Universal Global Imprints of Genome Growth and Evolution – Equivalent Length and Cumulative Mutation Density

Hong-Da Chen¹,², Wen-Lang Fan²,³, Sing-Guan Kong¹,², Hoong-Chien Lee¹,²,⁴,⁵ *

¹ Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan, ² Department of Physics, National Central University, Chungli, Taiwan, ³ Genomic Research Center, Academia Sinica, Taipei, Taiwan, ⁴ Cathay Medical Research Institute, Cathay General Hospital, Taipei, Taiwan, ⁵ National Center for Theoretical Science, Shinchu, Taiwan

Abstract

**Background:** Segmental duplication is widely held to be an important mode of genome growth and evolution. Yet how this would affect the global structure of genomes has been little discussed.

**Methods/Principal Findings:** Here, we show that equivalent length, or \( L_e \), a quantity determined by the variance of fluctuating part of the distribution of the \( k \)-mer frequencies in a genome, characterizes the latter's global structure. We computed the \( L_e \) of 865 complete chromosomes and found that they have nearly universal but \( (k\text{-dependent}) \) values. The differences among the \( L_e \) of a chromosome and those of its coding and non-coding parts were found to be slight.

**Conclusions:** We verified that these non-trivial results are natural consequences of a genome growth model characterized by random segmental duplication and random point mutation, but not of any model whose dominant growth mechanism is not segmental duplication. Our study also indicates that genomes have a nearly universal cumulative “point” mutation density of about 0.73 mutations per site that is compatible with the relatively low mutation rates of \((1 \sim 5) \times 10^{-3} \text{site/Mya}\) previously determined by sequence comparison for the human and *E. coli* genomes.


Editor: Josh Bongard, University of Vermont, United States of America

Received November 4, 2009; Accepted February 8, 2010; Published April 14, 2010

Copyright: © 2010 Chen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the National Science Council (ROC) (http://web1.nsc.gov.tw/mp.aspx?mp=7), Cathay General Hospital (http://www.cgh.org.tw/en/index.html), National Central University (http://www.ncu.edu.tw/e_web/index.php). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hlee@phy.ncu.edu.tw

Introduction

Evolution has many facets, and one that is particularly accessible to quantitative analysis is the evolution of genomic sequences. In particular, the study of point mutations (here used in the sense that includes relatively small insertions and deletions, or indels) on genes has led to deep understandings of many aspects of genome evolution [1,2]. Point mutation however cannot be the main force driving genome growth, because it does not give rise to gene duplication [3–8], and because the pace of evolution based on point mutation alone would be too slow. Gene duplication is a product of segmental duplication (SD). In fact, genomes are replete with vestiges of duplication [9–11], not only in the form of homologous genes, but also as transposons [12–14], pseudogenes [15–18], and many other types of coding and non-coding repeats [19–22]. There is also evidence of large-scale genomic rearrangements [23–27] and whole genome duplications [3,28–30]. This has led to the generally held view that SD is an important mode of genome growth and evolution.

If products of SD are so prevalent in genomes, we expect the SD’s in a genome, collectively, to leave a large imprint on the global structure of its host, one that is detectable using means not relying on sequence alignment, which in any case is not suitable for global studies. One may reasonably expect a study to understand the formation of such an imprint to yield useful insights into the global pattern of genome growth and evolution, yet no such effort has been made.

Here, we study the statistical properties of genomes by analyzing the distribution of the frequency of occurrence, or FD, of \( k \)-letter words, or \( k \)-mers, in the sequence. Although genomic FDs have been much studied before [31–36], the method and focus of the present study are both distinct from all previous studies. A novel approach we use, crucial to our ability to extract results presented here, is the separation of the contributions to the variance from the fluctuating part of an FD (FFD), and the non-fluctuating part (NFFD). We show that NFFD is entirely understood; it carries no statistical information other than the base composition of a sequence. A genomic sequence and its matching random sequence have essentially the same NFFD. The contribution from NFFD overwhelmingly dominates the variance (of an FD) of a random sequence in all cases and dominates the variance of a genome except when its base composition is approximately even. As a consequence, if the separation mentioned above is not carried out, then it is sometimes easy to distinguish genomic from random sequences and sometimes not, a situation that has confounded many previous studies. We will demonstrate that the very special
characteristics of genomic FFDs sharply distinguishes them from their random counterparts under all circumstances.

In this study we used the FFD to define the equivalent lengths ($L_k$'s; one for each $k$) of a sequence and discovered a universality in these quantities. We then identify these $L_k$'s and their small values, as a clear and distinct global imprints of genome growth and evolution. (The $L_k$ of a sequence is inversely proportional to the FFD part of the variance and is defined such that the $L_k$ of a random sequence is its own true length. Therefore, a sequence whose equivalent length is $L_k$ has the characteristic randomness of a random sequence of length $L_k$.) We computed the $L_k$ of about 900 complete chromosomes, all the complete sequences at the time of download from GenBank, for $k = 2$ to 10, and found some unexpected and useful results: Roughly, the complete set of about 7400 $k$-dependent whole-chromosome $L_k$'s is well represented by the universal formula $L_k = L_{2k} = 2^{a_k} (2^k - 2)$, where $L_{2k} \approx 310^{1.290} 130^{0.92}$ (base pair) and $a_k = 0.92$. The formula means that, for the smaller $k$'s, the universal genomic $L_k$ is only a small fraction of the genome length even for the shortest genomes. Another unexpected result is the small difference between the $L_k$'s of coding and non-coding parts. In our successful attempt to describe these results in a simple genome growth model driven by random segmental duplication, we obtained a universal cumulative point mutation density of $r = 0.73 \pm 0.07$ site for genomes. This value is compatible with the relatively low mutation rates previously determined by sequence comparison for the human and E. coli genomes [37–39].

