Mutations along human chromosomes: How randomly scattered are they?

José-Angel Oteo[*]
*Departament de Física Teòrica, Universitat de València, 46100 Burjassot, Valencia, Spain
and Institute for Integrative Systems Biology, 46980 Paterna, Valencia, Spain*

Gonzalo Oteo-García[†]
Department of Chemistry, Life Sciences and Environmental Sustainability, Università di Parma, 43121 Parma, Italy

(Received 16 June 2022; revised 19 September 2022; accepted 6 November 2022; published 7 December 2022)

The diversity of mutations in human chromosomes is nowadays very well documented. The mutations characterize populations in the world as well as genetic causes of diseases. In the approach that we follow, we study the patterns of gaps between mutations by means of the rescaled range analysis and the fractal dimension estimates. The results for chromosomes 1 to 22 and X indicate the existence of the so-called Hurst phenomenon in all of them. The interpretation of this outcome entails the presence of long-range correlations and we propose an explanation based on the genomic feature dubbed linkage disequilibrium, a nonrandom association of alleles at different loci. An unexpected outcome is the noteworthy uniform reduction in the Hurst phenomenon when considering the centimorgan metric instead of base position units. By contrast, such uniform reduction is not observed with the fractal dimension values.

DOI: 10.1103/PhysRevE.106.064404

I. INTRODUCTION

Genetic mutations are generally assumed to be a random phenomenon. The vast majority of them are considered to be neutral and do not have an impact on biological fitness of the organism. This is in opposition to the small fraction of beneficial and deleterious mutations that will be favored or purged through selection mechanisms. Genetic drift is another important mechanism by which genetic variants increase or decrease frequency over time, eventually getting fixed or lost. Based on minimum allele frequency (MAF) criteria, mutations are labeled as common (MAF greater than 5%), low-frequency (MAF between 1% and 5%), and rare (MAF less than 1%) [1,2].

There exists presently a large catalog of successful mutations in human chromosomes [3]. They include insertions, deletions, duplications, copy-number variants, inversions, translocations, and single-nucleotide polymorphism (SNP) [4,5]. When the region affected is longer than 50 base positions (BPs) it is typically referred to as structural variation. When the region affected consists of only one base pair, the variant is classified under the SNP category [6,7]. The catalog of SNP we are using has been generated based on a comparison against a consensus DNA sequence represented by a human reference genome [8] for which various versions are currently in use (GRCh37 and GRCh38) [9]. Variants can be defined by comparing any other sequence against the reference. As a rule of thumb, there is one SNP per thousand bases in human chromosomes. The SNP catalog is built up from DNA sequences provided by a large variety of human population samples (approximately 2500 from 26 populations). For instance, world populations are characterized by the allele frequencies of particular subsets of SNP. Moreover, they serve as identifying characteristics to track the history of human evolution [7] and in the process of identifying genetic elements which are responsible for diseases [10,11].

There are chromosome regions, termed hot spots, where SNPs tend to be observed. Other regions are SNP-free and some specific sites cannot afford mutations at all, for it would compromise essential biological functions. Thus, the probability of finding mutations along the chromosomes cannot be purely random. An instance of SNP distribution is given in Fig. 1, for chromosome 21, with two different metrics (to be explained below) in Figs. 1(a) and 1(b). Every line stands for a mutation. In Fig. 1(c) the lines are uniformly and randomly distributed for the sake of comparison. The three plots have the same number of lines in the space allotted. Compared to the random pattern, the distribution of SNP presents compelling differences in both metrics. Our goal is to quantify and to provide an interpretation of the degree of correlation of SNP locations in the human chromosomes 1 to 22 and X.

The mutation sequences we are analyzing are data structures dubbed a genetic map. A single individual exhibits only a subset of those mutations whose cardinal number is small if the individual is genetically close to the consensus sequence of the catalog and larger otherwise.

We analyze the data provided by the 1000 Genome Project [9] with some filters applied. The most direct approach consists in using the series of intervals between mutations as given by nucleotide BPs. Note that this separation has no real bearing on the physical or spatial distance between nucleotides in the chromosome. In contrast, the study of

[*] oteo@uv.es
[†] gonzalo.oteogarcia@unipr.it

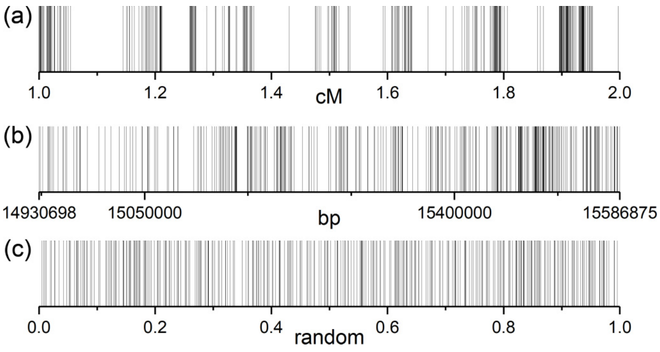

FIG. 1. Chromosome 21. (a) Location of consecutive SNPs between 1 and 2 cM. (b) Same SNP subset as in (a) referring to their site number in the chromosome. (c) Uniform random distribution sample of the same number of SNPs. The three plots have the same number of lines in the space allotted.

