POLITECNICO DI TORINO
Repository ISTITUZIONALE

Kullback–Leibler cluster entropy: a proxy of human chromosome complexity

Original
Kullback–Leibler cluster entropy: a proxy of human chromosome complexity / Gandino, F., Ferrero, R., Carbone, A.. - In: JOURNAL OF COMPUTATIONAL SCIENCE. - ISSN 1877-7511. - STAMPA. - 100:(2026). [10.1016/j.jocs.2026.102964]

Availability:
This version is available at: 11583/3015024 since: 2026-08-31T14:39:17Z

Publisher:
Elsevier

Published
DOI:10.1016/j.jocs.2026.102964

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

15 October 2026

Journal of Computational Science 100 (2026) 102964

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Kullback–Leibler Cluster Entropy: A proxy of human chromosome complexity

Filippo Gandino [a], Renato Ferrero [a], Anna Carbone [b,*]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy
[b] Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy

ARTICLE INFO

Keywords:
Information-theoretic measures
Long-range correlation
Kullback–Leibler divergence
Hurst exponent
Human reference genome

ABSTRACT

Long-memory correlation among individual processes and components yield information about how complex systems evolve. Well-established methods to discriminate correlated and anticorrelated sequences (e.g. box-counting (BC), rescaled range (R/S), detrended fluctuation (DFA), and detrended moving average (DMA) analysis) mainly operate via linear regressions of log–log data plots. The Kullback–Leibler cluster entropy (KLCE), a measure of divergence between probability distributions, has been recently proposed to estimate the scaling exponent of long-range correlated sequences. Here, the KLCE is illustrated on the 24 chromosomes of the T2T-CHM13+Y human reference genome, including centromeric regions and short arms last remaining gaps in the GRCh38 reference assembly. It is shown that the Kullback–Leibler cluster entropy can accurately detect such structurally complex nucleotide regions through the local variability of the correlation exponents. To test performance and prove accuracy in detecting the newly added strands, the outcomes are mapped against nucleotide positions of the GRCh38 assembly. Statistical tests are also included to prove KLCE robustness and accuracy against DFA and DMA performances. Potential applications of the Kullback–Leibler cluster divergence in machine-learning and deep-learning pipelines for genomic data are outlined.

1. Introduction

Information-theoretic measures, based on Shannon, Kullback–Leibler, Jensen–Shannon, Rényi entropy and their variants, are extensively adopted to quantify data features with the ultimate goal of achieving comprehensive and accurate description of systems complexity. The difference between entropy at increasing string length N and its asymptotic value ($N \to \infty$) quantifies complexity in relation to causal states and regularities [1,2]. Information-theoretic clustering refers to computational frameworks based on sample-by-sample estimates of cluster measures, related probability distributions and entropy operators [3]. The approach aims at identifying features that naturally emerge from large amounts of data, rather than partitioning into artificial subsets according to some distance criterion [4,5]. Standard clustering methods can be understood in a broad sense as measures of distance between data points (e.g., Euclidean distance, cosine similarity, Pearson's correlation). In contrast, information-theoretic clustering operates via measures of divergence among probability distributions of empirical data and models. As uncertainty increases in the sequence features, entropy and information content change. Hence, information-theoretic measures efficiently fit for the purpose of discovering patterns, classifying features, and quantifying complexity in terms of probability distribution models in a variety of fields [6].

Power-law distributions ($P(x) \propto x^{-\alpha}$) feature random variables x relevant to complex phenomena evolving over space and time. Estimates of the scaling exponent α have become rather common quantifiers of power-law distributed data. The scaling exponent α can be deduced from the slope of the linear regression of data histograms plotted on log–log scales, a procedure requiring caution to ensure accuracy and veracity of the outcomes [7]. Several methods (e.g., Box-Counting (BC), Rescaled range (R/S), Detrended Fluctuation (DFA), Detrended Moving Average (DMA) analysis [8–12]) estimate the scaling exponent by means of least-square regressions. Machine learning and deep learning algorithms have recently been introduced to automatically estimate scaling exponents and self-similarity features [13–17]. Approaches based on distance between empirical data $P(x)$ and reference $Q(x)$ distribution obeying power-laws that best fits the data have been adopted (e.g. the Kolmogorov–Smirnov distance $D = \max_{x \geq x_{\min}} |P(x) - Q(x)|$ [7]). A family of divergences has been introduced that generalizes Kullback–Leibler, Jensen–Shannon and Rényi to quantify the differences between basic probability assignments in

* Corresponding author.
E-mail address: anna.carbone@polito.it (A. Carbone).

https://doi.org/10.1016/j.jocs.2026.102964
Received 2 January 2026; Received in revised form 22 May 2026; Accepted 8 July 2026
Available online 24 July 2026
1877-7503/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

sequences [18,19]. The Kullback–Leibler cluster entropy (KLCE) has been put forward to infer optimal power-law distributions as illustrated on pairs of financial series and synthetic fractional Brownian motions (fBms) with assigned Hurst exponent in [20,21]. The probability distributions $P(x)$ and $Q(x)$ are obtained by ranking elementary clusters generated in the series. Then, the *minimum relative entropy principle* yields the best estimate of the correlation exponent and quantifies deviations from the assumed model.

Prominent examples of complex systems characterized by power-law distributions are found in genomics. Nucleotide patterns, emerging from seemingly random structures and highly complex interplay of nonlinear interactions among multiple heterogeneous components in DNA and RNA sequences, exhibit power-law scaling, self-organization and complex information processing [22]. Scaling and self-similarity reflect features of biological processes (such as duplication, segmentation, unzipping [23–28]), local structural properties (such as composition and mechanical strength [29–34]), frequency distribution of basepair clusters [35,36], segment similarity in pangenome minigraphs [37], rare phenotypic variants [38], gene expression in lung adenocarcinoma [39], departures of codon sequences from equal usage of synonymous codons [40], protein coding segments and transposable elements [41]. Deep learning analysis of scRNA-sequence data involves Kullback–Leibler (KL) divergence to quantify the extent of the error made by assuming a model in place of the correct probability distribution, when individual cell clustering is adopted to reveal subtypes and infer lineages based on intercellular relationships [42–44]. When direct clustering is performed, various error sources (intrinsic noise, high dimensionality, sparsity of scRNA-seq data) might lead to numerically unstable and inaccurate results [42]. Hence, specific numerical strategies must be put in place to tackle the challenges posed by such error sources. Deep clustering methods that employ denoising autoencoders and simultaneously reduce both dimensionality and clustering biases represent a promising direction in the scRNA-seq data research landscape [44]. Iterative data denoising by means of Laplacian smoothing and self-supervised discriminative embedding allows the correct identification of clusters through adaptively defined similarity thresholds [43].

In this work, the Kullback–Leibler cluster entropy [20,21] is applied to long-read sequences (T2T-CHM13+Y) recently assembled by the *Telomere-to-Telomere Consortium* from complete hydatidiform moles (CHM) [45,46]. The 24 complete linear chromosomes include centromeric, ribosomal sequences and euchromatic segmental duplications, highly complex and repetitive nucleotide regions (exceeding 100 kb) missing in the previous version of the human genome reference assembly (GRCh38) [47]. The inclusion of highly dynamic components, related to the chromosome rearrangements relevant to major diseases, offers unprecedented potential to quantify novel feature annotations and achieve a comprehensive description of human complexity. The KLCE approach operates through an adaptive threshold and incorporates a smoothing mechanism with the ability to mitigate bias and noisy issues.

