SCIENTIFIC REPORTS

There are amendments to this paper

OPEN

Informational laws of genome structures

Vincenzo Bonnici[1,2] & Vincenzo Manca[1,2]

Received: 01 February 2016
Accepted: 09 June 2016
Published: 29 June 2016

In recent years, the analysis of genomes by means of strings of length *k* occurring in the genomes, called *k*-mers, has provided important insights into the basic mechanisms and design principles of genome structures. In the present study, we focus on the proper choice of the value of *k* for applying information theoretic concepts that express intrinsic aspects of genomes. The value $k = \lg_2(n)$, where *n* is the genome length, is determined to be the best choice in the definition of some genomic informational indexes that are studied and computed for seventy genomes. These indexes, which are based on information entropies and on suitable comparisons with random genomes, suggest five informational laws, to which all of the considered genomes obey. Moreover, an informational genome complexity measure is proposed, which is a generalized logistic map that balances *entropic* and *anti-entropic* components of genomes and is related to their evolutionary dynamics. Finally, applications to computational synthetic biology are briefly outlined.

The study of complexity in Biology is an old topic that often reemerges in theoretical biological investigations[1–3]. The study of complexity has very important implications for any deep understanding of the informational organization that life chooses in the different species to realize their specific biological functionalities. Entropy is a fundamental scientific concept that is naturally related to complexity and was the basis of statistical physics founded by Ludwig Boltzmann and the essence of his famous H theorem, which related the arrow of time to Boltzmann's equation, where entropy is expressed in terms of mechanical microstates[4]. Essentially, the same function was the basis of the information theory founded by Claude Shannon in 1948[5], where entropy is defined on information sources, that is, probability distributions over finite sets of elements (symbols, words or signals). A genome is essentially a text; if read at pieces of length *k* (called *k*-mers), a genome becomes an information source. Therefore genomic *k*-entropies can be easily defined, and the concepts and results of information theory can be applied[6–10].

In recent years, many studies have approached the investigation of DNA strings and genomes by means of algorithms, information theory and formal languages[11–22], and methods were developed for investigating whole genome structures. In particular, dictionaries of words occurring in genomes, distributions defined over genomes, and concepts related to word occurrences and frequencies have been very useful and seem to characterize important genomic features relevant in biological contexts[23–30]. Dictionaries are, in essence, finite formal languages. In genome analyses based on dictionaries, concepts from formal language theory, probability, and information theory are naturally combined by providing new perspectives in the investigation of genomes, which may disclose the internal logics of their structures.

The set of all *k*-mers, occurring in a given genome is a particular dictionary. A point that is crucial in genome analyses based on *k*-mers is the value of *k* that is more adequate for specific investigations. This issue becomes extremely evident when computing the entropy of a genome. We prove that preferential lengths exist for computing entropies, and in correspondence with these lengths, some informational indexes can be defined that exhibit "informational laws" and characterize an informational structure of genomes. As we have already noticed, there is a long tradition in investigating genomes by using *k*-mers. However, comparing genomes of different lengths, by using the same value of *k* (usually less than 12) may result in the loss, in some cases, of important regularities. In fact, the genomic laws that we discover emerge when the values of *k* are suitably defined from the logarithmic length of the genomes.

When genomic complexity is considered, it is very soon clear that it cannot be easily measured by parameters such as genome length, number of genes, CG-content, basic repeatability indexes, or their combinations. Therefore, we follow an information theoretic line of investigation based on k-mer dictionaries and entropies[16,26,27,31–33], which is aimed at defining and computing informational indexes for a representative set of

[1]University of Verona, Department of Computer Science, University of Verona, Verona 37134, Italy. [2]Center for BioMedical Computing, University of Verona, Verona, 37134, Italy. Correspondence and requests for materials should be addressed to V.M. (email: vincenzo.manca@univr.it)

genomes. This task is not trivial when genome sizes increase, so a specific software package is used to this end[31]. Moreover, an aspect that is missing in classical Shannon's conceptual apparatus is relevant in our approach: random strings and pseudo-random generation algorithms, which now can be easily produced and analyzed[34]. In fact, it is natural to assume that the complexity of a genome increases with its "distance" from randomness[35,36], as identified by means of a suitable comparison between the genome under investigation and random genomes of the same length. This idea alone provides important clues about the correct k-mer length to consider in our genome analyses, because theoretical and experimental analyses show that random genomes reach their entropic maxima for k-mers of length $\lg_2(n)$, where n is the genome length. No assumption on the distribution of probability of k-mers is assumed or inferred (as in Markov Models-based approaches); rather, data processing is developed on the basis of the empirical distributions of k-mers computed over the investigated genomes.

To this end, two basic indexes are introduced, which we call *entropic* and *anti-entropic components*. These indexes, and other related indexes, are computed over the chosen seventy genomes, ranging from prokaryotes to primates. The obtained values suggest some laws of genome structure. These laws hold in all of the investigated genomes and motivate the definition of the genomic complexity measure *BB* proposed in the paper. This measure depends on the entire structure of a genome and considers, together, the components of genomes (e. g., repeats, CpG, long range correlations, surely affecting entropies) without considering them separately. Moreover, as demonstrated below, *BB* is related to phylogeny but does not coincide with phylogenetic ordering. Certainly, primate genomes are usually more complex than, say, bacterial or insect genomes, but the situation is surely more critical because evolution is always active and a bacterium that we sequence today is not a type of bacteria that firstly arose in the tree of life. For this reason, genomes that are phylogenetically older can cumulate, even along different paths, "distances" from their corresponding random genomes comparable with those gained by "more evolved" genomes.

Results

The results presented in this paper are based on comparing real genomes with random genomes of the same length. As we show, any genome $\mathbb{G}$ of length n defines a partition of $\lg_4(n)$ in two addends $AC(\mathbb{G})$ and $EC(\mathbb{G})$ such that $AC(\mathbb{G}) + EC(\mathbb{G}) = \lg_4(n)$.