chromosome meiotic recombination has introduced a more elaborate measure of genetic distance between SNPs in terms of the so-called centimorgan (cM) units. Given two different sites in a chromosome, 1 cM corresponds to 1% probability that it approximately breaks between those two sites during meiotic recombination. The key point is that although a meiotic recombination fraction can be experimentally estimated, it does not increase linearly with the separation between BPs, which prevents it from being a true genetic distance. However, a genetic distance d expressed in centimorgans is additive, namely, $d(AC) = d(AB) + d(BC)$, for three given sites, unlike recombination probabilities. The relationship between centimorgans and recombination probability is not linear, but logarithmic. A very elegant and short formulation of the issue was given long ago by Kosambi [12,13].

We study the SNP interval sequences using the so-called rescaled range R/S method and the fractal dimension [14]. The analyses are carried out first in terms of a series of SNP intervals measured in BPs and second using the SNP intervals given in centimorgan units.

The R/S analysis, when successful, provides an index $0 < H < 1$ whose value has an interesting interpretation. When the series under scrutiny originates from a simple random process, $H = 0.5$. Values $H \neq 0.5$ indicate the presence of long-range correlations in the series. Whenever $H > 0.5$, the series is said to be persistent, which means that large (small) values are most likely to be followed by large (small) values in the sequence. The case $H < 0.5$ is called antipersistent and presents the opposite behavior. This tool was introduced by Hurst [15] in the context of hydrology where it was found that a number of phenomena in nature (recorded as numerical series) yielded $H \simeq 0.73$, which is referred to as the Hurst phenomenon in the literature and H as Hurst exponent [14]. The mathematical origin of the Hurst phenomenon has intrigued mathematicians for years [16–18]. The crux is that although the distribution of intervals between SNPs on its own may follow a probability law of a random process, the distribution when the order of occurrence of these intervals becomes a factor cannot be one of simple random probability [16].

We present results for the Hurst exponent associated with mutations in human chromosomes 1 to 22 and X. The analysis in terms of nucleotide BPs indicates that all chromosomes exhibit the Hurst phenomenon. When the analysis is made in terms of centimorgan units, the value of H decreases significantly but still shows persistent character. An interpretation of these results is provided and a possible origin of the Hurst phenomenon in SNP patterns is suggested.

The computation of the fractal dimension D associated with the SNP sequences has been carried out by a procedure due to Higuchi [19]. Values $D \simeq 1$ are associated with regular curves and $D \simeq 1.5$ with pure noise. The D outcomes show less conclusive results than H, in both BP and centimorgan representations.

For the sake of completeness we briefly explain the R/S method in Sec. II and the relationship between centimorgan units and meiotic recombination probabilities in Sec. III. The latter is intended to clarify the differences between the two chromosome descriptions we use with the R/S analysis. Section IV gathers some statistical features of the chromosomes. In Secs. V and VI the results of the analyses in terms of BP and centimorgan units are given. A discussion and interpretation of the outcomes are in Sec. VII.

The data sets we use are available at the 1000 Genomes Project [3] for every human chromosome 1 to 22 and X, in ASCII format [20,21]. We have discarded mutations that are found in the populations with MAF less than 1%. Namely, we keep mutations with common and low-frequency allele frequencies. This is customary in some genetic studies and, in a way, successful mutations are defined by this threshold. Given a chromosome, the corresponding data file contains the site number where the SNP takes place and the genetic distance in centimorgans refers to the origin.

II. RESCALED RANGE ANALYSIS

Given a nucleotide site i in a chromosome, let $\xi_i \geqslant 0$ define the gap between two adjacent SNPs, either in BP or in centimorgan units. The index i runs from 1 to N, where $N + 1$ stands for the total number of SNPs.

We commence by considering a contiguous subsequence of length n inside the full sequence $\{\xi_i\}_1^N$. The computation we are describing will be replicated systematically for nonoverlapping SNP interval windows yielding in this way average values. The mean interval length between adjacent SNPs in the window is then

$$\langle \xi \rangle_n = \frac{1}{n} \sum_{k=1}^{n} \xi_k. \tag{1}$$

Following Feder [14], for fix n we calculate the series $\{X(i,n)\}_{i=1}^n$ of accumulated departures of adjacent SNP gaps from the mean SNP separation in the window

$$X(i,n) = \sum_{k=1}^{i} (\xi_k - \langle \xi \rangle_n), \quad i \leqslant n, \tag{2}$$

and then compute its range

$$R(n) = \max_{1\leqslant i\leqslant n} X(i,n) - \min_{1\leqslant i\leqslant n} X(i,n). \tag{3}$$