**Results**

Only FFD contains non-trivial information

A key to our approach to the analysis of genomic sequences is the decomposition of $CV^2 - CV$ is the coefficient of variation of an FD – into FFD and NFFD components (Methods). This is illustrated in Fig. 1, which shows the values of $CV^2$ for 2-mers; results for other $k$'s are similar. The full $CV^2$ of genomic sequences (Fig. 1(a)) differs from that of their matching random sequences (Fig. 1(b)) clearly only when $|p - 0.5| \leq 0.1$, where $p$ is the fractional A/T-content. (A genome and its matching random sequence have the same length and base composition.) The situation becomes much clearer when $CV^2$ is decomposed into its FFD and NFFD parts, $CV_{ff}^2$ and $CV_{nf}^2$, respectively. While the values of $CV_{nf}^2$ for the two type of sequences are almost indistinguishable (red) triangles, Fig. 1(c,d); the two “volcano” curves are identical, being both given by the theoretical prediction, Eq. (12)), the values of $CV_{ff}^2$ for genomes and random sequences are drastically different (blue) bullets, Fig. 1(c,d)). The genomic $CV_{ff}^2$ span a narrow band ranging from 0.01 to 0.1, while the random $CV_{ff}^2$ are several orders of magnitude smaller. In fact for random sequences the value of $CV_{ff}^2$ is well understood to be inversely proportional to sequence length (Eq. (13), and below). Clearly, if random sequences are used as controls to discuss the non-random properties of genomic sequences when the distinction between FFD and NFFD is not made, then it is possible that conflicting conclusions [32,40–43] may be drawn.

![Figure 1. Fluctuating and non-fluctuating parts of variance.](http://www.plosone.org/)

(a) Variances of 2-mer frequency distribution of 865 complete sequences. (b) Same as (a) but for 865 matching random sequences. Bottom: same data as in top plots, but with each variance split into non-fluctuating (triangles) and fluctuating (bullets) parts, for (c) genomes and (d) matching random sequences. The “volcanic” curves through the non-fluctuating data in (c) and (d) plot theoretical values given by Eq. (12).

doi:10.1371/journal.pone.0009844.g001
Genomic Ls is approximately a constant of sequence length

Throughout this paper we use Ls to denote generically the equivalent length of any sequence (Eq. (14), Methods), and reserve Lw for denoting entire sequences such as a complete chromosomes. Fig. 2 shows Ls versus segment length Ls for segments taken from the chromosomes of four model organisms: E. coli K12; C. elegans, Chr. (chromosome) 1; A. thaliana, Chr. 1; H. sapiens, Chr. 1, and matching random sequences. The computation is carried out only when Ls is at least four times 4^k, since for shorter lengths the systematic error becomes too large. It is seen that whereas the Ls of random sequences closely tracks Ls, as expected, the Ls of genomic sequences quickly levels off to a saturation value Lm(k). These results for Ls ≥ 5 kb may be summarized in terms of the scaling relation Ls ∝ (Ls)^γ. Then we have the two distinct classes γ ≈ 1 for random sequences and γ ≈ 0 for genomic sequences. This scaling relation is not the same as the long-range correlation and scale-invariance observed in binary analyses of long genomic sequences [44–46]. In Fig. 2 Ls is seen not to depend strongly on organism. For small k, Ls(k) is diminutive relative to genome length: ~ 0.35 and ~ 1.0 kb when k = 2 and 4, respectively, growing to ≤ 600 kb when k = 10. Within a genome, the apparent invariance of CV (not CVp) with respect to segment length was noted in [47–49] and the relation between Shannon information and a quantity similar to CVp was discussed in [50].

Whole chromosomes have nearly universal Lw(k)

A list of the 865 complete chromosomes studied here is given in Table S1, and a list of Lw(k)’s, k = 2 to 10, for the chromosomes is give in Table S2. Fig. 3 shows Lw(k), as a function of p (top panels) and chromosome length L (bottom panels), computed from the complete chromosomes for even k’s up to k = 10. Table 1 gives the Lw(k), k = 2 to 10, of chromosomes of seven model organisms. It is seen that Lw(k) has a clear dependence on k, is essentially independent of sequence length, and has a weak dependence on p. Fig. 4 gives Lw(k) for odd k’s averaged over categories of organisms and over chromosomes in model organisms (for more detailed results see Table S3). The k = 5 data reconfirms the absence in Lw of a systematic dependence on chromosome length (similarly for other k’s). In the k = 3 and 7 plots Lw’s are given separately for the whole chromosome, and genomic (gn) and intergenic (ig) regions. The unicellulars are seen to have the largest variation in Lw, especially for the ig and ex regions. This partly reflects the fact that this category includes two phylogenetically remote groups, protists and fungi. In contrast, the relatively small variation in the vertebrate Lw reflects the fact that, compared to organisms in other categories, vertebrates are phylogenetically very close. Two examples in opposite extremes are shown in the bottom panel of Fig. 4 (k = 7): the malaria causing parasite P. falciparum with especially small Lw’s, and the fungus S. pombe with relatively large Lw’s. This indicates that the chromosomes of P. falciparum and S. pombe are much less and much more random, respectively, than the genomic norm. Although such inter-category, inter-species and inter-regional differences are significant, they pale when compared with the difference between Lw and true chromosome lengths. Table 2 lists Lw(k), k = 2, 5, 7 and 10, averaged over all 865 sequences, for whole chromosome and the four types of concatenates.