Each nucleotide in the sequences of the T2T-CHM13+Y assembly is assigned a numerical value. The sequences are analyzed in terms of the Kullback–Leibler cluster entropy. Then the *minimum relative entropy* criterion provides accurate estimates of the correlation exponents at varying nucleotide positions and lengths along the 24 chromosomes. Annotated sets of correlation exponents along the sequences are systematically linked to the position of the recently completed genomic strands of the T2T-CHM13+Y and referred to the GRCh38 assembly. The newly added patches exhibit higher correlation exponents, consistently with the high degree of variability, complexity and dynamics compared to the GRCh38 reference genomes.

The manuscript is organized as follows. Section 2 briefly illustrates the main concepts, mathematical relationships, and computational steps of the *Kullback–Leibler Cluster Entropy*. Section 3 describes the main features of the empirical data to be analyzed, i.e. the T2T-CHM13+Y sequences. The RY procedure to map the nucleotide into numerical sequences and the generation of the empirical probability distribution P are briefly illustrated. The model probability distribution Q generated by artificial data is also described. In this work Fractional Brownian paths with different correlation exponents are used as the reference model for generating the model probability distribution. In Section 4 the main outcomes of the relative cluster entropy analysis are illustrated for the 24 chromosomes of the T2T-CHM13+Y. Section 5 reports the numerical validation of the KLCE algorithm and the comparison with the *Detrending Moving Average* and the *Detrending Fluctuation Analysis* algorithms [11,12]. Comments and future directions of the work are presented in Section 6 (*Discussion and Conclusions*).

2. Method

Consider a long-range correlated numerical sequence $y(x_i)$ of length N and its local average

$$\widetilde{y}(x_i, n) = \frac{1}{n}\sum_{n'=0}^{n-1} y(x(i-n')) \tag{1}$$

of length $N-n$, with $n \in (1, N)$. For each n, a partition $\{C\}$ is generated by the consecutive intersections of $\{y_i\}$ and $\{\widetilde{y}_{i,n}\}$ defined by the indices i where the difference $\epsilon_{i,n} = y_i - \widetilde{y}_{i,n} = 0$. The instances $\{y(x_i)\}$ occurring between two consecutive intersections x_{j-1} and x_j define the jth cluster in the partition $\{C\}$. Let

$$\ell_j \equiv \|x_j - x_{j-1}\| \tag{2}$$

be the *distance* between x_{j-1} and x_j. The frequency of the clusters of length ℓ_j is obtained by counting the number of clusters $\mathcal{N}(\ell_1, n)$, $\mathcal{N}(\ell_2, n), \ldots, \mathcal{N}(\ell_j, n)$ according to their length $\ell_1, \ell_2, \ldots, \ell_j$ for each n:

$$P(\ell_j, n) = \frac{\mathcal{N}(\ell_j, n)}{\mathcal{N}_C(n)} \tag{3}$$

with $\mathcal{N}_C(n) = \sum_{j=1}^{k(n)} \mathcal{N}(\ell_j, n)$ the number of clusters generated by the partition for each n. Let $k = \sum_{n=1}^{N} \mathcal{N}_C(n)$ indicate the total number of clusters for all the possible values of n. The normalization condition holds:

$$\sum_{n=1}^{N}\sum_{j=1}^{\mathcal{N}_C(n)} P(\ell_j, n) = 1 \quad , \tag{4}$$

where the index j runs over the unique cluster lengths obtained for each partition of size n, with $n \in (1, N)$.

The *relative cluster entropy* is defined in terms of the Kullback–Leibler functional as follows:

$$\mathcal{D}_{j,n}[P \parallel Q] = P(\ell_j, n)\log\frac{P(\ell_j, n)}{Q(\ell_j, n)} \quad , \tag{5}$$

where the cluster probability distributions $P(\ell_j, n)$ and $Q(\ell_j, n)$ indicate the cluster frequency obtained, respectively, from the empirical data sequence and the artificial data sequence defined through a suitable stochastic process assumed as a model. The artificial data sequence is generated with length N equal to that of the empirical data to satisfy the condition $\mathrm{supp}(P(\ell_j, n)) \subseteq \mathrm{supp}(Q(\ell_j, n))$. After generating the artificial sequence (e.g. a long-range correlated sequence by using a fractional Brownian motion, ARFIMA), the clusters are obtained and counted to yield $Q(\ell_j, n)$ using the same approach explained above for $P(\ell_j, n)$.

The functionals $\mathcal{D}_{j,n}[P \parallel Q]$ defined by Eq. (5) represent the individual components of the relative cluster entropy. After summing Eq. (5) over all possible values of the distance ℓ_j and moving window n, one has:

$$\mathcal{D}_C[P \parallel Q] = \sum_{n=1}^{N}\sum_{j=1}^{\mathcal{N}_C(n)} P(\ell_j, n)\log\frac{P(\ell_j, n)}{Q(\ell_j, n)} \quad , \tag{6}$$

which quantifies the information gained when the empirical probability distribution P is compared to the probability distribution Q taken as a reference.

To illustrate the inferential power of the *relative cluster entropy*, we consider Eqs. (5) and (6) in terms of continuous variables with the sums replaced by integrals. Let the probability distribution functions be in the form of power-laws:

$$P(\ell) \sim \ell^{-\alpha_1} \qquad Q(\ell) \sim \ell^{-\alpha_2} \tag{7}$$

with α_1 and α_2 denoting the scaling exponents. The relative entropy of the cluster in Eq. (6) can easily be estimated in terms of a definite integral with respect to the continuous variable ℓ. By calculating the integral over the interval $[1, \infty)$ one obtains the following:

$$D_C[P \parallel Q] = \log \frac{\alpha_1 - 1}{\alpha_2 - 1} + \frac{\alpha_1 - \alpha_2}{1 - \alpha_1} \quad , \tag{8}$$

which satisfies the relative entropy condition to be positively defined as $D_C[P \parallel Q] \geq 0$ over the whole range of α_1 and α_2, while $D_C[P \parallel Q] = 0$ for $\alpha_1 = \alpha_2$, i.e. for the empirical distribution P coincident with the model distribution Q. Therefore, interestingly, $D_C[P \parallel Q]$ turns out to be a function of the correlation exponents α_1 and α_2 only:

$$D_C[P \parallel Q] = D_C[\alpha_1 \parallel \alpha_2] \tag{9}$$

and thus can be adopted as a straightforward measure of divergence between the empirical and model correlation exponents.

To illustrate how to proceed with an empirical dataset, let $P(\ell_j, n)$ be the empirical probability distribution obtained by a sample-by-sample ranking of a set of random variables ℓ_j. $P(\ell_j, n)$ is assumed to behave as a power-law distribution with an unknown correlation exponent α_1. Then let $Q(\ell_j, n)$ be a reference probability distribution generated by artificial data according to a stochastic model with assigned correlation exponent α_2. Using the relative entropy functional Eq. (5), the empirical cluster probability distribution $P(\ell_j, n)$ is compared to the model probability distribution $Q(\ell_j, n)$. To implement the relative cluster entropy algorithm, the probabilities $P(\ell_j, n)$ and $Q(\ell_j, n)$ must be estimated on data sequences of the same length and numerical precision. To infer the optimal probability distribution $P(\ell_j, n)$, the *minimum relative entropy* principle is implemented non-parametrically. Specifically, the variance of the functionals $\mathcal{D}_{j,n}[P \parallel Q]$ around the value $\mathcal{D}_{j,n}[Q \parallel Q]$ under the null hypothesis for $P = Q$ is considered:

$$\sigma_D^2[P \parallel Q] \equiv \sum_{j,n} \left[\mathcal{D}_{j,n}[P \parallel Q] - \mathcal{D}_{j,n}[Q \parallel Q]\right]^2 \tag{10}$$

with the sum running over the total number of clusters obtained by the partition process. Taking into account that $\mathcal{D}_{j,n}[Q \parallel Q] = 0$, the previous relationship can be simplified as follows:

$$\sigma_D^2[P \parallel Q] \equiv \sum_{j,n} \left[\mathcal{D}_{j,n}[P \parallel Q]\right]^2 \quad . \tag{11}$$