To work with a dimensionless quantity, Hurst introduced the ratio R/S, the rescaled range, with S the standard deviation estimated from the n observations in the window

$$S(n) = \left[\frac{1}{n}\sum_{k=1}^{n}(\xi_k - \langle\xi\rangle_n)^2\right]^{1/2}, \tag{4}$$

and observed that for a number of real data records the rescaled range behaves as

$$R/S \equiv \langle R(n)\rangle/\langle S(n)\rangle = (n/\alpha)^H, \quad \alpha > 0, \tag{5}$$

with the Hurst exponent $H = 0.73 \pm 0.09$. Here $\langle R(n)\rangle$ and $\langle S(n)\rangle$ are averages of $R(n)$ and $S(n)$, respectively, for N/n nonoverlapping windows of length n.

In contrast, for records generated by statistically independent processes with finite variance, the rescaled range R/S behaves asymptotically as

$$R/S = (\pi n/2)^{1/2}. \tag{6}$$

A simple way to estimate H from data is as the slope of the linear regression of $\log(R/S)$ vs the window size $\log n$: $\log(R/S) = H\log n - H\log\alpha$. The constant α is characteristic of the data set at hand and has not received particular attention in the literature.

III. KOSAMBI'S CENTIMORGAN TRANSFORMATION

Given three consecutive loci a, b, and c, let y_1, y_2, and y_3 stand for the probabilities the chromosome breaks during meiosis in the intervals (a, b), (b, c), and (a, c), respectively, with $0 \leqslant y_i \leqslant 0.5$. In a meiotic process the recombination fractions y_i are not true additive distances. A way to map recombination probabilities into a genetic distance is explained next. Following Kosambi [12], the experimental evidence at that time was $y_3 \simeq y_1 + y_2$, only for small enough values y_i, whereas for intermediate values $y_3 \simeq y_1 + y_2 - y_1y_2$ and for large values $y_3 \simeq y_1 + y_2 - 2y_1y_2$. Kosambi posed the question of finding a single formula for the full range of values $0 \leqslant y_i \leqslant 0.5$, in the form

$$y_3 = y_1 + y_2 - \beta y_1 y_2, \tag{7}$$

with β to be specified. The idea [12] is to introduce a new variable x that depends continuously on y, $y = f(x)$, and be a true additive distance $x_3 = x_1 + x_2$. In addition, whenever x is small enough $f(x) \simeq x$. It is assumed that f is independent of the chromosome and of the site. Equation (7) reads then

$$f(x+h) = f(x) + f(h) - \beta f(x)f(h). \tag{8}$$

In the limit $h \to 0$, $f(h) \sim h$ and we get the differential equation

$$f'(x) = 1 - \beta f(x), \tag{9}$$

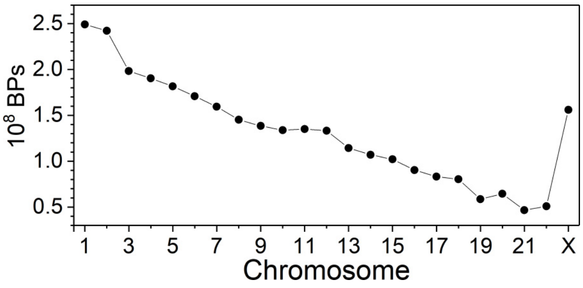

FIG. 2. Chromosome lengths in BPs.

i.e.,

$$\frac{dy}{dx} = 1 - \beta y, \quad y(0) = 0. \tag{10}$$

The next move requires an assumption about β. Kosambi proposed the simple choice $\beta = 4y$, which encompasses the cases above: $\beta \to 0$ for small values of y, $\beta \to 2$ for large ones since y cannot exceed 0.5, and $\beta \to 1$ for intermediate y, recovering the phenomenology at that time from small, medium, and large recombination fractions. The differential equation becomes

$$\frac{dy}{dx} = 1 - 4y^2, \quad y(0) = 0 \tag{11}$$

and is readily integrated

$$y = \frac{1}{2}\tanh(2x), \quad x = \frac{1}{4}\ln\frac{1+2y}{1-2y}. \tag{12}$$

This function maps the experimentally measured recombination probabilities $y \in [0, 0.5)$ into the new variable $x \in [0, \infty)$. By convention, $100x$ is expressed in centimorgans units.

IV. SOME FEATURES OF THE GENETIC MAP

The following is a brief description of the 23-chromosome genetic map that we are analyzing. The chromosome lengths, which vary between 50×10^6 and 250×10^6 BPs, are given in Fig. 2. Figure 3 presents the SNP genetic distance in centimorgans from a chromosome physical origin as a function of the SNP number. The detail shown for chromosome 1 in the inset may be found everywhere for all the curves. The irregular growth is reminiscent of the devil staircase shape [14] characteristic of fractal structures.