The fundamental informational components of genomes. We denote by $LG(\mathbb{G})$ the value $\lg_4(|\mathbb{G}|)$. Of course, $2LG(\mathbb{G}) = \lg_2(\mathbb{G})$. We call $LG(\mathbb{G})$ the logarithmic length of $\mathbb{G}$ and $2LG(\mathbb{G})$ the *double logarithmic length* of $\mathbb{G}$. When no possible confusion can arise, we avoid explicitly indicating $\mathbb{G}$, so we write in short LG, and consequently we denote the entropy $E_{2LG(\mathbb{G})}(\mathbb{G})$ over the $2LG(\mathbb{G})$-mers of $\mathbb{G}$ by $E_{2LG}(\mathbb{G})$ (analogous abbreviations are also adopted for other indexes). We also refer to the interval $[LG(\mathbb{G}),\ 2LG(\mathbb{G})]$ as the *critical entropic interval*. In the following, when $2LG$ is not integer, $E_{2LG}(\mathbb{G})$ denotes the linear interpolation between $E_{k_1}(\mathbb{G})$ and $E_{k_2}(\mathbb{G})$, where k_1, k_2 are the smallest integers such that $k_1 < 2LG < k_2$. In the case of the human genome, $2LG$ is between 31 and 32; in the genomes considered in this paper (from microbes to primates), it ranges between 16 and 36.

We prove, by using well-known results of information theory, that the values LG and $2LG$ have the following properties (see section *Methods*):

i). $2LG(\mathbb{G})$ is an upper bound to the values that entropy can reach over the genomes with the same length of $\mathbb{G}$;
ii). if k belongs to the critical interval $[LG(\mathbb{G}),\ 2LG(\mathbb{G})]$, and $|\mathbb{G}| = n$, then entropies E_k, for $k \leq n$, reach, on suitable genomes, the best approximations to $2LG(\mathbb{G})$ with an error close to zero, which is inferior to $|\lg_2(n/(n - \lceil \lg_2(n) \rceil))|$, being $\lceil x \rceil$ the closest integer greater than x.
iii). entropy $E_{2LG}(\mathbb{G})$ reaches its maximum in random genomes of length $n = |\mathbb{G}|$. This result follows from the fact that in random genomes of length n all $\lg_2(n)$-mers are hapaxes, that is, they occur once in the whole genome[37].

In conclusion, the maximum of $E_{2LG}(\mathbb{G})$ is almost equal to $2LG(\mathbb{G})$, and this maximum is reached by random genomes of length $|\mathbb{G}|$. It was realized that for all of the investigated genomes the following inequality immediately holds:

$$LG(\mathbb{G}) < E_{2LG}(\mathbb{G}) < 2LG(\mathbb{G}). \tag{1}$$

Therefore, we know that $E_{2LG}(\mathbb{G})$ belongs to the (open) real interval of bounds $LG(\mathbb{G})$ and $2LG(\mathbb{G})$. Then, we can define the following values $EC(\mathbb{G})$ and $AC(\mathbb{G})$, which we call *Entropic Component* and *anti-entropic Component* of $\mathbb{G}$, respectively:

$$EC(\mathbb{G}) = E_{2LG}(\mathbb{G}) - LG(\mathbb{G}) \tag{2}$$

$$AC(\mathbb{G}) = 2LG(\mathbb{G}) - E_{2LG}(\mathbb{G}). \tag{3}$$

Summing Equations (2) and (3), we obtain $AC(\mathbb{G}) + EC(\mathbb{G}) = LG(\mathbb{G})$. The value $EC(\mathbb{G})$ corresponds to the gap between the double logarithmic entropy $E_{2LG}(\mathbb{G})$ and the logarithmic length $LG(\mathbb{G})$, which is always positive according to the equations above. Moreover, $AC(\mathbb{G})$ is the gap between the double logarithmic length $2LG(\mathbb{G})$ and the entropy $E_{2LG}(\mathbb{G})$, which is positive because $2LG(\mathbb{G})$ is an upper bound to the entropies in the critical entropic interval. The term "anti-entropic" stresses an important difference with the analogous concept of *negh-entropy*, which is frequently used to denote the other side of the order/disorder dichotomy associated with

entropy (and its time arrow)[38–41]. In fact, in *anti-entropy*, no change of sign is involved, but a difference from an upper bound of the entropy is instead considered.

Informational genomic laws. Let us define $LX(\mathbb{G})$, called *lexical index*, as the ratio:

$$LX(\mathbb{G}) = |\mathbb{G}|/|D_m(\mathbb{G})| \tag{4}$$

The numerator is essentially the number of words of length $2LG$ occurring in random genomes, which as we already noticed are all hapaxes, and therefore, coincides with the number of possible occurrences of $2LG$-mers in $\mathbb{G}$. The denominator is the number of words of length $2LG$ occurring in $\mathbb{G}$. This ratio is related to the degree of order that $\mathbb{G}$ gains with respect to random genomes. In fact, in a random genome R, we have $LX(R) = 1$; therefore, in a real genome $\mathbb{G}$, $LX(\mathbb{G}) > 1$. The lexical index is smaller than the ratio $EC(\mathbb{G})/AC(\mathbb{G})$ but is greater than $LG(\mathbb{G})/EC(\mathbb{G})$. Moreover, by dividing and multiplying LX by $EC(\mathbb{G})$ and $LG(\mathbb{G})$, it is possible to obtain lower and upper bounds to $AC(\mathbb{G})$. The value $EH(\mathbb{G})$, given by $(EC(\mathbb{G}) - AC(\mathbb{G}))/LG(\mathbb{G})$, corresponds to the eccentricity of an ellipse associated with $\mathbb{G}$ (see Supplementary Information, Sup. Fig. 3). The product of $EH(\mathbb{G})$ with $LX(\mathbb{G})$ differs by 1 less than $AC(\mathbb{G})$. In conclusion, the following laws hold for all seventy investigated genomes:

$$EC(\mathbb{G}) > LX(\mathbb{G}) * AC(\mathbb{G}) \tag{5}$$

$$LX(\mathbb{G}) * EC(\mathbb{G}) > LG(\mathbb{G}) \tag{6}$$

$$LX(\mathbb{G}) * EC(\mathbb{G}) - LG(\mathbb{G})/LX(\mathbb{G}) > AC(\mathbb{G}) \tag{7}$$

$$EC(\mathbb{G}) - AC(\mathbb{G}) > LG(G)/LX(\mathbb{G}) \tag{8}$$

$$1 - AC(\mathbb{G}) < EH(\mathbb{G}) * LX(\mathbb{G}) < 1 + AC(\mathbb{G}). \tag{9}$$