The quantity σ_D^2 corresponds to the mean squared value of the area of the region between the curve $\mathcal{D}_C[P \parallel Q]$ and the horizontal axis $\mathcal{D}_C[Q \parallel Q] = 0$. For probability distributions in the form of power laws, σ_D^2 can be rewritten in terms of the correlation exponents α_1 and α_2:

$$\sigma_D^2[\alpha_1 \parallel \alpha_2] \equiv \sum_{j,n} \left[\mathcal{D}_{j,n}[\alpha_1 \parallel \alpha_2]\right]^2 \quad . \tag{12}$$

The quantity σ_D^2 turns out to be approximately a quadratic function of the difference between the exponents α_1 and α_2. It takes its minimum value ($\sigma_D^2 = 0$) for $\alpha_1 = \alpha_2$ with the asymmetry of the Kullback–Leibler entropy functional over the range of values α. The above relationships can be used to estimate the minimum value $\sigma_D^2 = 0$ of the relative cluster entropy corresponding to the condition $\alpha_1 = \alpha_2$ and ultimately infer the value of α_1 to quantify the features of the empirical data.

3. Data

The 24 chromosomes of the T2T-CHM13+Y human genome are the empirical data used to generate the probability distributions $P(\ell_j, n)$. Genomic sequences are recorded as natural language texts composed of adenine (A), thymine (T), cytosine (C), and guanine (G) nucleotides (FASTA files). Segments of ℓ_j nucleotides are referred to as ℓ-words. The first step is the numerical mapping of the FASTA sequences, i.e. the conversion of the four-letter nucleotide alphabet into a numerical format suitable for statistical analysis. In this work, the sequence of nucleotide bases is mapped according to the RY rule:purine (A,G) and pyrimidine (C,T) bases are mapped, respectively, to positive and negative random numbers generated by a Gaussian distribution with zero mean and standard deviation equal to 1. Then the sequence of positive and negative numbers is summed and a random walk $y(x_i)$ (*DNA walk*) with x_i the nucleotide position is obtained. Several coding rules have been adopted to map large DNA nucleotide contigs to numerical sequences [10]. The RY coding rule has the advantage of keeping the nonstationarity of the numerical series to a minimum. On average, the concentrations of A and T are each approximately 30%, while those of G and C are approximately 20%, thus the concentrations of purines (A+G) and pyrimidines (C+T) are very close to 50% along the sequence. Coding nucleotides into a balanced distribution of positive and negative values yields a numerically unbiased series. Conversely, unbalanced coding rules would amplify the variations of the local density distribution of the bases along the sequences, increasing nonstationarity and bias of the corresponding random walk. Dedicated Python scripts have been implemented to map the 24 chromosomes FASTA files (text format) to numerical sequences. For illustration purposes, thirty bases of the chromosome 1 (T2T-CHM13+Y) are shown as text symbols A,G,C,T at the bottom of Fig. 1, blue squares represent the DNA walk $y(x)$ and red continuous line is the local average $\widetilde{y}(x)$.

Next, a suitable stochastic process is assumed as a model to generate the artificial data sequences needed to obtain the probability distributions $Q(\ell_j, n)$ to be entered into Eq. (5). As stated above, the chromosomes of the human reference genome are used as empirical sequences. Chromosomes are characterized by long-range correlations. Hence, for the purpose of generating sequences mimicking the long-range correlated chromosomes, fractional Brownian motions are considered as the generating stochastic process model. Then, the probability distributions $Q(\ell_j, n)$ are power laws with scaling exponent $\alpha = 2 - H$ where H is the Hurst exponent of the fractional Brownian motion. Using $\alpha_1 = 2 - H_1$ and $\alpha_2 = 2 - H_2$, Eq. (8) takes the form:

$$D_C[P \parallel Q] = \log \frac{1 - H_1}{1 - H_2} + \frac{H_1 - H_2}{1 - H_1} \quad , \tag{13}$$

which is positively defined over the entire range of H_1 and H_2 values and vanishes for $H_1 = H_2$. Interestingly, $D_C[P \parallel Q]$ turns out to be a function only of the Hurst exponents H_1 and H_2 of the empirical and model sequences:

$$D_C[P \parallel Q] = D_C[H_1 \parallel H_2] \tag{14}$$

and can be taken as an inferential tool of H_1 the unknown Hurst exponent of the empirical sequence. To infer H_1, model probability distributions $Q(\ell_j, n)$ are generated by fractional Brownian motions with H_2 varying in the range $[0, 1]$, then the condition $D_C[P \parallel Q] = 0$ yields $H_1 = H_2$.

Probability distributions $P(\ell_j, n)$ and $Q(\ell_j, n)$ are estimated by counting the clusters generated respectively in the chromosome and artificial sequences according to the definition given in Eq. (3). The relative entropy of the clusters is then estimated by entering the values of the probability distributions $P(\ell_j, n)$ and $Q(\ell_j, n)$ in the functional $D_{j,n}[P \parallel Q]$ defined by Eq. (5). Individual functionals $\mathcal{D}_{j,n}[P \parallel Q]$ are estimated at different cluster lengths by varying the indices j and n.

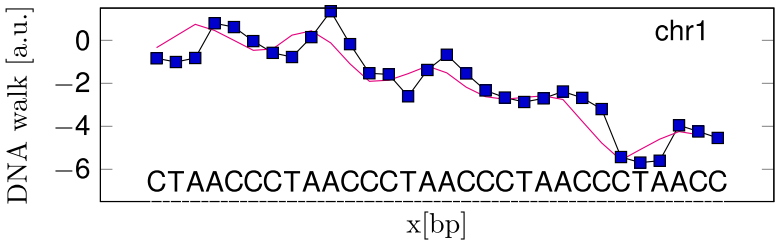

Fig. 1. The letter sequence at the bottom corresponds to the first 30 bases of the human chromosome 1. A numerical sequence is obtained according to the RY rule: each purine (A, G) and pyrimidine (C, T) base is mapped respectively to a positive or negative random number drawn from a Gaussian distribution with zero mean and variance +1. The DNA walk $y(x)$ (blue squares) is obtained by summing the numerical sequence of positive and negative random numbers. Continuous line (red) shows the moving average $\widetilde{y}(x_i, n)$ estimated by using the DNA walk (blue squares) over moving average windows $n = 3bp$.

Interestingly, $\mathcal{D}_{j,n}[P \parallel Q]$ takes positive or negative values depending on the difference between the Hurst exponents H_1 and H_2, namely:

$$\mathcal{D}_{j,n}[P \parallel Q] \begin{cases} > 0 & \text{for} \quad H_1 > H_2, \\ = 0 & \text{for} \quad H_1 = H_2, \\ < 0 & \text{for} \quad H_1 < H_2 \end{cases} \tag{15}$$

After summing the quantities $\mathcal{D}_{j,n}[P \parallel Q]$ over ℓ_j and n by using Eq. (11), the minimum relative entropy principle is applied thus yielding the optimal inference for the Hurst exponent H_1.