Next we present some results concerning the statistics of SNP gaps in chromosomes, in both BP and centimorgan units. Their distributions in the genomic map are given in Figs. 4 and 5, both in log-log scales, in BP and centimorgan units, respectively. The 23 curves in BP units do not collapse into a universal curve, not even when they are rescaled with respect to their own average gap length. The X-chromosome curve exhibits a shape slightly different from the rest, a feature that will reappear below in the R/S analysis outcomes. The SNP gap range from 1 to 10 BPs presents a differentiated trend in all curves which could be interpreted as a kind of mutation repulsion in the BP representation.

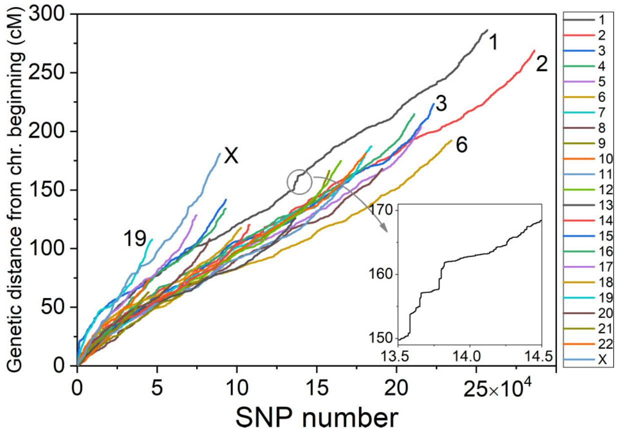

FIG. 3. SNP distance from the chromosome origin (in centimorgans). Only some chromosome curves are labeled for the sake of clarity. The inset shows the devil staircase–like shape of the chromosome 1 curve.

The different curves in centimorgan units seem to display power laws, including an approximately common crossover around 0.001 cM. However, the explicit interpretation of that crossover scale is not easy. In addition, the ranges produced by these power laws are short.

In order to contrast the presence of correlations in the SNP distributions we generated surrogate chromosomes in which the gap sequence $\{\xi_i\}$ has been replaced with $\{\xi_{\sigma(i)}\}$, where σ stands for a random permutation of the indices. The resulting sequences have the same mean and higher moments than the originals; however, the eventual ordering correlations of the sequence have been dropped out. Had all the correlations detectable by H and/or D been removed, then $H \simeq 0.5$ and/or $D \simeq 1.5$. The presence of correlations may be explicitly illustrated computing the ratios $\{r_k\}_1^{N-1}$ of two consecutive SNP gaps, in BP units, as

$$r_k = \min\{\xi_k, \xi_{k+1}\}/\max\{\xi_k, \xi_{k+1}\} \in [0, 1] \qquad (13)$$

for chromosomes 1 to 22 and X and then the corresponding probability distribution, say, $P(r)$. The same computation is carried out for surrogate data. Figure 6 shows the differences between both estimates. The proportions of many unequal contiguous pairs of SNP gaps, say, $r < 0.2$, are much higher in shuffled gaps. In the balance, the presence of similar pairs of gaps is more frequent in real chromosomes. Thus, real chromosomes exhibit higher diversity in ratios of pairs of gaps with respect to the versions of surrogate data.

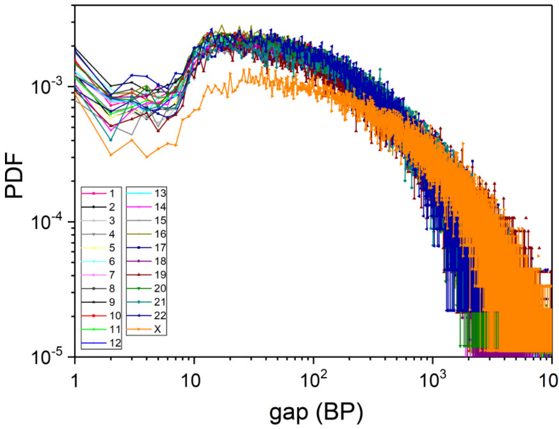

FIG. 4. Estimated distribution of SNP gaps in BP units. Here PDF denotes probability distribution function.

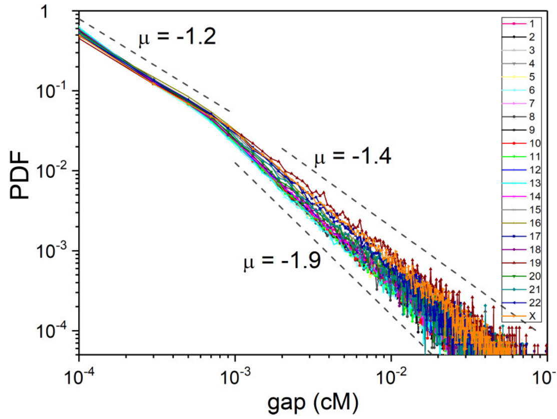

FIG. 5. Estimated distribution of SNP gaps in centimorgan units. Here μ stands for the slope of the dashed lines, given for visual reference.

