev. A **43**, 5240 (1991).
[8] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[9] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[10] X. Lu, Z. Sun, H. Chen, and Y. Li, Phys. Rev. E **58**, 3578 (1998); M. de Sousa Vieira, *ibid.* **60**, 5932 (1999).
[11] W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver, Genome Res. **8**, 916 (1998).
[12] A. Fukushima, Ph.D. thesis, Nara Institute of Science and

Technology, 2003; A. Fukushima, T. Ikemura, M. Kinouchi, T. Oshima, Y. Kudo, H. Mori, and S. Kanaya, Gene **300**, 203 (2002).
[13] G. Cuny, P. Soriano, G. Macaya, and G. Bernardi, Eur. J. Biochem. **115**, 227 (1981); G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier, Science **228**, 953 (1985); G. Bernardi, Gene **241**, 3 (2000).
[14] O. Clay, N. Carels, C. Douady, G. Macaya, and G. Bernardi, Gene **276**, 15 (2001); O. Clay and G. Bernardi, *ibid.* **276**, 25 (2001).
[15] W. Li, P. Bernaola-Galván, P. Carpena, and J. L. Oliver, Comput. Biol. Chem. **27**, 5 (2003).
[16] O. Clay, Gene **276**, 33 (2001).
[17] W. Li, Gene **300**, 129 (2002).
[18] W. Li, Complexity **3**, 33 (1997).
[19] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[20] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Gene **300**, 105 (2002).
[21] Z.Abellah *et al.*, Nature (London) **431**, 931 (2004); J. Schmutz *et al.*, *ibid.* **429**, 365 (2004).
[22] A. F. A. Smit and P. Green, computer code REPEATMASKER, University of Washington, available at http://repeatmasker.genome.washington.edu/
[23] P. J. Daniell, Suppl. J. R. Stat. Soc. **8**, 88 (1946).
[24] J. R. Korenberg and M. C. Rykowski, Cell **53**, 391 (1988); P. Medstrand, L. N. van de Lagemaat, and D. L. Mager, Genome Res. **12**, 1483 (2002); M.-A. Hakimi, D. A. Bochar, J. A. Schmiesing, Y. Dong, O. G. Barak, D. W. Speicher, K. Yokomori, and R. Shiekhattar, Nature (London) **418**, 994 (2002); J. S. Han, S. T. Szak, and J. D. Boeke, *ibid.* **429**, 268 (2004).
[25] O. Clay, C. J. Douady, N. Carels, S. Hughes, G. Bucciarelli, and G. Bernardi, Eur. Biophys. J. **32**, 418 (2003).
[26] G. Macaya, J. P. Thiery, and G. Bernardi, J. Mol. Biol. **108**, 237 (1976).
[27] W. Li, *Progress in Bioinformatics* (Nova Science, Hauppauge, in press).
[28] J. Beran, *Statistics for Long-Memory Processes* (Chapman and Hall, London, 1994).
[29] B. Audit and C. A. Ouzounis, J. Mol. Biol. **332**, 617 (2003).
[30] P. Bernaola-Galván, J. L. Oliver, P. Carpena, O. Clay, and G. Bernardi, Gene **333**, 121 (2004).
[31] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 7th ed. (Iowa State University Press, Ames, 1980).
[32] R. F. Voss and J. Clarke, Nature (London) **258**, 317 (1975); K. J. Hsu and A. Hsu, Proc. Natl. Acad. Sci. U.S.A. **88**, 3507 (1991).
[33] W. Li and D. Holste, Comput. Biol. Chem. **28**, 393 (2004).
[34] W. Li and D. Holste, Fluct. Noise Lett. **4**, L453 (2004).
[35] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler, Science **304**, 1321 (2004).
[36] P. Pevzner and G. Tesler, Genome Res. **13**, 37 (2003).
[37] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).
[38] A. Meyer and Y. van de Peer, J. Struct. Funct. Genomics **3**, vii (2003).
[39] G. Bernardi, Annu. Rev. Genet. **23**, 637 (1989); **29**, 445 (1995).
[40] G. Bernardi, *Structural and Evolutionary Genomics* (Elsevier, Amsterdam, 2004).
[41] M. J. Lercher, A. O. Urrutia, and L. D. Hurst, Nat. Genet. **31**, 180 (2002); M. J. Lercher, A. O. Urrutia, A. Pavlicek, and L. D. Hurst, Hum. Mol. Genet. **12**, 2411 (2003); R. Versteeg, B. D. van Schaik, M. F. van Batenburg, M. Roos, R. Monajemi, H. Caron, H. J. Bussemaker, and A. H. van Kampen, Genome Res. **13**, 1998 (2003).
[42] Y. Niimura and T. Gojobori, Proc. Natl. Acad. Sci. U.S.A. **99**, 797 (2002).
[43] P. A. Dijkwel and J. L. Hamlin, Int. Rev. Cytol. **162A**, 455 (1995); S. V. Razin, I. I. Gromova, and O. V. Iarovaia, *ibid.* **162B**, 405 (1995).
[44] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore, Hum. Mol. Genet. **10**, 211 (2001); S. Saccone, C. Federico, and G. Bernardi, Gene **300**, 169 (2002).
[45] Human Genome browser, available at http://genome.ucsc.edu/