To summarize, the above equation set can be used to quantify the divergence of an empirical sequence with correlation exponent H_1 compared to a fractional Brownian motion with assigned Hurst exponent H_2 taken as a model. It is worth recalling that Hurst exponent takes values in the range $[0, 1]$, with $H = 0.5$ indicating a simple random process, while $0 \leq H < 0.5$ and $0.5 < H \leq 1$ correspond, respectively, to negatively and positively correlated processes.

4. Results

The above detailed KLCE analysis has been performed on multiple pairs of nucleotide strands along chromosomes and artificial fractional Brownian motion sequences of equal length N. Probability distributions $P(\ell_j, n)$ and $Q(\ell_j, n)$ are obtained by counting clusters generated by using moving average windows varying in the range $2 < n < 1000$. As the ratio $P(\ell_j, n)/Q(\ell_j, n)$ quickly becomes negligible as ℓ_j increases, the sum in Eq. (11) is calculated up to a threshold cluster length $\ell_j(th) \leq 100$. For illustrative purposes, the values σ_D^2 estimated according to Eq. (11) on three nucleotide strands of the human chromosome 1 are shown in Fig. 2. The strands are $N = 100$ kbases-long beginning respectively at nucleotide coordinates chr1: 500–100500 (continuous line), chr1: 1267000–1367000 (circles) and chr1: 1282000–1382000 (triangles). The minimum of the curves shown in Fig. 2 corresponds to the condition $H_1 = H_2$ in the relative cluster entropy and thus yields the optimal estimate of the correlation exponent $H_1 = 0.51$, $H_1 = 0.53$ and $H_1 = 0.65$ respectively for each strand.

The entire chr1 is 248.3 Mbases long, hence a total number $N_s = 2483$ of 100 kbases-long strands needed to cover chr1 have been analyzed, yielding 2483 values of the Hurst exponent corresponding to each strand. The values of the Hurst exponent obtained for all the chr1 strands are shown in the top-left panel of Fig. 3. The procedure has been applied to 100 kbases-long nucleotide strands for all 24 chromosomes. The results obtained for chromosomes chr1-12 and chr13-Y are shown, respectively, in Fig. 3 and Fig. 4. Strong fluctuations of the Hurst exponent are systematically observed corresponding to the newly added nucleotide strands shown by the gold lines at the bottom of each figure.

Number of strands N_s, mean value H and standard deviation σ_H of the Hurst exponent for the entire chromosomes, start and end coordinates of the newly added nucleotides in the T2T-CHM13+Y assembly are reported in Table 1. The mean Hurst exponent is always larger than 0.5, indicating that the nucleotide strands exhibit positive correlation all throughout the 24 chromosome sequences.

5. Validation

In this section, significance, robustness and accuracy of the Kullback–Leibler cluster entropy method are discussed. A comparison with the Detrending Fluctuation Analysis and the Detrending Moving Average algorithms is also included. There are a few reasons that make the comparison of KLCE computationally and statistically significant with reference to those methods: (1) DMA and DFA are widely adopted by the scientific community to discriminate between correlated and anticorrelated sequences in terms of scaling exponents; (2) KLCE and DMA implement the same partition procedure to generate the ensembles of elementary random variables (the moving average cluster lengths ℓ_j with windows of size n); (3) DFA is based on a nearly similar partition procedure. The DFA operates via the division of the sequence in boxes of size n that can be assimilated to the moving average filter windows n. The range of sizes n is a critical parameter for the performance of the algorithm. Hence, it is convenient that statistical significance, accuracy and robustness are compared by varying the parameter n over the same range. Before discussing the numerical validation outcomes, for the sake of completeness the main computational steps of the DMA and DFA are briefly recalled.

The *Detrending Moving Average* algorithm is based on the estimation of the generalized variance defined for one-dimensional sequences $y(x_i)$ of relevance to this work as:

$$\sigma^2_{DMA} = \frac{1}{N - n_{max}} \sum_{i=n}^{N} [y(x_i) - \widetilde{y}(x_i, n)]^2 \quad , \tag{16}$$

with $\widetilde{y}(x_i, n)$ defined by Eq. (1). Moving average polynomials of orders higher than Eq. (1) can be used, resulting in more accurate filtering of the sequence at the expense of higher computational complexity of the algorithm [48]. The DMA algorithm revolves around the following main steps: the moving average $\widetilde{y}(x_i, n)$ is estimated with different values of n, then the difference $y(x_i) - \widetilde{y}(x_i, n)$ for each n is entered in Eq. (16); the quantity σ^2_{DMA} is plotted on log–log scales. It has been shown that for data distributed according to a power-law with scaling exponent α, the following relationship holds:

$$\sigma_{DMA} \sim n^{\alpha} \quad . \tag{17}$$

Hence, the regression of log–log plots of σ_{DMA} vs. the moving average window n yields the slope α.

The *Detrending Fluctuation Analysis* begins with the division of the sequence $y(x_i)$ into boxes of equal length n. In each box, $y(x_i)$ is fitted by a polynomial $\widehat{y}(x_i, n)$. A different order of the polynomial can be used resulting in linear, parabolic, cubic fit and so on. Next, the difference $y(i) - \widehat{y}_n(i)$ between the sequence $y(i)$ and the fit $\widehat{y}_n(i)$ is estimated for each box of length n. Finally, for each box n, the quantity :

$$\sigma^2_{DFA} \equiv \frac{1}{N-1} \sum_{i=1}^{N} \left[y(i) - \widehat{y}_n(i) \right]^2 \tag{18}$$

is calculated. For scale-invariant sequences with power-law correlations, a linear relationship between function σ_{DFA} and the box size n is

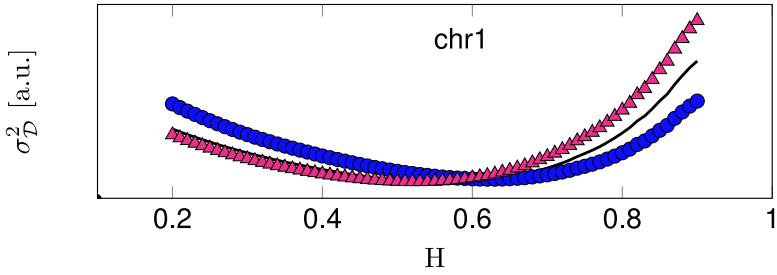

Fig. 2. Variance of the Kullback–Leibler cluster entropy estimated according to Eq. (11) over three nucleotide sequences of the T2T-CHM13+Y chromosome 1. The three sequences have same length (100 kbases). The probability distributions $P(\ell_j, n)$ and $Q(\ell_j, n)$ are obtained by counting clusters generated by using moving average windows varying in the range $2 < n < 1000$. The threshold cluster length $\ell_j(th)$ used in Eq. (11) is equal to 20bases. The three strands refer to nucleotide coordinates ranging respectively from chr1: 500–100500 (continuous black line), chr1: 1267000–1367000 (blue circles), chr1: 1282000–1382000 (magenta triangles). The minimum value in each curve corresponds to the condition $H_1 = H_2$ and provides the optimal Hurst exponent, found respectively equal to $H = 0.51$ (triangles), $H = 0.55$ (continuous line) and $H = 0.65$ (circles).