V. THE *R*/*S* ANALYSIS OUTCOMES

We have carried out the R/S analysis with both real and surrogate chromosomes. Figure 7 shows in log-log scale the outcomes of the power law (5). The results of the two sequence representation types are given, with SNP gaps in (i) BPs (closed symbols) and (ii) centimorgans (open symbols, with triangles for shuffled sequences). The dashed line has a slope of $\frac{1}{2}$ and corresponds to Eq. (6). The first observation is that all chromosome sequences, given either in centimorgans or in BPs, exhibit a similar slope. The Hurst exponents obtained by a linear fit to pooled data are $H(\text{BP}) = 0.784(3)$ and $H(\text{cM}) = 0.636(2)$. For the corresponding surrogate sequences, $H_s(\text{BP}) = 0.515(1)$ and $H_s(\text{cM}) = 0.526(1)$, very close to a pure random pattern. The specific Hurst exponents for every chromosome are in Fig. 8. Every slope is obtained from a linear fit to a set of seven to nine points, depending on the chromosome length. The mean values of the 23 estimated exponents are $\langle H(\text{BP})\rangle = 0.769$, with standard

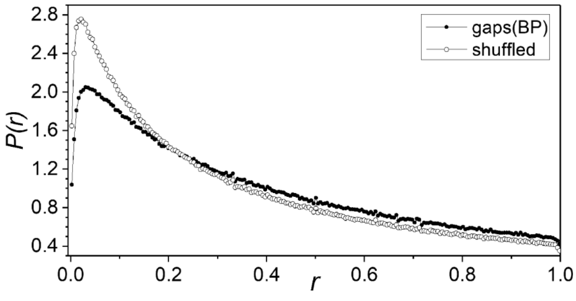

FIG. 6. Probability distribution for the ratio (13) of adjacent SNP gaps measured in BPs, estimated from chromosomes 1 to 22 and X (closed circles) and from surrogate chromosomes (open circles).

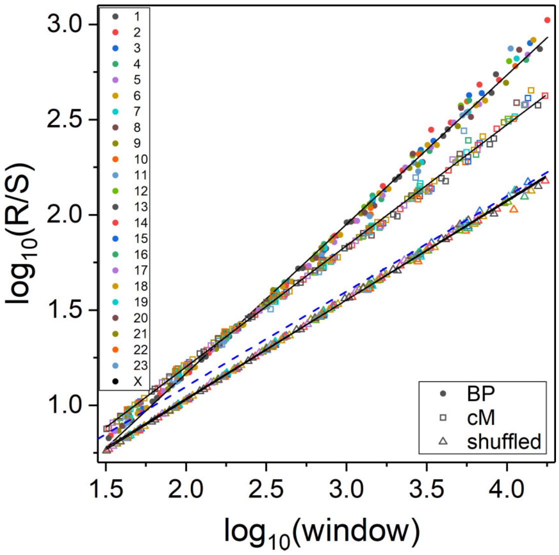

FIG. 7. Plot of the R/S analysis for chromosomes 1 to 22 and X from SNP gaps in centimorgans (open squares) and BPs (closed circles). Triangles stand for surrogate chromosomes. Color codes the chromosome. The dashed line is for Eq. (6). Solid lines are best linear fits to pooled data.

deviation (s.d.) equal to 0.019, and $\langle H(\mathrm{BP})\rangle = 0.636$, with s.d. equal to 0.016.

The main observation is that the mapping from BP-based separation to centimorgan-based genetic distance reduces the correlation among SNP gaps along the chromosomes, albeit not completely. Thus, the Hurst phenomenon exhibited by the SNP gap sequences in BPs is depleted when the R/S analysis is done in terms of centimorgans. All this said, it is interesting to observe the peculiarity of the X chromosome whose H value decrease is the shortest one. A possible explanation is that the X chromosome is somewhat different for the way it takes part in meiosis. Recombination of the X chromosome only occurs in females. Thus, it experiences half the amount of recombination with respect to the rest. The lower share in meiosis might have lead to a different degree of statistical correlations.

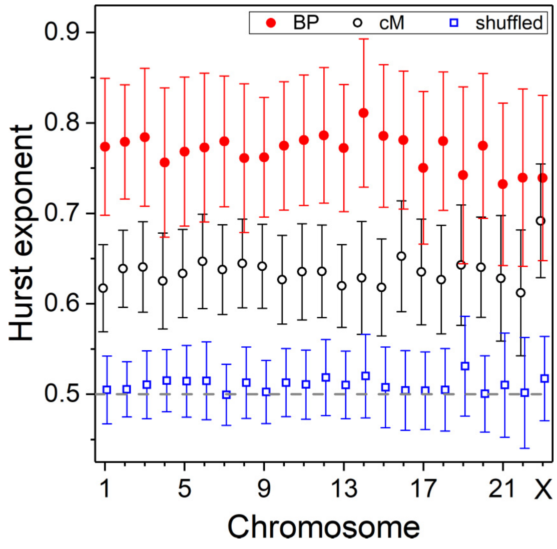

FIG. 8. Hurst exponents for chromosomes 1 to 22 and X from SNP gaps in centimorgans (open circles) and BPs (closed circles). Open squares stand for surrogates chromosomes. Error bars are 95% CI.