Biobit: a measure of genomic complexity. As we already noticed, AC is an index measuring the informational distance between genomes and random genomes with the same length. This means that the more biological functions a genome $\mathbb{G}$ has acquired, the further the genome is from randomness. However, if we directly identify the complexity of $\mathbb{G}$ with $AC(\mathbb{G})$, we obtain some biologically inconsistent results. For example, *Zea mays* has an LG value of 15.4701 but an AC value of 3.6678 (primates have AC less than 1). These types of anomalies suggested to us that AC is surely related to the biological complexity of a genome, but this complexity is not a linear function of AC because also the EC component also has to be considered in a more comprehensive definition of complexity. Our search focused on a function that combines AC with EH, which is strictly related to EC. If x briefly denotes the *anti-entropic fraction* $AF = AC/LG$, it is easy to verify that because $EC = LG - AC$, then $EH = (EC - AC)/LG = (1 - 2x)$; therefore, the product $AC * EH$ can be represented by:

$$LGx(1 - 2x).$$

This function (after a simple change of variables) is a type of logistic map $ax(1 - x)$, with a constant, and x variable ranging in [0, 1], which is very important in population dynamics.

If we generalize $x(1 - 2x)$ in the class of functions $x^{\gamma}(1 - 2x)^{\delta}$, with γ and δ positive rationals weighting the two factors, then we discover that these functions have maxima for values approaching to zero when $\gamma \leq 1$ decreases and δ increases. Therefore, because AC is supposed to have a predominant role in the complexity measure, we define $BB_{\gamma,\delta}$ as $BB_{\gamma,\delta} = x^{\gamma}(1 - 2x)^{\delta}$ by choosing the values of the exponents in such a way that maxima of $BB_{\gamma,\delta}$ fall close to the values that the anti-entropic fraction AF assumes for the most part in genomes with high values of AC (almost all of them have medium horizontal eccentricity; see Supplementary Information, Sup. Table 2). No genome on our list reaches the maximum of the chosen function because their AF value is always smaller (suboptimal genomes) or greater (super-optimal genomes) than the value where the maximum is reached.

In conclusion, we conjecture that the genomic complexity is a non-linear function of AC having the form (apart from a multiplicative constant):

$$\sqrt{LG}\left(\frac{AC}{LG}\right)^{\gamma}\left(\frac{EC}{LG}\right)^{\delta} \tag{10}$$

In particular, the following definition, which is an instance of (10), was supposed to be the most appropriate ($\frac{AC}{LG} = AF$ and $\frac{EC}{LG} = 1 - AF$):

$$BB(\mathbb{G}) = \sqrt{LG(\mathbb{G})}\sqrt{AF(\mathbb{G})}(1 - 2AF(\mathbb{G}))^3. \tag{11}$$

In Fig. 1, the biobit values, together with the other described informational indexes, of the seventy genomes are visualized in a diagram. In Fig. 2 a flowchart is given that, in general terms, expresses the main stages for computing the BB measure of a given genome.

A further law could be associated with the biobit index, according to which genomes *evolve* by increasing the value of the BB function. This means that an ordering, denoted by $\succeq$ (a reflexive, antisymmetric, and transitive relation), can be defined such that:

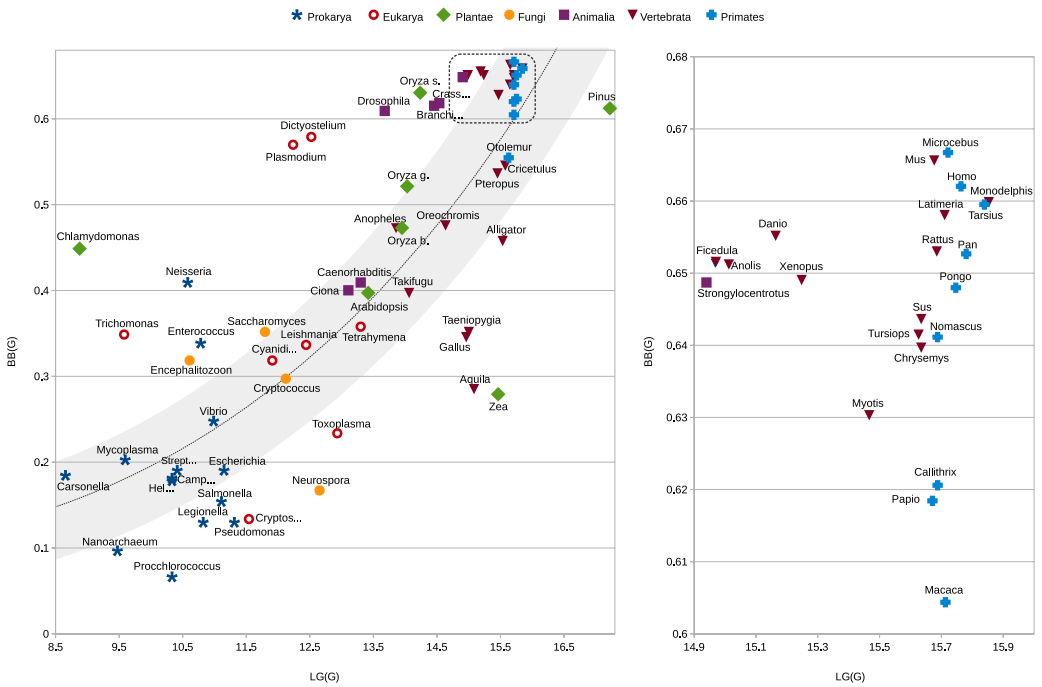

Figure 1. The left side of the figure shows the 70 analyzed genomes plotted on a Cartesian plane with their logarithmic length *LG*(G) as the abscissa and their biobit value *BB*(G) as the ordinate. An enlargement of the top-right region, which is highlighted with a dashed line, is shown on the right side of the image.