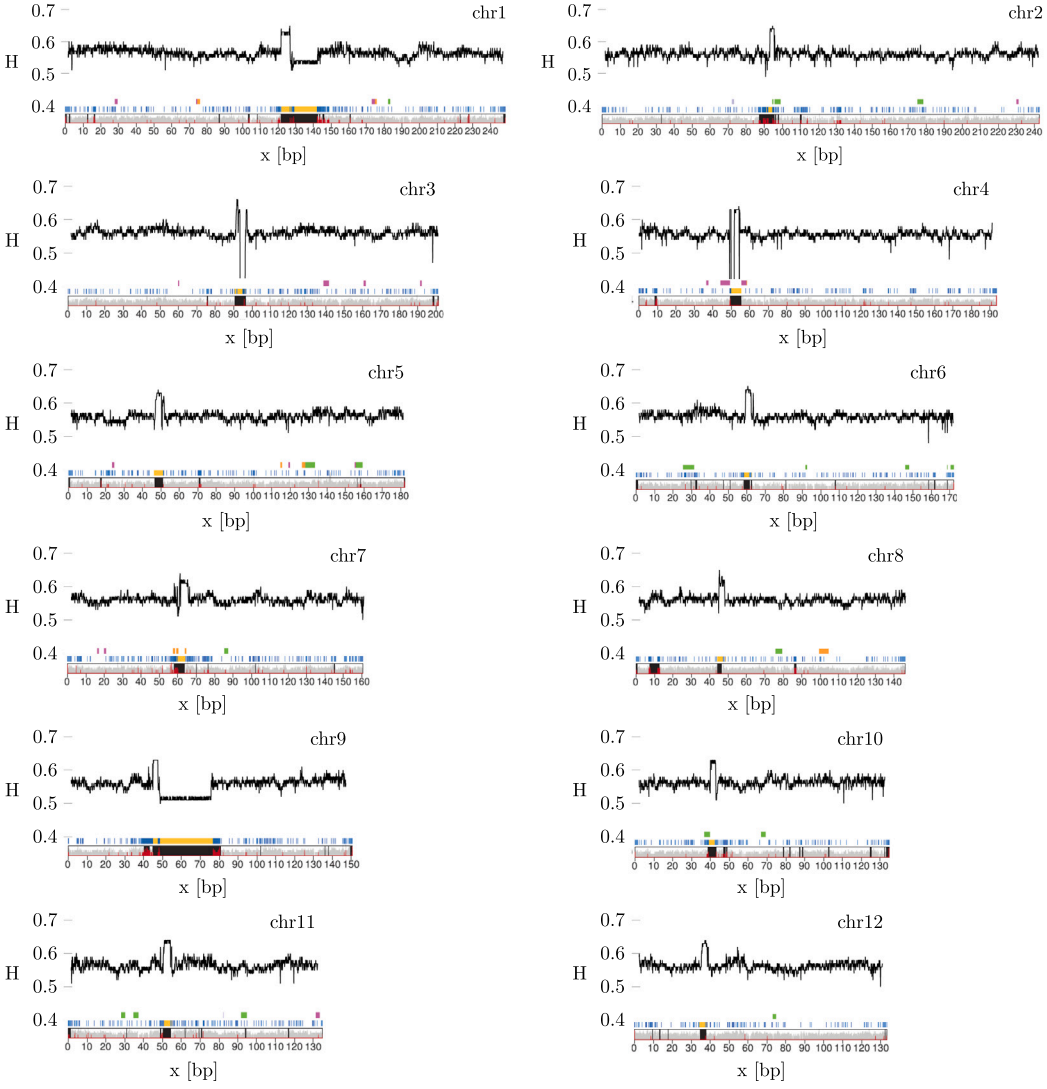

Fig. 3. Hurst exponents vs. nucleotide positions for equal length (100kbases) strands of T2T-CHM13+Y chrs 1–12. Each value of the Hurst exponent is obtained by taking the minimum cluster entropy divergence of the nucleotide strand. The nucleotide position of the T2T-CHM13+Y newly added regions are shown against the GRCh38 assembly at the bottom of each panel.

Table 1
Chromosome number (chr); number of 100kbases long nucleotide strands (N_s); mean value (H) and standard deviation (σ_H) of the Hurst exponent for the whole chromosome; start and end nucleotide coordinates of the newly added T2T-CHM13+Y strands. The average H is yield by the relative cluster entropy described in Section 2 for each of the N_s nucleotide regions.

chr	N_s	H	σ_H	start	end	chr	N_s	H	σ_H	start	end
1	2483	0.5643	0.0172	116 796.16	147 241.28	13	1135	0.5397	0.0637	0	23171058
2	2426	0.5612	0.0142	85 991 672	996 730.6	14	1011	0.5623	0.0208	0	17765925
3	2011	0.5579	0.0341	85 805 192	101 415.17	15	997	0.5585	0.0327	0	23279251
4	1935	0.5541	0.0297	44 705 247	598 706.4	16	963	0.5635	0.0177	30 848 291	57219476
5	1820	0.5610	0.0151	42 077 197	545 966.9	17	842	0.5734	0.0194	18 892 710	32487230
6	1721	0.5614	0.0155	53 286 920	660 586.2	18	805	0.5634	0.0234	10 965 698	25933550
7	1605	0.5617	0.0147	55 414 368	687 144.6	19	617	0.5762	0.0200	19 655 572	34768168
8	1462	0.5606	0.0136	39 243 541	513 250.6	20	662	0.5700	0.0197	21 383 653	37969531
9	1506	0.5544	0.0241	39 952 789	816 940.3	21	450	0.5498	0.0412	0	17078862
10	1347	0.5636	0.0153	34 633 784	466 645.0	22	513	0.5590	0.0518	0	20739833
11	1351	0.5660	0.0181	46 061 948	594 134.5	X	1542	0.5574	0.0179	52 820 107	65927026
12	1333	0.5636	0.0159	29 620 490	422 024.2	Y	624	0.4805	0.1164	57 264 655	62460029