No relationship is observed between the chromosome length and the Hurst exponent.

VI. FRACTAL DIMENSION ANALYSIS OUTCOMES

The computation of the fractal dimension D associated with a time series, or in general to a sequence of points $\{i, x(i)\}$, $i = 1, 2, \dots, n$, allows a classification of the complexity of that signal and may provide interesting information. The D index is sometimes interpreted as a measure of the roughness of the pattern.

We have carried out the D computation for the SNP gap sequences of chromosomes following a technique by Higuchi [19,22]. The idea consists in measuring the cumulative absolute distance reached by the curve on the ordinate axis. The length is measured on point subsets of the curve which are k units apart from each other, namely, $x(m), x(m+k), x(m+2k), \dots$. The explicit formula that defines the length of the curve, with m and k given, reads

$$L_m(k) = \frac{n-1}{k^2\left\lceil\frac{n-m}{k}\right\rceil}\sum_{i=1}^{\lceil (n-m)/k\rceil} |x(m+ik) - x[m+(i-1)k]|, \tag{14}$$

with n the total number of points. The length is then averaged over all possible initial values m to give $\langle L(k)\rangle$. Then, if $\langle L(k)\rangle \propto k^{-D}$, the curve is fractal with dimension D. For instance, a curve obtained with pure noise should give $D = 1.5$, whereas for a regular curve $D = 1$. Under the hypothesis of self-affinity of the pattern, there is a simple algebraic relationship between the Hurst exponent and the fractal dimension [14,23]

$$H + D = 2. \tag{15}$$

In Fig. 9 we have plotted in doubly logarithmic scale $\langle L(k)\rangle$ against k, both in BP and centimorgan units. The higher (lower) points in every column set correspond to the longer (shorter) chromosomes. Straight solid lines are linear fits to the pooled chromosome data. The pluses and crosses stand for shuffled data and have been horizontally shifted by a factor 2 for clarity. The dashed lines are linear fits to these two sets. The fractal dimension obtained for every single chromosome by a linear fit is in Fig. 10. Closed and open symbols are for real and surrogate chromosomes, respectively. The average fractal dimension for the 23 chromosomes reads $\langle D(\mathrm{BP})\rangle = 1.18$ with s.d. equal to 0.05 and $\langle D(\mathrm{cM})\rangle = 1.20$ with s.d. equal to 0.02. Unlike the Hurst exponent, these results do not exhibit a uniform change when considering the chromosomes individually.

For real chromosomes, if we use the mean Hurst exponent values estimated above, then the algebraic identity (15) is approximately preserved in the BP case, whereas it is not in centimorgan units. This is illustrated in Fig. 10, where the black and red dashed lines stand for the inferred fractal dimensions $2 - \langle H\rangle$. The centimorgan representation is then a

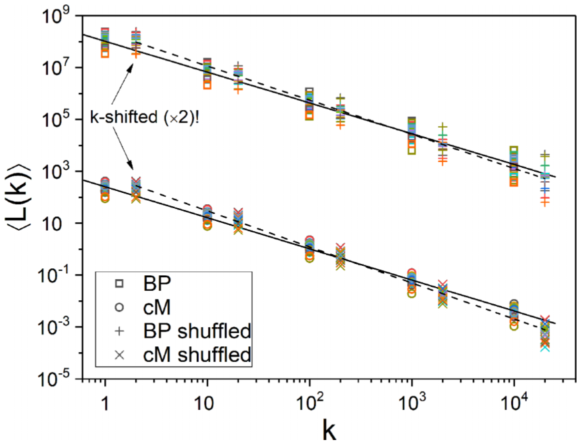

FIG. 9. A log-log plot to determine the fractal dimension of SNP gap sequences in BP (squares) and centimorgan (circles) units. Pluses and crosses are for shuffled data and have been horizontally shifted by a factor 2 for more clarity. Every color stands for a different chromosome (legend not provided). Solid and dashed lines are linear fits to pooled data and to one run of shuffled data, respectively.

real data instance where the Hurst index and fractal dimension separate, a case thoroughly discussed and illustrated in [23] for a Gaussian random process of the Cauchy class.