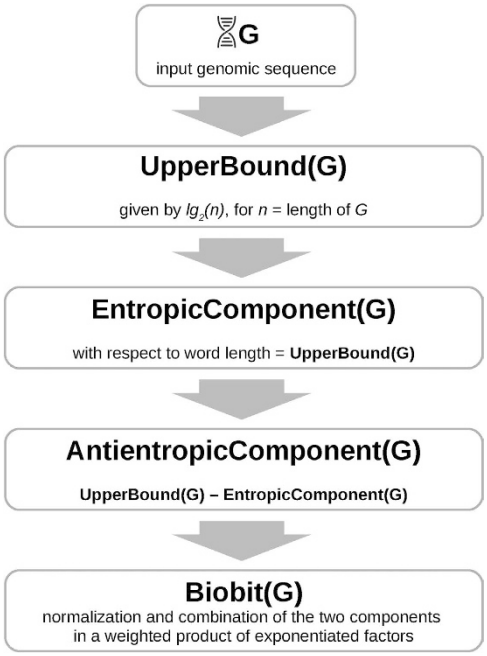

Figure 2. A flowchart of the computational steps involved in calculating *BB*(G). Given an input genome $\mathbb{G}$, an upper bound of maximum entropy is calculated, its value equals $2LG(\mathbb{G})$, and the value also defines the appropriate word length. Then the entropic and anti-entropic components are computed as, respectively, $EC(\mathbb{G})$ and $AC(\mathbb{G})$ and are successively normalized and combined by a weighted product into $BB(\mathbb{G})$.

$LG = \lg_4(\lvert\mathbb{G}\rvert)$	=	Logarithmic length
$EC = E_{2LG}(\mathbb{G}) - LG$	=	Entropic component
$AC = 2LG - E_{2LG}(\mathbb{G})$	=	anti-entropic component
$LX = \lvert D_{2LG}\rvert / \lvert\mathbb{G}\rvert$	=	Lexical index
$AF = AC/LG$	=	anti-entropic fraction
$EH = (EC - AC)/LG$	=	Horizontal eccentricity
$BB = \sqrt{LG}\sqrt{AF}\,(1 - 2AF)^3$	=	Biobit

Table 1. Main informational genomic indexes. $|D_{2LG}|$ is the number of $2LG$-mers occurring in $\mathbb{G}$, and $|\mathbb{G}|$ is the length of $\mathbb{G}$.

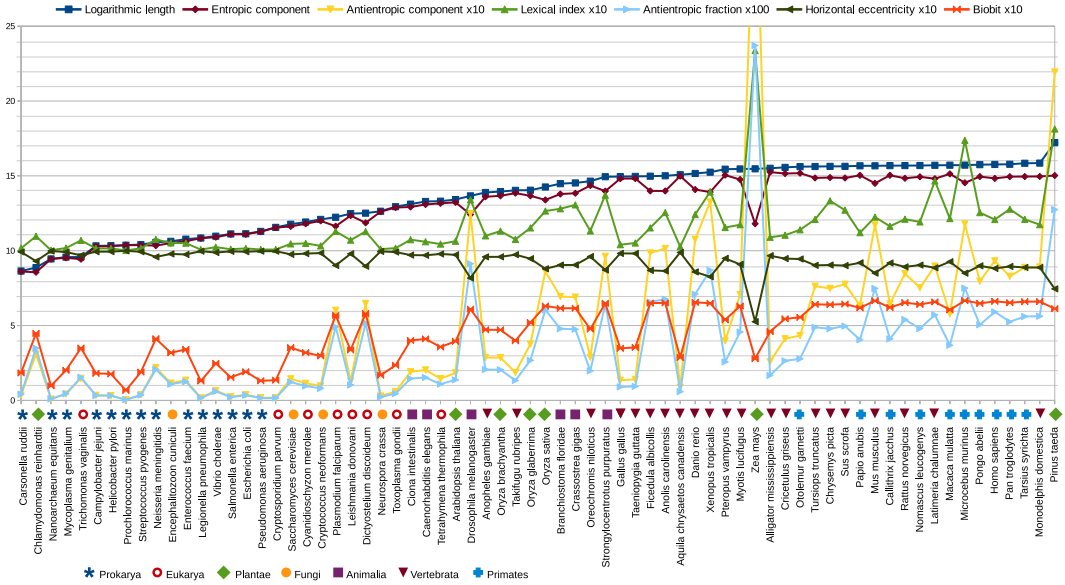

Figure 3. A chart of the main informational indexes. Some measures have been rescaled, by applying a factor of ten ($\times 10$) or one hundred ($\times 100$) to their value, to obtain a comprehensive overview. Species are arranged on the horizontal axis according to their genome length (increasing from left to right).

$$BB(\mathbb{G}_2) \geq BB(\mathbb{G}_1) \Rightarrow \mathbb{G}_2 \succeq \mathbb{G}_1. \tag{12}$$

Table 1 reports the main informational indexes based on the two entropic components of the logarithmic length of genomes. Figure 3 depicts graphically the values of these informational indexes for all of the investigated genomes (see Supplementary Information, Sup. Table 4, for the exact numerical values). The lengths of genomes are naturally linearly ordered, thus allowing us to arrange them along the x-axis. Apart from the EC curve, which is quite coincident with LG, the other indexes presents peaks that correspond to the genomes differing only slightly in lengths but differing greatly in other indexes.

It is interesting that, in essence, biological evolution is anti-entropic because the AC component, representing the tendency toward order, increases with the increase of biological functionalities, under the constraint of keeping the ratio AC/EC under a threshold, as expressed by the factor $(1 - 2x)^3$ of BB.

A 3D-visualization of our seventy genomes, by means of the AC, LX, BB informational indexes (see Supplementary Information, Sup. Fig. 2), reveals that genomic complexity does not coincide with classical phylogenetic classifications, as argued in the next section.

Discussion

We think that our informational indexes, and the laws relating them, confirm a very simple and general intuition. If life is information represented and elaborated by means of (organic) molecules, then the laws of information necessarily have to reveal the deep logic of genome structures.