Summary of genomic data

We summarize the trends of genomic data: (a) Lw(k) increases with k. (b) For given k, Lw has no systematic dependence on L and has a weak dependence on p. (c) For given k, Lw for different organisms are of the same order of magnitude. (d) Within a genome, Lw differs little among chromosomes. (e) There is remarkable agreement between the gn and ex data sets. (f) There is not a significant difference between the Lw(k)’s for coding (ex and gn) and non-coding (in and ig) regions, and the agreement between the two regions improves when that fact that coding regions tend to be GC-rich is taken into account (Text S1 and Fig.

![Figure 2. Segmental equivalent lengths from four model organisms](https://example.com/figure2.png)

**Figure 2. Segmental equivalent lengths from four model organisms.** Equivalent length Ls versus sequence length Ls for genomic (hollow symbols) and matching random (solid symbols) sequences. Genomic segments are from E. coli (○), worm (C. elegans (chromosome) I, □), mustard (A. thaliana I, △), and human (H. sapiens I, □). Each Ls in the form of mean ± SD is averaged over the maximum number of non-overlapping segments (of length Ls) in the chromosome or, if the chromosome is longer than 2Ls, 20 randomly selected segments.

doi:10.1371/journal.pone.0009844.g002
S1). We remark that in splicing the $gn$ concatenate genes in positive and negative orientations from a single strand of DNA are concatenated, without inverting the negatively oriented genes (Methods). Similarly for the $ex$ concatenate.

**Discussion**

Universal $L_e$ is not a result of inter-chromosome similarity in $k$-mer-content

Fig. 5 shows intra-chromosome $k$-mer-content similarity plots (Methods) for six representative chromosomes. In the plots, a small value of $\eta_{am}$ ($\leq 0.2$, black-blue) indicates high degree of similarity, and a large value ($\geq 1$, cyan to red) indicates the opposite. A general trend is that local $k$-mer-content within a chromosome is fairly homogeneous [51,52] on a scale as small as 50 kb. When $k$-mer-contents of coding and non-coding parts show a significant difference, as is seen in the case of P. falciparum, M. stadtmanae, and E. coli, it is mainly caused by the $gn$ part being substantially richer in GC content than the $ig$ part (Table 3). Nevertheless, because $L_e$ is defined such that first-order dependence in base composition is removed, within a chromosome the $L_e$'s for the $gn$ and $ig$ parts and for the whole chromosome generally have similar values (Table S3, SI).

Fig. 6 compares the intra-E. coli plot with inter-chromosome plots of E. coli versus seven other organisms whose phylogenetic distances to E. coli range from close to remote. The approximate monochromaticity of each plot reconfirms our previous observation that $k$-mer-content within a chromosome has a high degree of homogeneity (on a scale of 100 kb). We see close correlation between phylogenetic distance and the shades (colors) of the seven inter-chromosome plots. Fig. 7 gives the mean $\eta_{am}$ for the plots.
and P-values from Student t-tests for the null assumption that the inter-chromosome plots are the same as the intra- E. coli plot. These results verify that the observed near universal value in $L_c$ is not cause by similarity in $k$-mer-content among chromosomes. As an aside, we note that in Fig. 6 the plot for S. pombe indicates a ~ 100 kb $\tilde{g}$ segment around the 1.1 Mb site has extraordinary low similarity with respect to all other regions of the chromosome. This could be the result of a non-geic horizontal/lateral transfer [53,54] and suggests that similarity plots may be useful for locating such events.

A universal formula for $L_c$

The 7360 pieces of data in the “All” set in Table 2 is well represented by the empirical formula,

$$L_c^{(w)}(k; p) = L_{c2} \exp (\frac{k - 2}{a(p)}); (2 \leq k \leq 10)$$ (1)

$$a(p) = \frac{a_0}{1 + e \tan \left( \frac{p^2 + (1-p)^2 - 0.5}{n} \right)}$$ (2)
where \( a_0 = 0.92, L_{c0} = 310^{+700}_{-200} \text{ kb}, \) and \( \epsilon = 0.50 \pm 0.05. \) The central values of the formula are shown as solid lines in Fig. 3 and listed as the entries in the row labeled \( L_{e}^{\text{avg}} \) in Table 2. The denominator in Eq. (2) represents the residual \( p \)-dependence indicated in the data in Fig. 3; it works well even for chromosomes with large \( p \approx 0.5 \) (Table S4, SI). For the vast majority of genomic \( L_{e} \), \( \chi^2 \equiv \ln^2(L_{e}(k)/L_{e}^{\text{avg}}(k;p)) \) (Text S1) is less than 1 (Fig. S2) and, averaged over the 7360 pieces of data in the “All” set, \( \langle \chi^2 \rangle \approx 0.43. \) This means that on average the genomic \( L_{e} \) is within a factor of two of \( L_{e}^{\text{avg}}(k;p). \) In recognizing that genomes as a category exhibit such a non-trivial common feature which is itself the manifest of an underlying but yet undefined cause, we say genomes belong to a universality class. It is realized that Eq. (1) cannot be extended to \( k \) much greater than 10 (and not even to 10 for some of the smaller chromosomes), because a meaningful value for \( L_{e}(k) \) may be extracted only when a sequence is at least \( 4^{k+1} \) bases long.