For surrogate chromosomes the value $D = 1.5$, characteristic of pure noise, is not reached. The points in Fig. 10 are averages of a number of shuffled data runs. The 23 chromosomes' averaged fractal dimensions are $\langle D_s(\mathrm{BP})\rangle = 1.31$ with s.d. equal to 0.02 and $\langle D_s(\mathrm{cM})\rangle = 1.39$ with s.d. equal to 0.01. Thus, although the ordering correlations have been eliminated by shuffling, others remain detected by D.

VII. DISCUSSION

The direct study of DNA segments by a rescaled range, or by related methods, has received attention in the past [24–26]. Note, however, that the analyses of DNA sequences and of single-nucleotide polymorphisms are completely different from each other. The former is carried out on a symbolic series made up of nucleotide symbols {A,C,G,T} which are mapped on numbers to allow the analysis. The latter are the mutations that punctuate DNA sequences which give rise to the numeric sequences of gap lengths. Interestingly, the R/S analysis of DNA segments carried out in [26] provides the value $H = 0.6145$ for a DNA strand of chromosome 20, indicating persistence in the sequence. The origin of the correlations is attributed to the presence of the so-called motifs and Alu repeats in the DNA sequence. An issue these works have addressed concerns the possible characterization of coding and noncoding regions by means of this type of analysis. It could be interesting to R/S analyze segments of the genetic map corresponding to coding genome regions. However, coding DNA strands for only 3% of genome and mutations takes place one every 1000 BPs, which makes the length of records too short for the analysis.

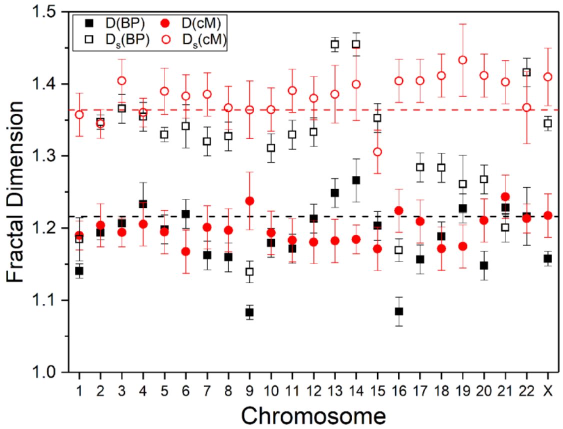

FIG. 10. Estimates of fractal dimension for every chromosome SNP gap sequence given in BP (squares) and centimorgans units (circles). Open symbols are for shuffled data. Dashed horizontal lines are the D estimates obtained from the algebraic relationship (15) using the Hurst exponent estimates as input. Error bars are 95% CI.

Here we have studied the distribution of mutations in chromosomes collected from a large number of human populations [3]. The outcomes indicate the existence of the Hurst phenomenon in the human genetic map. When the separation between mutations is computed in terms of BPs, $\langle H(\mathrm{BP})\rangle = 0.78$. Using the genetic distance in terms of centimorgans, the Hurst exponent is uniformly depleted to $\langle H(\mathrm{cM})\rangle = 0.64$, which amounts a meaningful correlation reduction. This is an unforeseen effect of the mathematical centimorgan mapping, which was originally devised to merely provide a true genetic distance measure in chromosomes, because BP counts and recombination probabilities are inappropriate.

The complexity ranking provided by D is more difficult to interpret than H in the case at hand. The estimated mean values $\langle D(\mathrm{BP})\rangle$ and $\langle D(\mathrm{cM})\rangle$ are quite similar, although they present substantial variability when chromosomes are considered individually. Unlike the uniform decrease of $H(\mathrm{BP})$ to $H(\mathrm{cM})$ for all chromosomes, the variations of $D(\mathrm{BP})$ to $D(\mathrm{cM})$ are quite whimsical. In particular, for 10 out of 23 chromosomes it turns out that $D(\mathrm{BP}) > D(\mathrm{cM})$, which renders the D description of these chromosomes more random in the BP representation than in centimorgan units. In addition, the SNP gap shuffling does not drive the fractal dimension close to $D = 1.5$, the nominal value for noise. The fractal analysis does not disclose the nature of the correlations involved; however, the fact that small scales correspond to fractal dimension whereas large scales are associated with the Hurt coefficient [23] indicates that the centimorgan-mapping effect is different in both scales. This observation is connected with features revealed in Sec. IV, namely, short gap intervals show in BPs a different trend in Fig. 4 and the approximate crossover in Fig. 5 seems also to separate short and long ranges.