The laws presented in the previous section represent universal aspects of genome structure and may rarely hold for strings of the same lengths that are not genomes. Therefore, the genomic complexity measure BB, obtained by means of informational indexes, is not a mathematical trick but must to be related to the way genomes are organized and to the way in which the genomes were generated. Figure 2 shows the values of BB along the 70 investigated genomes, and it is clear that BB is related to the evolutionary positions of organisms. However, our approach has an important biological implication in clarifying the difference between phylogenesis and genomic complexity, which are related but different concepts. In fact, several cases have been found (see Fig. 2 and Sup. Fig. 2

in Supplementary Information) where organisms that are phylogenetically more primitive than others, for example bacteria, have biobit values greater than those of "more evolved" organisms. The reason could be the following. A bacterium that we sequence today is an evolutionary product of some primitive bacterium. Even if we do not know the path from the bacterium's (possibly unknown) ancestor to the bacterium, its complexity along this path grew over time because its evolutionary age is the same as H. sapiens (even along different branches). The genomic complexity of $\mathbb{G}$ is, in a sense, a measure of the relevant steps from random genomes to $\mathbb{G}$. Surely, these steps reflect the evolutionary pressure and the biological interactions and competitions among species. However, if we forget this perspective, we lose an important aspect of evolutionary dynamics. This is why complexity-driven classifications that completely agree with phylogenesis are almost impossible. For example, we found that bacteria associated with human diseases have BB values significantly higher than others phylogenetically comparable to them. The BB measure is a sort of absolute distance from random, whereas phylogenesis concerns similarity or dissimilarity between species. Therefore, a very natural question arises, which suggests the development of the presented theory. Can entropic divergences (Kullback-Leibler divergence or similar concepts) be applied to phylogenetic analysis of genomes by means of "common words" and their probability distributions in the compared genomes? Finally, what is the applicability of our indexes in the identification of informational features that are relevant in specific pathological genetic disorders? Of course, these questions deserve specific investigations; however, our informational indexes with the related laws, and computational tools, provide a framework on which these informational analyses may be fruitfully set. We argue that it is almost impossible that functional changes do not correspond to precise informational alterations in the relationships expressed by the genomic laws. The challenge is in discovering the specific keys of these correspondences.

We developed some computational experiments showing a direct applicability of informational indexes and related genomic laws to the emergent field of synthetic biology. In fact, recent experiments on minimal bacteria[42] are based on the search for genome sequences obtained by manipulating and reducing some real genomes. It has been proved that after removing some parts of the *M. mycoides* genome, the resulting organism, JCVI-syn3.0 (531 kilobase pairs, 473 genes), is able to survive and has a genome smaller than that of any autonomously replicating cell found in nature (very close to *M. genitalium*). Of course, in this manner a better understanding of biological basic functions is gained, which directly relates with the investigated genome (removing essential portions results in life disruption). On the basis of this principle, we considered *M. genitalium* and removed some portions of its genome through a greedy exploration of the huge space of possibilities. At every step of our genome modifications (of many different types), we checked the validity of our genomic laws. We found that, after removing portions of the genome, some of our laws do not hold in the resulting sequences (see Supplementary Information, Sup. Table 6). Of course, these methods need to be carefully analyzed and validated with other examples and comparisons. However, a clear indication seems to emerge about the applicability of informational indexes and laws, possibly after suitable improvements to support and complement the development of genome synthesis and analysis, in the spirit of new trends in synthetic biology.

The starting point of our investigation was the comparisons of real genomes with random genomes of the same length. To accomplish this purpose, the right length of k-mers equal to the double logarithmic length of genomes was identified as being more appropriate for this comparison because for this length random genomes reach their maximum entropy. The difference between entropies was considered a measure of the order acquired by real genomes and corresponded to their capability of realizing biological functions. This intuition was supported by the values of indexes that we computed for an initial list of genomes. In fact, Sup. Table 3 in Supplementary Information provides *AC* values that, apart from two evident exceptions, seem to confirm the increasing of the *AC* value in accordance with the macroscopic biological complexity of organisms (independently from length, number of genes, or other typical genomic parameters). However, when we extended our analysis by including other genomes[43], we found *AC* values that were anomalous with respect to those already collected. In particular, plants provided extreme values, with no coherence with our interpretation of the *AC* index. To solve this puzzle, we considered a more comprehensive framework where *AC* and *EC* values interact in a trade-off between order and randomness. Genomes deviate from randomness, though to some extent, because genomes need a level of randomness that is sufficient to keep their evolutionary nature, based on a random exploration of new possibilities of life (filtered by natural selection).

In this picture, the two quantities $EC(\mathbb{G})$ and $AC(\mathbb{G})$ seem to correspond to the informational measure of two important aspects of genomes: *evolvability* and *programmability* (in the sense of[2]). Evolvability measures the random component of genomes, whereas programmability measures the order that genomes gain with respect to pure random genomes by acquiring biological functions. The non-random meaning of *AC* can be mathematically characterized in terms of Kullback-Leibler entropic divergence between the probability distribution of words of $\mathbb{G}$ and the probability distribution of the same words in random genomes[44].

Genome evolution is realized through an interplay of programmability and evolvability. The anti-entropic component *AC* cannot increase beyond a percentage of the logarithmic length because $LG = AC + EC$ and therefore increase of *AC* implies a decrease of *EC* by reducing the evolutionary ability. Therefore, the only way to increase *AC*, by keeping a good balance of the two components, is to increase the value of *LG*, i. e., the genome length, which explains why genomes increase their length during evolution. However, this increase is only indirectly correlated with biological complexity, as apparent in Fig. 1 (see also Supplementary Information, Sup. Table 3).

The definition of genomic complexity, in terms of a nonlinear function of *AC*, is related to the balance between *AC* and *EC* values. Some of the genome entropic laws continue to also hold for k-mers with $k > \lg_2(|\mathbb{G}|)$, but almost none of the laws continue to hold when $k < \lg_2(|\mathbb{G}|)$. For example, for $k = 6$ and $k = lg_4(|\mathbb{G}|)$, the values of *AC* completely lose the logic that they have for $k = lg_2(|\mathbb{G}|)$, by showing dramatic changes with respect to $k = lg_2(|\mathbb{G}|)$, on which our indexes are based (see Supplementary Information, Sup. Table 4). Of course, we could compare real and random genomes also for values shorter than $\lg_2(|\mathbb{G}|)$, but in this case, we need to gener-

ate random genomes and compute the corresponding entropies, whereas for $k = \lg_2(|\mathbb{G}|)$, we do not need such generations and computations, because we know, by theoretical arguments (see Proposition 3) that in random genomes, entropies at double logarithmic lengths can be assumed to be equal to $\lg_2(|\mathbb{G}|)$.

Our investigation can be compared to the astronomical observations measuring positions and times in the orbits of celestial objects. Kepler's laws arose from the regularities found in planetary motions, and from Kepler's laws, the laws of mechanics emerged. This astronomical comparison, which was an inspiring analogy, revealed a surprising coincidence when ellipses were introduced in the representation of entropic and anti-entropic components. Kepler's laws were explained by Newton's dynamical and gravitational principles. Continuing our analogy, probably deeper informational principles are the ultimate reason for the laws that we found.