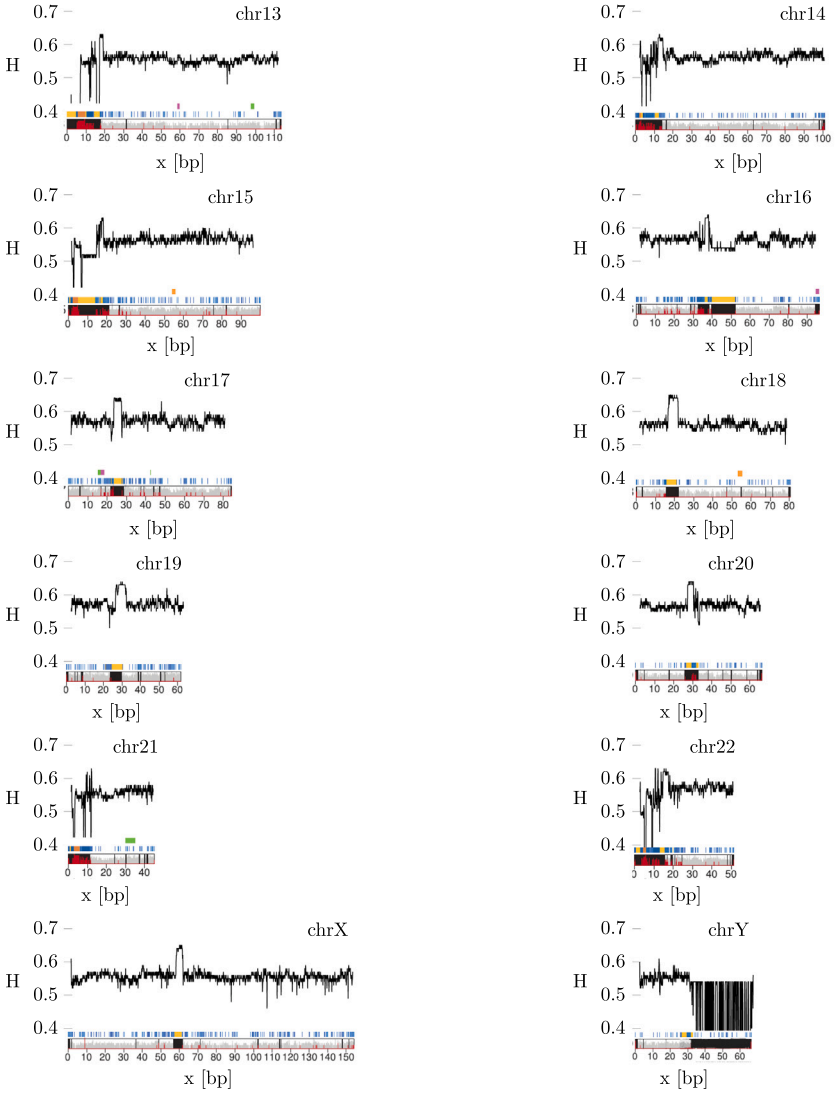

Fig. 4. Hurst exponents vs. nucleotide positions for strands of equal length (100kbases) of the T2T-CHM13+Y chrs 13-Y. Same as Fig. 3.

observed:

$$\sigma_{DFA} \sim n^{\alpha}. \tag{19}$$

The log–log plot of σ_{DFA} vs. the box size n yields a straight line with slope α.

The comparison of DMA, DFA and KLCE algorithms reported in this section are implemented on sets of 100 artificial sequences (fBms) generated by the *Wood–Chan algorithm* [49]. Outcomes are illustrated in the histograms of Fig. 5. The histograms refer to the outcomes yielded by using 100 fBms with length respectively equal to $N = 10000$ (a,b,c,d) and $N = 100000$ (e,f,g,h). The fBms are characterized by

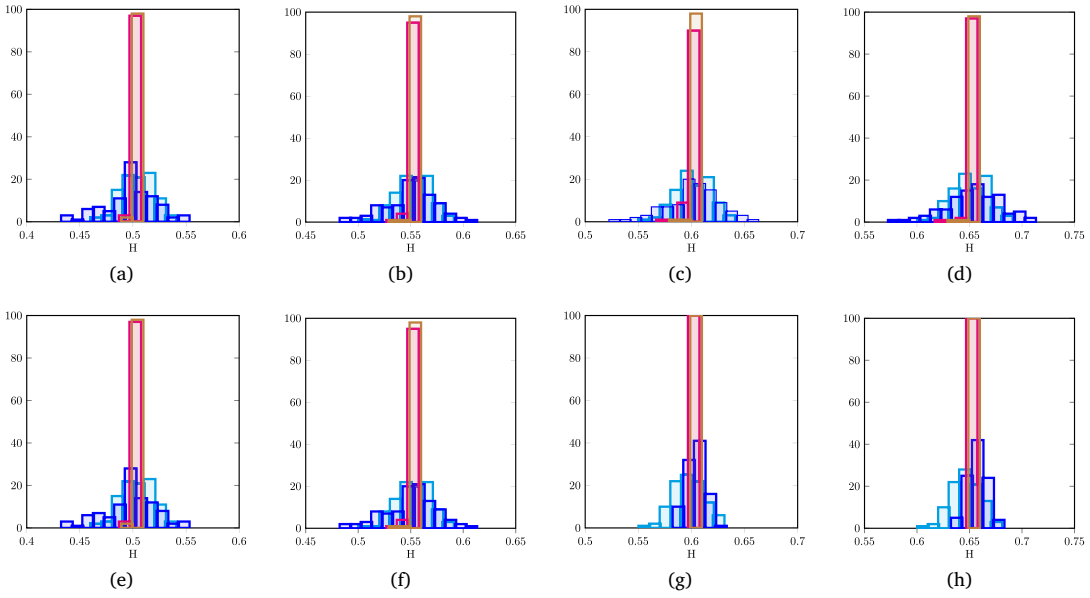

Fig. 5. Outcomes of KLCE, DMA and DFA algorithms implemented on 100 samples of artificially generated sequences (fBMs) with assigned Hurst exponents. The sample sequences' lengths are respectively equal to $N = 10000$ (a, b, c, d) and $N = 100000$ (e, f, g, h). Nominal Hurst exponents are: $H = 0.5$ (a,e); $H = 0.55$ (b,f); $H = 0.60$ (c,g); $H = 0.65$ (d,h). Colors are respectively: blue (DMA), cyan (DFA), orange (KLCE20) and brown (KLCE30). KLCE20 and KLCE30 refer to the σ_D estimated by using Eq. (11) respectively with cluster length thresholds $\ell_j(th) = 20$bases and $\ell_j(th) = 30$bases.

Table 2
Success probability π, mean μ, variance σ and Detrending Moving Average, Detrending Fluctuation Analysis, and Kullback–Leibler cluster entropy, corresponding to the histograms in Fig. 5. The success probability π is estimated as the ratio of the number of the nominal outcomes to the total outcomes.

	DMA			DFA			KLCE (20)			KLCE (30)		
	π (%)	μ	σ	π (%)	μ	σ	π (%)	μ	σ	π (%)	μ	σ
a	22	0.5085	0.0051	28	0.4996	0.0057	97	0.4997	0.0004	98	0.4998	0.0003
b	22	0.5566	0.0052	20	0.5428	0.0071	95	0.5494	0.0007	98	0.5497	0.0004
c	24	0.6057	0.0053	20	0.6038	0.0057	90	0.5988	0.0013	98	0.5997	0.0004
d	23	0.6547	0.0055	15	0.644	0.0077	97	0.6495	0.0006	98	0.6497	0.0004
e	28	0.5053	0.0046	47	0.5076	0.0051	100	0.5065	0.0066	100	0.5065	0.0066
f	24	0.5535	0.0047	43	0.5543	0.0036	100	0.5500	0	100	0.5500	0
g	25	0.6012	0.0048	32	0.6066	0.0037	100	0.6000	0	100	0.6000	0
h	28	0.6504	0.0047	25	0.6597	0.0037	100	0.6500	0	100	0.6500	0

nominal Hurst exponent: $H = 0.50$ (a,e); $H = 0.55$ (b,f); $H = 0.60$ (c,g); $H = 0.65$ (d,h). The KLCE and DMA outcomes are obtained with moving average windows n ranging over the same interval, namely n from 2 to 1000 with steps given by $i \cdot 10^{kk}$ with $kk = 0, 1, 2$ and $i \in [2, 10]$. The same n values are used to generate the boxes in the DFA procedure to make the comparison among the three methods computationally equivalent. The KLCE outcomes refer respectively to threshold cluster lengths $\ell_j(th) = 20bases$ and $\ell_j(th) = 30bases$ considered in the estimation of the Eq. (11). Mean μ, variance σ and success probability π for the outcomes of the KLCE, DMA and DFA algorithms shown in Fig. 5 are summarized in Table 2. The success probability π is defined as the ratio of the number of exact to total number of outcomes. One can note that π for KLCE is higher than DMA and DFA in all the analyzed cases. Furthermore, on average, the KLCE error is one order of magnitude lower than those of the DMA and DFA.