The Hurst exponent was historically introduced in a time-series context. A theoretical explanation of the Hurst phenomenon was then given on the basis of a fractional Gaussian process in which the value of a point in the time series is affected by all the precedents [14,27]. However, the SNP gap patterns that we have analyzed come from a physical alignment and there is no preferred sense involved. The question is then to ascertain where the SNP correlations come from. A well-known source of correlations in chromosomes is the linkage disequilibrium mechanism. It originates in the way meiotic recombination takes place. During meiosis homologous chromosomes pair and undergo reciprocal genetic exchange, termed crossover. A pair of nearby SNPs is more likely to stay in the next generation than a distant one because the probability that the chromosome breaks between the nucleotides during meiosis is, as rule of thumb, proportional to the separation. The proportionality is modified by the presence of chromosomal regions with particularly low (jungles) or high (deserts) recombination rates [28]. This is a realization of the linkage disequilibrium effect. The fact that SNPs are not fully random when measured in centimorgans but more random than measured in BPs would suggest that meiotic breaks are a significant but not the only contributor to correlations.

ACKNOWLEDGMENTS

The work of J.-A.O. was partially funded by Project No. PID2019-109592GB-I00/AEI/10.13039/501100011033 from the Spanish Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación. The work of G.O.-G. benefited from the framework of the COMP-HUB Initiative, funded by the Departments of Excellence program of the Italian Ministry for Education, University and Research.

[1] G. Gibson, Rare and common variants: Twenty arguments, Nat. Rev. Genet. **13**, 135 (2012).

[2] L. Bomba, K. Walter, and N. Soranzo, The impact of rare and low-frequency genetic variants in common disease, Genome Biol. **18**, 77 (2017).

[3] 1000 genomes project, https://www.internationalgenome.org/.

[4] L. Feuk, A. R. Carson, and S. W. Scherer, Structural variation in the human genome, Nat. Rev. Genet. **7**, 85 (2006).

[5] C. Alkan, B. P. Coe, and E. E. Eichler, Genome structural variation discovery and genotyping, Nat. Rev. Genet. **12**, 363 (2011).

[6] P. Sudmant, T. Rausch, E. Gardner, R. Handsaker *et al.*, An integrated map of structural variation in 2,504 human genomes, Nature (London) **526**, 75 (2015).

[7] A. Auton, G. Abecasis, D. Altshuler, R. Durbin *et al.*, A global reference for human genetic variation, Nature (London) **526**, 68 (2015).

[8] S. Nurk, S. Koren, A. Rhie, M. Rautiainen *et al.*, The complete sequence of a human genome, Science **376**, 44 (2022).

[9] Human Genome Overview, Genome Reference Consortium, https://www.ncbi.nlm.nih.gov/grc/human.

[10] B. S. Shastry, SNPs in disease gene mapping, medicinal drug development and evolution, J. Hum. Genet. **52**, 871 (2007).

[11] P. D. Thomas and A. Kejariwal, Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects, Proc. Natl. Acad. Sci. USA **101**, 15398 (2004).

[12] D. D. Kosambi, The estimation of map distances from recombination values, Ann. Eugenic. **12**, 172 (1943).

[13] D. D. Kosambi, *D.D. Kosambi: Selected Works in Mathematics and Statistics*, edited by R. Ramaswamy (Springer India, New Delhi, 2016), pp. 125–130

[14] J. Feder, *Fractals* (Plenum, New York, 1988).

[15] H. Hurst, Long-term storage capacity of reservoirs, Trans. Am. Soc. Civ. Eng. **116**, 770 (1951).

[16] J. Sutcliffe, S. Hurst, A. Awadallah, E. Brown, and K. Hamed, Harold Edwin Hurst: The Nile and Egypt, past and future, Hydrol. Sci. J. **61**, 1557 (2016).

[17] P. O'Connell, D. Koutsoyiannis, H. Lins, Y. Markonis, A. Montanari, and T. Cohn, The scientific legacy of Harold Edwin Hurst (1880–1978), Hydrol. Sci. J. **61**, 1571 (2016).

[18] T. Graves, R. Gramacy, N. Watkins, and C. Franzke, A brief history of long memory: Hurst, Mandelbrot and the road to ARFIMA, 1951–1980, Entropy **19**, 437 (2017).

[19] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D **31**, 277 (1988).

[20] Beagle 4.0, https://faculty.washington.edu/browning/beagle/b4_0.html.

[21] Genetic Map, https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/.

[22] L. Liehr and P. Massopust, On the mathematical validity of the Higuchi method, Physica D **402**, 132265 (2020).

[23] T. Gneiting and M. Schlather, Stochastic models that separate fractal dimension and the Hurst effect, SIAM Rev. **46**, 269 (2004).

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

[25] A. Rosas, E. Nogueira, Jr., and J. F. Fontanari, Multifractal analysis of DNA walks and trails, Phys. Rev. E **66**, 061906 (2002).

[26] A. Provata, C. Nicolis, and G. Nicolis, DNA viewed as an out-of-equilibrium structure, Phys. Rev. E **89**, 052105 (2014).

[27] B. B. Mandelbrot and J. W. Van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Rev. **10**, 422 (1968).

[28] A. Yu, C. Zhao, Y. Fan, W. Jang *et al.*, Comparison of human genetic and sequence-based physical maps, Nature (London) **409**, 951 (2001).