Methods

The seventy investigated genomes include prokaryotes, algae, amoebae, fungi, plants, and animals of different types. In Sup. Table 5 of Supplementary Information, source data bases, assembly identifiers, genome lengths, and percentages of unknown nucleotides are given. Basic concepts from information theory, probability theory, and formal language theory can be found in classical texts in these fields[5,45,46].

Basic definitions and notation. Strings are finite sequences of contiguous symbols. Mathematically, strings are functions from a set of positions, viewed as a subset of the set $\mathbb{N}$ of natural numbers, $\{i \in \mathbb{N} | 1 \geq i \geq n\}$ to a set of symbols, called *alphabet*. The number n is called the length of the string. We denote generic strings with Greek letters (possibly with subscripts) and reserve λ for the empty string (useful for expressing mathematical properties of strings). The length of a string α is denoted by $|\alpha|$, and $\alpha[i]$ is the symbol occurring in α at position i, whereas $\alpha[i, j]$ is the string occurring in α between the positions i and j (both included).

Let us consider the genomic alphabet of four symbols (characters, or letters, associated with nucleotides) $\{a, c, g, t\}$. The set $\{a, c, g, t\}^*$, as usual, denotes the set of all possible strings over $\{a, c, g, t\}$. A genome $\mathbb{G}$ is representable by a string of $\{a, c, g, t\}^*$, where symbols that occur, from the first to the last position, are written in the order that they occur, from left to right, according to the standard writing system of Western languages, and according to the chemical orientation 5′–3′ of DNA molecules.

Substrings $\mathbb{G}[i, j]$ of length k, where $1 \leq i \leq j \leq |\mathbb{G}|$, are also called *k-words*, *k-factors*, *k-mers* of $\mathbb{G}$ (k may be omitted, when it is not relevant). We remark that the absolute value notation $|-|$ used for string length has different meaning when applied to sets or multisets. In fact, for a finite set A, then $|A|$ denotes its cardinality, whereas for a finite multiset X (set of elements that possibly occur in many "identical" copies, with no relevance for occurrence order) $|X|$ denotes its size (the sum of the elements of X each counted all the times that the element occurs).

A *dictionary of* $\mathbb{G}$ is a set of strings occurring in $\mathbb{G}$. We denote by $D_k(\mathbb{G})$ the dictionary of all k-mers occurring in G. It is easy to verify that the number of occurrences of k-mers in $\mathbb{G}$ is $|\mathbb{G}| - k + 1$ ($|\mathbb{G}|$ is the length of $\mathbb{G}$) and corresponds to the maximum cardinality $|D_k(\mathbb{G})|$ reachable by a dictionary of k-mers within genomes of the same length of $\mathbb{G}$.

A word α of D can occur in $\mathbb{G}$ many times. We denote by $mult_{\mathbb{G}}(\alpha)$ its *multiplicity* in $\mathbb{G}$, that is, the number of times α occurs in $\mathbb{G}$. A word of $\mathbb{G}$ with multiplicity greater than 1 is called a *repeat* of $\mathbb{G}$, whereas a word with multiplicity equal to 1 is called a *hapax* of $\mathbb{G}$. This term is used in philological investigation of texts, but it is also adopted in document indexing and compression[37]. The values of word multiplicities can be normalized if we divide the word multiplicities by the sum of the multiplicities of all the words occurring in $\mathbb{G}$. This normalization corresponds to replacing multiplicities with frequencies, which can be seen as percentages of multiplicity.

Many important indexes related to characteristics of genome dictionaries can be defined on genomes. For example, $mrl(\mathbb{G})$ is the length of the longest repeats of $\mathbb{G}$. Of course, $mrl(\mathbb{G}) + 1$ is the minimum length, such that k-mers with k greater than $mrl(\mathbb{G})$ are all hapaxes.

Shannon used the term *information source* as synonymous with discrete probability distribution to introduce the notion of (information) *entropy*. Given a distribution of probability p, over a finite set A, its entropy is given by $-\sum_{x \in A} p(x) \lg_2(p(x))$. We remark that if $-\lg_2(p(x))$ is considered to be the information associated with the occurrence of $x \in A$ (the more improbable x is, the more its occurrence is informative), then entropy is the mean (in a probabilistic sense) quantity of information emitted by the information source (A, p).

An intrinsic property of entropy is its *Equipartition Property*, that is, in the finite case, the fact that entropy reaches its maximum value $\lg_2(|A|)$, when p is equally distributed, that is, when $p(x) = 1/|A|$, for all $x \in A$ ($|A|$ is the number of elements of A).

A genome $\mathbb{G}$ is any sequence over the alphabet $\{a, c, g, t\}$. This definition includes real genomes and ideal genomes, with no biological meaning, which are important in the mathematical analysis of genomes, as "material points" are essential in physics for discovering motion laws. Any subsequence of contiguous symbols of $\mathbb{G}$ is called a string, word, or k-mer of $\mathbb{G}$ (k explicitly expresses the length).

The *empirical k-entropy* $E_k(\mathbb{G})$ of $\mathbb{G}$ is given by (the adjective empirical refers to the use of frequencies):

$$E_k(\mathbb{G}) = -\sum_{\alpha \in D_k(\mathbb{G})} p(\alpha) \lg_2(p(\alpha)). \tag{13}$$

We remark that the entropy $E_k(\mathbb{G})$ is computed only with the k-mers occurring in $\mathbb{G}$ (see definition of $D_k(\mathbb{G})$). The computation of $E_k(\mathbb{G})$ becomes prohibitive when $\mathbb{G}$ has length of order 10^9 and $k > 20$. Therefore, we used suffix arrays[47] in the computation of genomic dictionaries.