### A universal formula for the standard deviation from the fluctuating part in \( k \)-mer frequency

The short genomic \( L_{e} \) (relative to actual chromosome length) is a direct consequence of the genomic \( CV_{p} \) being much larger than its random-sequence counterpart. If we approximate \( a(p) \) in Eq. (1) by \( a_0 \) and approximate the factor \( b_k \) in Eq. (14) (Methods) by unity, then through Eq. (14) we convert Eq. (1) to a universal formula for the \( m \)-set-averaged standard deviation for the \( k \)-mer FFD:

\[
\sigma_{\beta}(k) \approx 0.14^{+0.05}_{-0.04} 10^{-k/2} L,
\]

where \( L \) is the sequence length. The formula is meant to be applicable so long as \( L \) is several times greater than \( 4^{k}. \) For sequences with \( p \approx 0.5, \sigma_{\beta} \) reduces to the usual variance. Note that for random sequences \( \sigma_{\beta}(k) \sim L^{1/2} 4^{-k/2}. \) Since \( L \) is large, genomic \( \sigma_{\beta} \) can be orders of magnitude greater than its random counterpart. For instance, for the 4.6 Mb chromosome, the \( k = 4 \) values for \( \sigma_{\beta} \) given by Eq. (3), the actual chromosome \( (m \)-averaged), and a random sequence are 6440 b, 6230 b, and 134 b, respectively, and for the 228 Mb human chromosome 1, the corresponding values are 319,000 b, 380,000 b, and 943 b, respectively. To give statistical meaning to such differences, Table 4 examines universal genomes of

![Figure 4. 4.4 Equivalent lengths for complete chromosomes and concatenates.](image)

**Table 2. Average genomic equivalent lengths.**

<table>
<thead>
<tr>
<th>Category</th>
<th>( L_{c} ) (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>359.7±33.1</td>
</tr>
<tr>
<td>gn (41.8%)</td>
<td>317.2±25.3</td>
</tr>
<tr>
<td>ex (59.6%)</td>
<td>462.5±37.6</td>
</tr>
<tr>
<td>in (31.8%)</td>
<td>348.6±37.6</td>
</tr>
<tr>
<td>( L_{c}^{\text{avg}}(p=0.5) )</td>
<td>310.7±35.9</td>
</tr>
<tr>
<td>RSD model</td>
<td>597.2±316.4</td>
</tr>
</tbody>
</table>

\( L_{c}(k), k = 2, 5, 7 \text{ and } 10, \) averaged over 865 chromosomes. Total sequence lengths is about 2.2 \times 10^{9} \) bases. Abbreviations: All, complete chromosome; gn, genes; ig, intergenic; ex, exons; in, introns. Percentage given indicates portion of complete sequence. \( L_{c}^{\text{avg}} \) is defined in Eq. (1) and RSD results are averaged over 200 model sequences. See Table S4 for \( L_{c}(k) \) of other \( k \) values.

doi:10.1371/journal.pone.0009844.t002
various lengths and gives the fractions of 2-mers and 9-mers (in the genomes) whose frequencies have P-values that are less than $P_n$ – the P-value corresponding to $n$ standard deviations away from the expected frequency in a random sequence – for $n = 3, 6, \text{ and } 8$, respectively. Because $\frac{\sigma_r(k)}{\sigma_{\text{ran}}(k)} \sim L^{1/2}/(0.4)^{1/2}$, the fraction increases with decreasing $k$ and increasing $L$ (for a given $n$). For instance, for a sequence 4.6 Mb long (length of $E. \text{ coli}$ chromosome), fourteen of the sixteen 2-mers have $P \leq P_8 (= 1.3 \times 10^{-15})$, whereas only 26,000 of the 262,144 9-mers are so. In comparison, for a sequence 226 Mb long (length of human chromosome 1), all sixteen 2-mers and 213,000 of the 9-mers are so.

**Segmental duplication shortens $l_u$**

We now discuss probable causes for the formation of the universality class. We first list some general properties of the ratio $\rho$ of $l_u$ to the sequence length $l$: if the sequence is (nearly) random then $\rho(=l_u/l) \approx 1$; if it is far less random than a random sequence of length $l$ then $\rho \ll 1$; if it is essentially ordered then $\rho \approx 0$; if it is sequence 226 Mb long (length of human chromosome 1), all sixteen 2-mers and 213,000 of the 9-mers are so.