The sources of errors that come into play over the main computational steps of the three algorithms are now briefly discussed. DMA and DFA errors intrinsically come from the differences $y(x_i) - \widetilde{y}(x_i, n)$ and $y(x_i) - \widehat{y}(x_i, n)$ appearing respectively in Eqs. (16) and (18). Each value $y(x_i)$ is obtained as a mean value plus a random variable output of a Gaussian number generator (here the RY and the fBms generators). $\widetilde{y}(x_i, n)$ and $\widehat{y}(x_i, n)$ provide smoothed values of $y(x_i)$. The differences $y(x_i) - \widetilde{y}(x_i, n)$ and $y(x_i) - \widehat{y}(x_i, n)$ are proportional to the sum of the individual random variables yield by the Gaussian number generator over n. KLCE errors intrinsically come from the measures of the cluster lengths, i.e. the differences $\ell_j \equiv \|x_j - x_{j-1}\|$ appearing in Eq. (2). The intersections x_j and x_{j-1} correspond to the sequence items where $y(x_i) = \widetilde{y}(x_i, n)$ i.e. the points where the errors on the $y(x_i)$ and $\widetilde{y}(x_i, n)$ exactly compensate each other. Hence the cluster lengths ℓ_j are affected less by the errors of the random Gaussian number generators compared to the quantities $y(x_i) - \widetilde{y}(x_i, n)$ and $y(x_i) - \widehat{y}(x_i, n)$. The cluster lengths ℓ_j are affected by the finite size effects of the window n which in turns are reflected in the empirical and model distributions. Further inaccuracy might come from the algorithms used to generate the artificial sequences used as model. Thus testing the algorithms outcomes by using different artificial generator might be of help. Finite size effects due to the length N of the empirical or model sequences, the box/window n are main causes of inaccuracy and biases. Least-square regression of the log–log plots of Eqs. (16) and (18) cause errors when DMA and DFA analysis are applied. Hence, caution should be taken by using at least three decades of n values to ensure that the distributions are consistent with a power-law, increasing the number of analyzed samples N_s and range of sequence lengths N consistent with the range of n values.

6. Discussion and conclusions

The Kullback–Leibler cluster entropy has been adopted as a tool to quantify the degree and extent of the long-range correlations exhibited by the 24 chromosomes of the recently published T2T-CHM13+Y complete human genome assembly. A systematic analysis has been performed considering nucleotide strands of the same length along all 24 chromosomes. Several divergence values obtained for different values of the parameters n and ℓ_j have been summed over values of n, with the same interval of cluster length ℓ_j. Typical results corresponding to three strands of chromosome 1 are shown in Fig. 2. Values of the Hurst exponent inferred for the N_s strands of all the 24 chromosomes are shown in Fig. 3 and Fig. 4. In Table 1, the mean values of the Hurst exponent averaged over all N_s strands of the 24 chromosomes are reported together with the nucleotide coordinates of the newly added strands. Numerical validation of the KLCE techniques is reported in Section 5 together with the comparison with DMA and DFA.

It is worth noting that for the five acrocentric chromosomes strong deviations of the correlation exponent are observed at the beginning of the sequences, consistently with the coordinates of the newly added strands reported in Table 1 (chr13: 0–23171058; chr14: 0–17765925; chr15: 0–23279251; chr21: 0–17078862; chr22: 0–20739833 respectively). The deviations, quantified in terms of standard deviation $\sigma_H = 0.0637$, $\sigma_H = 0.0208$, $\sigma_H = 0.0327$, $\sigma_H = 0.0412$, $\sigma_H = 0.0518$, respectively, take higher values compared to the other chromosomes. High density of satellite repeats, segmental duplications and dynamic complexity have prevented sequencing of the short arms of the five acrocentric chromosomes until the recent completion of the T2T-CHM13+Y assembly, despite their biological relevance. Similarly, a high variability of the Hurst exponent is observed in the newly added fragments of the chromosome Y over the range of nucleotide coordinates chrY: 57264655-62460029, as also indicated by the higher standard deviation value $\sigma_H = 0.1164$ compared to the other chromosomes. The relative cluster entropy approach shows that the newly included segments of the T2T-CHM13+Y display higher variability and stronger correlation compared to the adjacent strands already present in the GRCh38 assembly. These results are consistent with the higher level of complexity and repetitions in the newly added parts. The annotated self-similarity with the related power-law correlations and Hurst exponent across different chromosomal segments can be adopted as an inferential criterion for assessing complexity and heterogeneity of chromosome structure. The power-law scaling of the long-range correlated clusters obtained in the linear sequence of reference have meaningful implications from a biological perspective. Regions of genomes characterized by high levels of heterogeneity and disorder are more likely to duplicate and mutate. Overall, the outcomes presented confirm the robustness and specificity of the Kullback–Leibler cluster entropy measure in quantifying the occurrence of regions with repetitive content and duplication in genomic sequences as a complementary tool to standard alignment methods.

Further development and application areas of the proposed method can be envisaged in the context of deep-learning approaches [17,42–44], where the challenges of adaptive thresholding and onset of spurious correlation among the samples need to be tackled. Future research directions may involve estimating the relative cluster entropy $D_C[P \parallel Q]$ and quantifying the *divergence* between the empirical probability distribution of the pangenome minigraph segments $P(s)$ and the cluster distribution $P(\ell, n)$ of the reference linear sequences. The minimum relative entropy principle could be adopted as a stringent optimization criterion for accurate inference about the structure of the pangenome.