A *Bernoullian*, or random, genome is a synthetic genome generated by means of casual (blind) extractions (with insertion after extraction) from an urn containing four types of balls, in equal numbers of copies, completely

length	*min*	*max*	*sd*	*avg*	$lg_2(\|R\|)$
1,000	9	15	1.07	10.2	9.97
100,000	15	20	0.95	16.67	16.61
200,000	16	21	0.86	17.78	17.61
500,000	18	23	0.91	19.09	18.93
1,000,000	18	24	0.96	20.14	19.93
10,000,000	22	26	0.97	23.49	23.25
20,000,000	23	27	0.93	24.31	24.25
30,000,000	24	30	1.14	25.08	24.84
50,000,000	24	31	1.17	25.86	25.58
75,000,000	25	29	0.85	26.44	26.16
100,000,000	25	30	1.02	26.89	26.58

Table 2. For each genome length, 100 trials were performed. The minimum, the maximum and the average, together with the standard deviation, of $mrl+1$ was computed for each trial set. With a good approximation $\lg_2(|R|) \approx avg(mrl(R)+1)$.

identical apart from their colors, denoted by the genomic letters *a*, *c*, *g*, *t*. Pseudo-Bernoullian genomes can be generated by means of (pseudo) random generators available in programming languages (by suitable encoding of genomic symbols). We denote by RND_n the class of Bernoullian genomes of length *n*.

The computations of the main informational indexes, given in Table 1, extract the set of $\lg_2(|\mathbb{G}|)$-mers occurring in the considered genomes, where $\lg_2(|\mathbb{G}|)$ varies from 16 to 36, by means of a dedicated software, based on suffix arrays, called InfoGenomics Tools (shortly IGTools)[31], which is an efficient suite of interactive tools mainly designed for extracting *k*-dictionaries, computing on them distributions and set-theoretic operations, and finally evaluating empirical entropies E_k, and informational indexes, for different and even very large values of *k*.

In Supplementary Information, a 3D-visualization (Sup. Fig. 2) of 70 genomes is given with respect to *BB*, *AC*, *LX* axes, where Principal Component Analysis is applied for a better visualization. A taxonomy tree of the 70 genomes has been built via the NCBI taxonomy[48] (see Supplementary Information, Sup. Fig. 1).

Mathematical Backgrounds. In the following, some propositions are given, which were essential to the identification of parameters on which information entropies are computed. Let us start with the following question. Given a genome length *n* and a value $k \leq n$, which is the maximum value of $E_k(\mathbb{G})$ in the class of genomes of length *n*? We answer to the question above with Proposition 3, which is based on two Lemmas.

Lemma 1 *Given a genome* $\mathbb{G}$ *of length n, if* $k = mrl(\mathbb{G}) + 1$, *then* $E_k(\mathbb{G})$ *is the maximum value that* E_k *can reach in the class of all possible genomes of length n.*

Proof. The minimum value of *k* such that all *k*-mers are hapaxes of $\mathbb{G}$ is $mrl(\mathbb{G}) + 1$. Therefore, if $k = mrl(\mathbb{G}) + 1$, then $E_k(\mathbb{G})$ is maximum, according to the entropy Equipartition Property, because we have the maximum number of words occurring once in $\mathbb{G}$, and all these words have the same probability of occurring in $\mathbb{G}$. □

Lemma 2 *If R is a random genome of length n, then*

$$\lfloor \lg_2(n) \rfloor - 1 \leq mrl(R) + 1 \leq \lceil \lg_2(n) \rceil .$$

Proof. Let RND_n the class of random genomes of length *n*. If $k = mrl(R)+1$, the probability that a *k*-mer occurs in $R \in RND_n$ is $(n-k+1)/4^k$, and the probability that it occurs exactly once in *R* (being all *k*-mer hapaxes) is $1/(n-k+1)$. Therefore, by equating these two probabilities we get:

$$(n - k + 1)/4^k = 1/(n - k + 1) \tag{14}$$

that is:

$$(n - k + 1)^2 = 4^k \tag{15}$$

that implies (*k* has to be an integer) that the minimum length *k* for having all hapaxes in *R* is:

$$k = \lceil \lg_4(n - k + 1)^2 \rceil \tag{16}$$

whence

$$k = \lceil \lg_2(n - k + 1) \rceil \tag{17}$$

that is

$$k = \lceil \lg_2(n - \lceil \lg_2(n - k + 1) \rceil + 1) \rceil \tag{18}$$

therefore

$$\lceil \lg_2(n - \lceil \lg_2(n) \rceil) \rceil \leq k \leq \lceil \lg_2(n) \rceil \tag{19}$$

that implies the asserted inequality. □

Table 2 shows an experimental validation of Lemma 2. It confirms that $\lg_2(|R|)$ results to be a good estimation of the average of $mrl(R)+1$ in $RND_{|\mathbb{G}|}$.

Proposition 3 *In the class of genomes of length n, for every $k < n$, the following relation holds*

$$E_k(\mathbb{G}) < \lg_2(n). \tag{20}$$

Moreover, random genomes of length n have entropies differing from the upper bound $\lg_2(n)$ less than $\lg_2(n/(n - \lceil \lg_2(n) \rceil))$ (close to zero).

Proof. According to Lemma 1, $E_k(\mathbb{G})$ reaches its maximum, when $k = mrl(\mathbb{G}) + 1$. In this case:

$$E_k(\mathbb{G}) = \lg_2(n - k + 1) \tag{21}$$

therefore, the difference $\lg_2(n) - E_k(\mathbb{G})$ is given by:

$$\lg_2(n) - \lg_2(n - k + 1) = \lg_2(n/(n - k + 1)). \tag{22}$$

If $\mathbb{G}$ belongs to the class of random genomes of length n, according to Lemmas 1 and 2, the maximum entropy is given by $E_k(\mathbb{G})$, for $k = mrl(\mathbb{G}) + 1$, with $\lfloor \lg_2(n) \rfloor - 1 \leq k \leq \lceil \lg_2(n) \rceil$. Therefore, by substituting in equation (22) the upper bound of k, giving the upper bound of $\lg_2(n/(n - k + 1))$, we get: $\lg_2(n) - \lg_2(n - k + 1) \leq \lg_2(n/(n - \lceil \lg_2(n) \rceil + 1) < \lg_2(n/(n - \lceil \lg_2(n) \rceil)$. □