**Table 3. Intra-chromosome similarity indexes.**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Length (Mb)/p</th>
<th>Average $\eta_{\text{max}}$</th>
<th>chr-chr</th>
<th>gn-gn</th>
<th>ig-ig</th>
<th>gn-ig</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. pombe Chr. 1</td>
<td>2.45/0.64</td>
<td>0.648</td>
<td>0.569</td>
<td>0.615</td>
<td>0.647</td>
<td></td>
</tr>
<tr>
<td>E. cuniculi (genome)</td>
<td>2.50/0.53</td>
<td>0.527</td>
<td>0.481</td>
<td>0.450</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>P. falciparum Chr. 13</td>
<td>2.73/0.82</td>
<td>0.801</td>
<td>0.742</td>
<td>0.641</td>
<td>2.11</td>
<td></td>
</tr>
<tr>
<td>M. stadtmanae</td>
<td>1.77/0.73</td>
<td>0.805</td>
<td>0.782</td>
<td>0.757</td>
<td>2.52</td>
<td></td>
</tr>
<tr>
<td>S. glossinidius morsitans</td>
<td>4.17/0.46</td>
<td>0.638</td>
<td>0.510</td>
<td>0.635</td>
<td>0.729</td>
<td></td>
</tr>
<tr>
<td>E. coli K12</td>
<td>4.64/0.50</td>
<td>0.517</td>
<td>0.481</td>
<td>0.548</td>
<td>1.63</td>
<td></td>
</tr>
</tbody>
</table>

Compositions and average regional similarity indexes of sequences shown in Fig. 6; chr, chromosome; gn, gene; ig, intergenic.

doi:10.1371/journal.pone.0009844.t003
the n-fold replication of a random sequence, then \( \rho \approx 1/n \). We illustrate how segmental duplication can cause a sequence to have \( \rho \) much less then one, by considering the effect of a generalization of the operation of replication on \( l_e \). To be specific we label XY a concatenate composed of X and Y. If Y is a coarse-grained rearrangement of X, then, provided the scale of the rearrangements is not too small, \( l_e(X) \approx l_e(Y) \) and concatenating X and Y is similar to doubling X by replication, hence \( l_e(XY) \) will be nearly equal to \( l_e(X) \).

In general, if the k-mer-contents of X and Y are similar, then (provided the sequences are sufficiently long) we expect \( l_e(XY) \approx l_e(X) \approx l_e(Y) \). Conversely, if the k-mer-contents of X and Y are significantly different, then we expect \( l_e(XY) \approx \min(l_e(X), l_e(Y)) \) (see Text S1 for an expanded discussion, including formulas given in Table S5).

Results for testing these simple rules with real sequences are shown in Table 5. We expect agreement with theory to improve with increasing sequence length (\( l_e \)). The first two rows of results in Table 5 verify that for random sequence \( \rho \) is always close to one, or \( l_e \approx l_0 \). The results for AA and BB show that concatenating two equal-length segments from the same chromosome is indeed like doubling a sequence by replication. Chromosomes labeled C have k-mer-contents relatively more similar to A (Figs. 4 and 5), therefore \( l_e(AC) \approx l_e(A) \) as expected. Chromosomes labeled D and B have k-mer-contents more dissimilar to A, therefore \( l_e(X) \approx l_e(X) \approx \pi \). The case of AD, where D is H. sapiens chr. 1, is not an exception to the rule even for \( k = 2 \), because \( l_e(D) < l_e(A) \). In the bottom portion of Table 5 the approximate relation \( l_e \approx n^2 l_0 \) (Table S3; \( l_0 \) is the equivalent length of the genomic portion and \( n \) is the ratio of the

---

**Table 4. P-values for k-mer distribution in universality class.**

<table>
<thead>
<tr>
<th>Fraction of k-mers whose P-value is less than P(<em>k), P(</em>{9k}), or P(_{15k})</th>
<th>k = 2 (( L_e = 310 ) b)</th>
<th>k = 9 (( L_e = 194 ) kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (Mb)</td>
<td>P &lt; P(_k)</td>
<td>P &lt; P(_{9k})</td>
</tr>
<tr>
<td>0.8</td>
<td>0.953</td>
<td>0.906</td>
</tr>
<tr>
<td>4.6</td>
<td>0.980</td>
<td>0.960</td>
</tr>
<tr>
<td>30</td>
<td>0.992</td>
<td>0.985</td>
</tr>
<tr>
<td>226</td>
<td>0.997</td>
<td>0.994</td>
</tr>
</tbody>
</table>

P-values for k-mer distribution given by Eq. (1) (at \( p = 0.5 \)). Null theory assumes genomes are random sequences. The P-values P\(_k\) = 2.7 \times 10^{-1} \( F_{10} \), P\(_{9k}\) = 2.0 \times 10^{-9}, and P\(_{15k}\) = 1.3 \times 10^{-15} correspond to \( z \)-values of three, six, and eight, respectively.

doi:10.1371/journal.pone.0009844.t004
length of the concatenate to the that of the genomic portion) is seen to hold: \( L/RX \approx 4L/X \) (X being A or B), \( L/RAB \approx 2.3L/(AB) \), and \( L/R(R'X) \approx 9L/X \).