CRediT authorship contribution statement

Filippo Gandino: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Data curation. **Renato Ferrero:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation. **Anna Carbone:** Writing – review & editing, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] James Ladyman, Karoline Wiesner, What is a Complex System? Yale University Press, 2020.
[2] James P. Crutchfield, Between order and chaos, Nat. Phys. 8 (1) (2012) 17–24.
[3] Erhan Gokcay, Jose C. Principe, Information theoretic clustering, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 158–171.
[4] Nguyen Xuan Vinh, Julien Epps, James Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, J. Mach. Learn. Res. 11 (2010) 2837–2854.
[5] James Bailey, Michael E. Houle, Xingjun Ma, Relationships between tail entropies and local intrinsic dimensionality and their use for estimation and feature representation, Inf. Syst. 118 (2023) 102245.
[6] Zoubin Ghahramani, in: Olivier Bousquet, Ulrike von Luxburg, Gunnar Rätsch (Eds.), Unsupervised learning, in: Advanced Lectures on Machine Learning, Springer Berlin Heidelberg, 2004, pp. 72–112.
[7] Aaron Clauset, Cosma Rohilla Shalizi, Mark E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.
[8] Richard F. Voss, Fractals in nature: from characterization to simulation, in: The Science of Fractal Images, Springer, 1988, pp. 21–70.
[9] C-K Peng, Sergej V. Buldyrev, Ary L. Goldberger, Shlomo Havlin, Francesco Sciortino, Michael Simons, H. Eugene Stanley, Long-range correlations in nucleotide sequences, Nature 356 (6365) (1992) 168–170.
[10] Sergey V. Buldyrev, Ary L. Goldberger, Shlomo Havlin, Chung-Kang Peng, Michael Simons, H. Eugene Stanley, Generalized Lévy-walk model for DNA nucleotide sequences, Phys. Rev. E 47 (6) (1993) 4514.
[11] Limei Xu, Plamen Ch Ivanov, Kun Hu, Zhi Chen, Anna Carbone, H. Eugene Stanley, Quantifying signals with power-law correlations: a comparative study of detrended fluctuation analysis and detrended moving average techniques, Phys. Rev. E—Statistical, Nonlinear, Soft Matter Phys. 71 (5) (2005) 051101.
[12] Anna Carbone, Algorithm to estimate the Hurst exponent of high-dimensional fractals, Phys. Rev. E—Statistical, Nonlinear, Soft Matter Phys. 76 (5) (2007) 056703.
[13] Naor Granik, Lucien E. Weiss, Elias Nehme, Maayan Levin, Michael Chein, Eran Perlson, Yael Roichman, Yoav Shechtman, Single-particle diffusion characterization by deep learning, Biophys. J. 117 (2) (2019) 185–192.
[14] İbrahim Avcı, Hüseyin Lort, Buğce E. Tatlıcıoğlu, Numerical investigation and deep learning approach for fractal–fractional order dynamics of Hopfield neural network model, Chaos Solitons Fractals 177 (2023) 114302.
[15] Junhao Bian, Zhiqin Ma, Chunping Wang, Tao Huang, Chunhua Zeng, Early warning for spatial ecological system: fractal dimension and deep learning, Phys. A 633 (2024) 129401.
[16] Yen-Ching Chang, Deep-learning estimators for the hurst exponent of two-dimensional fractional Brownian motion, Fractal Fract. 8 (1) (2024) 50.
[17] Armando Delgadillo-Jimenez, Carina Toxqui-Quitl, Raúl Castro-Ortega, Aldo Aguilar-Vallejo, Alfonso Padilla-Vivanco, Anna Carbone, Deep learning approach to fractal surfaces: application to satellite images of urban areas, EPJ Data Sci. (2026).
[18] Fuyuan Xiao, Weiping Ding, Witold Pedrycz, A generalized f-divergence with applications in pattern classification, IEEE Trans. Knowl. Data Eng. 37 (4) (2025) 1556–1570.
[19] Lang Zhang, Fuyuan Xiao, Belief Rényi divergence of divergence and its application in time series classification, IEEE Trans. Knowl. Data Eng. 36 (8) (2024) 3670–3681.
[20] Anna Carbone, Linda Ponta, Relative cluster entropy for power-law correlated sequences, SciPost Phys. 13 (3) (2022) 076.
[21] Linda Ponta, Anna Carbone, Kullback–Leibler cluster entropy to quantify volatility correlation and risk diversity, Phys. Rev. E 142 (3) (2025) 205–210.
[22] Karoline Wiesner, James Ladyman, Complex systems are always correlated but rarely information processing, J. Phys.: Complex. 2 (4) (2021) 045015.
[23] Tom Misteli, The self-organizing genome: principles of genome architecture and function, Cell 183 (1) (2020) 28–45.
[24] Siwei Chen, Laurent C. Francioli, Julia K. Goodrich, et al., A genomic mutational constraint map using variation in 76,156 human genomes, Nature 625 (7993) (2024) 92–100.

[25] Rafael Plana Simões, Ivan Rodrigo Wolf, Bruno Afonso Correa, Guilherme Targino Valente, Uncovering patterns of the evolution of genomic sequence entropy and complexity, Mol. Genet. Genom. 296 (2021) 289–298.
[26] Dolores Bernenko, Sang Hoon Lee, Ludvig Lizana, Exploring 3D community inconsistency in human chromosome contact networks, J. Phys.: Complex. 4 (3) (2023) 035004.
[27] Junyi Li, Li Zhang, Huinian Li, Yuan Ping, Qingzhe Xu, Rongjie Wang, Renjie Tan, Zhen Wang, Bo Liu, Yadong Wang, Integrated entropy-based approach for analyzing exons and introns in dna sequences, BMC Bioinformatics 20 (2019) 1–7.
[28] Subhash Kak, Self-similarity and the maximum entropy principle in the genetic code, Theory Biosci. 142 (3) (2023) 205–210.
[29] Yuri L. Orlov, Vladimir N. Potapov, Complexity: an internet resource for analysis of DNA sequence complexity, Nucleic Acids Res. 32 (suppl_2) (2004) W628–W633.
[30] William Salerno, Paul Havlak, Jonathan Miller, Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments, Proc. Natl. Acad. Sci. 103 (35) (2006) 13121–13125.
[31] Sergio Arianos, Anna Carbone, Cross-correlation of long-range correlated series, J. Stat. Mech. Theory Exp. 2009 (03) (2009) P03037.
[32] Pedro Bernaola-Galván, Pedro Carpena, Cristina Gómez-Martín, Jose L. Oliver, Compositional structure of the genome: a review, Biology 12 (6) (2023) 849.
[33] Lonnie Baker, Charles David, Donald J. Jacobs, Ab initio gene prediction for protein-coding regions, Bioinform. Adv. 3 (1) (2023) vbad105.
[34] Yuriy L. Orlov, Nina G. Orlova, Bioinformatics tools for the sequence complexity estimates, Biophys. Rev. 15 (5) (2023) 1367–1378.
[35] Anna Carbone, Information measure for long-range correlated sequences: the case of the 24 human chromosomes, Sci. Rep. 3 (1) (2013) 2721.
[36] Sajia Akhter, Barbara A. Bailey, Peter Salamon, Ramy K. Aziz, Robert A. Edwards, Applying Shannon's information theory to bacterial and phage genomes and metagenomes, Sci. Rep. 3 (1) (2013) 1033.
[37] Renato Ferrero, Filippo Gandino, Anna Carbone, Information theoretic clustering of the human pangenome minigraph, Pattern Recognit. Lett. 191 (2025) 117–123.
[38] Yang Xiang, Xinrong Xiang, Yumei Li, Identifying rare variants for quantitative traits in extreme samples of population via Kullback–Leibler distance, BMC Genet. 21 (2020) 1–9.
[39] Mengxi Wei, Zhi Zheng, Weiping Zhang, A penalized regression-based biclustering approach in gene expression data, J. Syst. Sci. Complex. 38 (4) (2025) 1766–1783.
[40] Andrew Hart, María Paz Cortés, Mauricio Latorre, Servet Martinez, Codon usage bias reveals genomic adaptations to environmental conditions in an acidophilic consortium, PloS One 13 (5) (2018) e0195869.
[41] Dimitris Polychronopoulos, Labrini Athanasopoulou, Yannis Almirantis, Fractality and entropic scaling in the chromosomal distribution of conserved noncoding elements in the human genome, Gene 584 (2) (2016) 148–160.
[42] Vladimir Yu Kiselev, Tallulah S. Andrews, Martin Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, Nature Rev. Genet. 20 (5) (2019) 273–282.
[43] Jinxin Xie, Shanshan Ruan, Mingyan Tu, Zhen Yuan, Jianguo Hu, Honglin Li, Shiliang Li, Clustering single-cell RNA sequencing data via iterative smoothing and self-supervised discriminative embedding, Oncogene 43 (29) (2024) 2279–2292.
[44] Yi Zhang, Xi Feng, Yin Wang, Kai Shi, Deep learning powered single-cell clustering framework with enhanced accuracy and stability, Sci. Rep. 15 (1) (2025) 4107.
[45] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al., The complete sequence of a human genome, Science 376 (6588) (2022) 44–53.
[46] Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W. Hook, Sergey Koren, Mikko Rautiainen, Ivan A. Alexandrov, et al., The complete sequence of a human Y chromosome, Nature (2023) 1–11.
[47] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al., Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly, Genome Res. 27 (5) (2017) 849–864.
[48] Sergio Arianos, Anna Carbone, Christian Türk, Self-similarity of higher-order moving averages, Phys. Rev. E—Statistical, Nonlinear, Soft Matter Phys. 84 (4) (2011) 046113.
[49] Andrew T.A. Wood, Grace Chan, Simulation of stationary Gaussian processes in $[0, 1]^d$, J. Comput. Graph. Statist. 3 (4) (1994) 409–432.