References

1. Conrad, M. *Adaptability* (PlenumPress, 2001).
2. Conrad, M. The price of programmability. In *A half-century survey on The Universal Turing Machine*, 285–307 (Oxford University Press, 1988).
3. Holland, J. & Mallot, H. Emergence: from chaos to order. *Nature* **395,** 342–342 (1998).
4. Cercignani, C. *The Boltzmann Equation and Its Application* (Springer, 1988).
5. Shannon, C. E. A mathematical theory of communication. *Bell Sys Tech J* **27,** 623–656 (1948).
6. Pincus, S. M. Approximate entropy as a measure of system complexity. *P Nat Acad Sci* **88,** 2297–2301 (1991).
7. Crochemore, M. & Vérin, R. Zones of low entropy in genomic sequences. *Computers & chemistry* **23,** 275–282 (1999).
8. Vinga, S. & Almeida, J. S. Local Renyi entropic profiles of DNA sequences. *BMC bioinformatics* **8,** 393 (2007).
9. Koslicki, D. Topological entropy of dna sequences. *Bioinformatics* **27,** 1061–1067 (2011).
10. Wang, D., Xu, J. & Yu, J. KGCAK: a K-mer based database for genome-wide phylogeny and complexity evaluation. *Biol direct* **10(1),** 1–5 (2015).
11. Head, T. Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *B Math Biol* **49,** 737–759 (1987).
12. Deonier, R. C., Tavaré, S. & Waterman, M. *Computational genome analysis: an introduction* (Springer, 2005).
13. Manca, V. & Franco, G. Computing by polymerase chain reaction. *Math Biosci* **211,** 282–298 (2008).
14. Searls, D. B. Molecules, languages and automata. In *Grammatical Inference: Theoretical Results and Applications*, 5–10 (Springer, 2010).
15. Vinga, S. Information theory applications for biological sequence analysis. *Brief Bioinform*, doi: 10.1093/bib/bbt068 (2013).
16. Manca, V. *Infobiotics: information in biotic systems* (Springer, 2013).
17. Gatlin, L. L. The information content of DNA. *J Theor Biol* **10(2),** 281–300 (1966).
18. Kraskov, A. & Grassberger, P. MIC: mutual information based hierarchical clustering. *Info Theor Stat Learn*, 101–123 (Springer, 2009).
19. Campbell, A., Mrázek, J. & Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *P Nat Acad Sci* **96(16),** 9184–9189 (1999).
20. Ebeling, W. & Jiménez-Montaño, M. A. On grammars, complexity, and information measures of biological macromolecules. *Math Biosci* **52(1),** 53–71 (1980).
21. Weiss, O., Jiménez-Montaño, M. A. & Herzel, H. Information content of protein sequences. *J Theor Biol* **206(3),** 379–386 (2000).
22. Holste, D., Grosse, I. & Herzel, H. Statistical analysis of the DNA sequence of human chromosome 22. *Phys Rev E* **64(4),** 041917 (2001).
23. Fofanov, Y. *et al.* How independent are the appearances of n-mers in different genomes? *Bioinformatics* **20,** 2421–2428 (2004).
24. Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics* **9(1),** 517 (2008).
25. Chor, B. *et al.* Genomic dna k-mer spectra: models and modalities. *Genome Biol* **10,** R108 (2009).
26. Castellini, A., Franco, G. & Manca, V. A dictionary based informational genome analysis. *BMC genomics* **13,** 485 (2012).
27. Bonnici, V. & Manca, V. Recurrence distance distributions in computational genomics. *Am J Bioinformat Comput Biol* **3,** 5–23 (2015).
28. Wen, J., Chan, R. H., Yau, S.-C., He, R. L. & Yau, S. S. k-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546,** 25–34 (2014).
29. Almirantis, Y., Arndt, P., Li, W. & Provata, A. Editorial: Complexity in genomes. *Comp Biol Chem* **53,** 1–4 (2014).
30. Hashim, E. K. M. & Abdullah, R. Rare k-mer DNA: Identification of sequence motifs and prediction of CpG island and promoter. *J Theor Biol* **387,** 88–100 (2015).
31. Bonnici, V. & Manca, V. Infogenomics tools: A computational suite for informational analysis of genomes. *J Bioinfo Proteomics Rev* **1,** 8–14 (2015).
32. Manca, V. *Infogenomics: genomes as information sources*. Chap. 21, 317–324 (Elsevier, Morgan Kauffman, 2016).
33. Manca, V. Information theory in genome analysis. In *Membrane Computing*, LNCS 9504, 3–18 (Springer, 2015).
34. Knuth, D. *The art of computer programming*, volume 2: Seminumerical algorithms (Addison-Wesley, 1998).
35. Kong, S. G. *et al.* Quantitative measure of randomness and order for complete genomes. *Phys Rev E* **79(6),** 061911 (2009).

36. Jiang, Y. & Xu, C. The calculation of information and organismal complexity. *Biol Direct* **5(59),** 565 (2010).
37. Witten, I. H., Moffat, A. & Bell, T. C. *Managing gigabytes: compressing and indexingdocuments and images* (Morgan Kaufmann, 1999).
38. Wiener, N. *Cybernetics or control and communication in the animal and the machine* (Hermann, Paris, 1948).
39. Schrödinger, E. *What Is Life? The Physical Aspect of the Living Cell and Mind* (Cambridge University Press, 1944).
40. Brillouin, L. The negentropy principle of information. *J Appl Phys* **24,** 1152–1163 (1953).
41. Volkenstein, M. V. *Entropy and information* (Springer, 2009).
42. Venter, J. C. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351,** 6280 (2016).
43. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
44. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann Math Stat*, 79–86 (1951).
45. Feller, W. *An Introduction to Probability Theory and Its Applications* (Wiley & sons, 1968).
46. Rozenberg, G. & Salomaa, A. *Handbook of Formal Languages: Beyonds words* vol. 3 (Springer, 1997).
47. Abouelhoda, M. I., Kurtz, S. & Ohlebusch, E. Replacing suffix trees with enhanced suffix arrays. *J Discrete Algorithms* **2,** 53–86 (2004).
48. Federhen, S. The NCBI taxonomy database. *Nucleic acids res* **40,** D136–D143 (2012).

Acknowledgements

The authors are grateful to Andres Moya for his support and help, and to Rosalba Giugno for her important suggestions.

Author Contributions

V.M. conceived the theoretical and mathematical setting of the paper, V.B. developed the software and computations, and V.M. and V.B. analyzed the results. V.M. wrote the paper, and both authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Bonnici, V. and Manca, V. Informational laws of genome structures. *Sci. Rep.* **6**, 28840; doi: 10.1038/srep28840 (2016).