**Artificial sequences generated by RSD growth model exhibit universal \( L_x \)**

We show that a very simple growth model, the minimum random segmental duplication (RSD) model \([49]\) (Methods; Text S1), generates chromosome-length sequences that have \( L_x \)'s very close to the universal \( L_x^{(w)} \) given by Eq. (1). In the model, simple segmental duplication (SD) serves to represent the numerous modes of DNA copying processes known to occur in genomes \([9,11,55,56]\), and point mutation represents all small non-duplicating events. We consider random events because it is the simplest assumption and because it generates sequences with a reasonable degree of homogeneity \([51,52]\). (It is known that genomes have long-range correlations that require tandem SDs to generate \([46,57]\). Since tandem duplications do not affect \( L_x \), for simplicity they are not given special treatment in this study.) The three parameters of the model are \( L_0 \) (initial length), \( d \) (average duplicated segment length), and \( r \) (cumulative point mutation per-base density) (Methods). \( L_x \) generated by the model is insensitive to sequence length provided it is longer than 0.5 Mb, allows a generous range in \( d \) and a tighter range in \( r \), and is highly sensitive to \( L_0 \) (Fig. S3, S7). (Because RSD will at least initially cause \( L_x \) to be longer than \( L_0 \) and because \( L_x (k+2) \approx \frac{\text{300}}{b} \), \( L_0 \) must be significantly less than 300 b.) Fig. 8 shows that, at \( L_0 = 64 \), the model admits a basin of good values delimited by \( d = 120 \) to 5000 and \( r = 0.65 \) to 0.80. \( L_x \)'s of model sequences obtained using the “best set” of parameters \( L_0 = 64 \), \( d = 1000 \), and \( r = 0.73 \) are shown in the right panel in Fig. 8, where the lines represent the universality class \( L_x^{(w)} \) (Eq. (1)). The \( \chi^2 \)'s for these \( L_x \)'s is 0.18 and implies that on average, the model \( L_x \), and \( L_x^{(w)} \) agree to within a factor of 1.6. This small \( \chi^2 \) can easily be increased to match that of the genomic data (\( \chi^2 = 0.43 \)) by using model parameters that cover suitable ranges of values centered around the best values.

The range of \( d \) within the basin of good values seems biologically realistic, for it is consistent with the range of the characteristic lengths of genes. The isolated basin near \( d = 30 \), \( r = 0.3 \) allows copious duplication of regulatory sequences, including microRNAs \([58]\), that are much shorter than genes. The considerable size of the main basin implies that it is easily accessible in an evolutionary selective process. On the other hand, that \( \chi^2 \) increases sharply outside the basin of good values demonstrates that even in the context of the RSD model it is very easy to generate sequences that are far outside the universality class.

**Rates of genome growth and duplication**

The parameters of the RSD model are compatible with rates of genome growth and duplication determined using sequence comparison \([37–39]\). In a model where a genome grows at a constant per-time rate \( \lambda \), we have \( \lambda = (t_2 - t_1)^{-1} \log (L_2/L_1) \) where \( L_x \) is the length of the genome at time \( t_i \) (Eq. (16), Methods). For human we can take \( t_2 \) to be the current time because the human genome has grown 15% to 20% in the last 50 Mya (10^9 years) \([39]\). The ancestors of eubacteria and archaea-eukarya diverged ~3.4 Gya (10^9 years) ago \([59–61]\), and before that prokaryotes most likely evolved as communities \([62–64]\), and hence had a different growth regime than later times. The smallest bacterial genome is about 0.2 Mb, we take \( L_0 \) to be from 0.05 to 0.2 Mb and \( L_2 = 3 \) Gb. Then \( \lambda = 2.7 \approx 3.7 \). These rates imply the human genome grew ~14~20% in the last 50 Mya, in agreement with \([39]\). If we assume the growth is purely SD and take the length of duplicated segment \( d \) to be 500 b to 2 kb, then the rate of SD events is \( \mu_{SD} = \lambda_b/d = 1.4 \approx 7.4 \) Mb/Mya. These values are comparable to the estimates of 3.9 Mb/Mya (from animal gene duplication rate of ~0.01 per gene per Mya \([6]\) and human coding region ~3% of genome), and 2.8 Mb/Mya (from human retrotransposition event rate \([39]\)).

**Cumulative mutation density and mutation rates**

The parameter \( r \) in the RSD model, the cumulative point mutation density, is related to the (per-site per-time) rate density \( \mu_p \) of “point mutations” — including small deletion and insertion but excluding SD — by \( \mu_p \approx r/\lambda \) (Eq. (19), Methods). If we take the best value \( r = 0.73 \) from the RSD model then \( \mu_{p,SD} = 0.98 \approx 1.4 \times 10^{-3} \) site/Mya. This agrees well with the value \( \mu_{SD,hs} \approx 1 \times 10^{-3} \) site/Mya \([37–39]\) determined by sequence comparison.

We cannot assume the \( E. coli \) genome is still growing, as the human genome appears to be. Instead, like most bacteria \( E. coli \) probably acquired its full length in antiquity, not too long after ancestors of eubacteria and archaea-eukarya diverged \([61]\). If we assume \( E. coli \) acquired its current length of 4.6 Mb about 0.4 to 0.6 Gya after that, then with \( L_1 \) as before, we have \( \lambda_{ec} = 5.4 \approx 11/ \) Mya, and \( \mu_{p,ec} = 2.0 \approx 4.0 \times 10^{-3} \) site/Mya. Fortuitously or perhaps this range of rates represent an equilibrium value, it is compatible with the sequence-comparison \( E. coli \) rate of


\( \mu_{r,sc} \approx 5 \times 10^{-3} / \text{site/Mya} \) based on mutations that (putatively) occurred in the last 0.5 Gya or less [37,38]. There is evidence that natural selection does cause genomes to have a relatively low and stable mutation rate. For instance, laboratory measured spontaneous mutation rates of E. coli [63], C. elegans [63,66], and Drosophila [65,67] tend to be two or three orders of magnitude higher than the characteristic rates of \( \approx 0.001 / \text{site/Mya} \) of wild types.

Presumably the same selective force is what causes the \( L_s \)'s, hence the cumulative mutation density \( r \), of coding and non-coding regions of a chromosome to be nearly equal. Such a force must be acting for otherwise we expect non-coding regions to have a significantly higher \( r \), which is not the case.

**Materials and Methods**

**Complete genome sequences**

A total of 855 complete chromosomes were downloaded from the genome database [68] on 2006/10/01. The set is composed of 467 prokaryotic chromosomes (435 eubacteria and 32 archaeca) and 398 chromosomes from 28 eukaryotes, including: 12 unicellulars (A. fumigatus (8 chromosomes), C. albicans (1), C. glabrata (13), C. neoformans (14), D. hansenii (7), E. canadensis (11), E. gossypii (7), Klyveromyces lactis (6), S. cerevisiae (16), S. pombe (3), Y. lipolytica (6), P. falicarpur (14)), 5 insects (A. gambiae (3), A. mellifera (16), C. elegans (6), D. melanogaster (4), T. castanea (10)), 2 plants (A. thaliana (5), O. sativa (12), 9 vertebrates (B. taurus (30), C. familiaris (39), D. rerio (25), G. gallus (30), H. sapiens (24), M. mulatta (21), M. musculus (21), P. troglodytes (25), R. norvegicus (21)). The complete list of sequences, accession numbers, lengths and other properties relevant to this study are given in Table S1.

**Partition of \( k \)-mers into \( m \)-sets**

We always speak of single-stranded sequences. We refer to a \( k \)-base nucleic word as a \( k \)-mer and denote the set of all \( \tau \approx 4^k \) types of \( k \)-mers by \( S \). Given a sequence, we count the frequency of occurrence (or frequency) \( f_\tau \) of each \( k \)-mer-type \( \tau \) in \( S \) using an overlapping sliding window of width \( k \) and slide one [36]. Then the sum of the frequencies is \( \sum_{\tau \in S} f_\tau = L - k + 1 \), here approximate by \( L \), and the mean frequency is \( f = L/\tau \). Let the fractional AT- and GC-content of a sequence be \( p \) and \( q = 1 - p \), respectively. We say a sequence has an even-base composition when \( p \) is equal to or very close to 0.5, otherwise it has biased base composition. Owing to Chargaff's second parity rule [69] \( p \) is an accurate and efficient classifier of base composition for statistical analysis. The \( k \)-mers in a sequence are naturally partitioned into \( k+1 \) “\( m \)-sets”, \( S_m \), \( m = 0, 1, \ldots, k \), where each \( k \)-mer in \( S_m \) has \( m \) and only \( m \) AT's; \( \bigcup_m S_m = S \). For example, in the case of \( k = 2 \), \( S_0 \) is the set \{CC, CG, GC, GG\}; \( S_1 \) is the set \{CA, CT, GA, GT, AC, AG, TC, TG\}; and \( S_2 \) is the set \{AAAT, TA, TT\}. The number of types of \( k \)-mers in \( S_m \) is \( \tau_m = 2^k / \binom{k}{m} \), which satisfies the sum-rule \( \sum_m \tau_m = \tau = 4^k \). These relations derive from the binomial expansion (for given \( k \))

\[ \tau = (2 + 2)^k = 2^k (1 + 1)^k = 2^k \sum_{m=0}^{k} \binom{k}{m} = \sum_{m=0}^{k} \tau_m. \] (4)

Let \( L_m = \sum_{\tau \in S_m} f_\tau \) be the sum frequency of the \( k \)-mers in \( S_m \). Then \( \sum_m L_m = L \) and the mean frequency of the \( k \)-mers in \( S_m \) is \( f_m = L_m / \tau_m \). The large-\( L \) limit of \( f_m \) for a random sequence, \( f_m^{(ran)} \), is obtained from the binomial expansion

\[ L = \tau f = f 4^k (p + q)^k = \sum_{m=0}^{k} \binom{k}{m} (2^k p^m q^{k-m} f) = \sum_{m=0}^{k} \tau_m f_m^{(ran)}. \] (5)

That is,

\[ f_m^{(ran)} \equiv \lim_{L \to \infty} f_m^{(ran)} = 2^k p^m q^{k-m} f. \] (6)

Depending on \( p \), \( f_m^{(ran)} \) can vary wildly, all collapsing to \( f \) when \( p = 0.5 \). Eq. (6) not only provides an highly accurate estimate of the value of \( f_m \) for genome-size random sequences, it also gives a reasonable estimate for genomic \( f_m \) (Table 6).

**Fluctuation in occurrence frequency**

The coefficient of variation of the occurrence distribution is \( CV = \sigma / \mu \), where \( \sigma \) is the standard deviation. For random events of equal probability, here translated to \( k \)-mer frequencies of a (long) random sequence with even-base composition, the distribution is Poisson and \( \sigma^2 = \mu \), hence \( CV^2 = f^{-1} \tau / L \), which tends to zero in the large-\( L \) limit. This no longer holds when the random sequence has a biased base composition. As controls we consider random sequences that match genomes, namely those whose lengths and base compositions are the same as their genomic counterparts. In particular, such sequences obey Chargaff’s second parity rule [69] in that their A and T, and C and G, separately...
have nearly equal probabilities. For any sequence whose \(k\)-mers are partitioned into \(m\)-sets, using a generalization of the parallel axis theorem, we write as follows:

\[
\begin{align*}
\sigma^2 &= \lim_{t \to 0} \sum_{m=0}^{k} \frac{(f_m - \bar{f})^2}{m} \\
&= \lim_{t \to 0} \sum_{m=0}^{k} \frac{\tau_m (f_m - \bar{f})^2 + 2(\bar{f} - f_m)^2}{m} + \sum_{m=0}^{k} \frac{(f_m - \bar{f})^2}{m}.
\end{align*}
\]

The second term vanishes upon summing over \(u \in S_m\), so \(\sigma^2\) is composed of two parts,

\[
\sigma^2 = \sigma_{nf}^2 + \sigma_{fl}^2,
\]

a non-fluctuating part determined by average frequencies \(\bar{f}\) and \(\bar{f}_m\),

\[
\sigma_{nf}^2 = \sum_{m=0}^{k} \frac{\tau_m (f_m - \bar{f})^2}{m},
\]

and a fluctuating part determined by the fluctuation of \(f_u\) (in an \(m\)-set) around an average frequency,

\[
\sigma_{fl}^2 = \sum_{m=0}^{k} \frac{\sum_{u \in S_m}(f_u - f_m)^2}{\tau} = \sum_{m=0}^{k} \frac{\tau_m \sigma_{fl}^2}{m}.
\]

Thus,

\[
CV^2 = (\bar{f}/\bar{f})^2 = (\sigma_{nf}/\bar{f})^2 + (\sigma_{fl}/\bar{f})^2 = CV_{nf}^2 + CV_{fl}^2.
\]

The non-fluctuating, or “non-statistical”, part, \(CV_{nf}^2\), has a well-defined value in the large-\(L\) limit, obtained by replacing \(f_m\) by \(f_m^{(x)}\) in Eq. (9):

\[
\left(\frac{CV^{(x)}}{L}\right)^2 = \lim_{L \to \infty} CV_{nf}^2 = \sum_{m=0}^{k} \left(2^{-k} \frac{m}{k} \left(2^k p^m q^{k-m} - 1\right)^2\right)
\]

\[
= 2^k (p^2 + q^2)^k - 1,
\]

which has a strong dependence on \(p\) and vanishes \(p = 0.5\). Because genomes are large, \(CV^{(x)}\) gives an accurate description of \(CV_{nf}^2\) for genome-size random sequences; it also happens to do almost as well for genome (Fig. 1). Owing to the existence of this term, the \(CV\) for a genomic sequence may be much greater than that of its matching random sequence (when \(p \approx 0.5\); see, e.g., Fig. 9 (A)); or quite similar (when \(p\) differs significantly from 0.5; see, e.g., Fig. 9 (B)). Because \(CV_{nf}^2\) hardly depends on the distribution of the \(k\)-mers, it should be considered a background in \(CV^2\) in relation to the signal which is \(CV_{fl}^2\).

For a random sequence, the frequency distribution in the subset \(S_m\) is nearly Poisson, hence \(\sigma_{nf}^2 \to f_m\) in the large-\(L\) limit. Therefore, from Eq. (10),

\[
\lim_{L \to \infty} CV_{nf}^2 = \frac{1}{f_m^2} \lim_{L \to \infty} \frac{\sigma_{fl}^2}{f_m^2} = \frac{1}{\bar{f}} \sum_{m=0}^{k} \frac{\tau_m f_m}{m} = \frac{1}{\bar{f}} \bar{f} = 1
\]

\[
= \bar{f} (\text{random sequence}),
\]

which is exactly the limit expected of \(CV^2\) for an even-base \((p = 0.5)\) random sequence. In other words, for random sequences \(CV_{fl}^2\), but not \(CV\), has the correct large-\(L\) limit expected of a random system. The right-hand side does not depend on \(p\), which is a reflection of the fact that for genome as well as random sequences, \(CV_{fl}\) has at most a weak \(p\)-dependence; the main \(p\)-dependence having been removed when \(CV_{nf}^2\) is subtracted from \(CV^2\). Because (for random sequences) \(CV_{fl}\) decreases with increasing \(L\) but \(CV_{nf}\) does not, there is a crossover value of \(L\) beyond which \(CV_{fl}^2\) becomes the leading term in \(CV^2\) (when \(p < 0.5\)). When \(p = 0.7\), this crossover value is 42, 316 and 2851 (bases) for \(k = 2, 4, \text{and} 6\), respectively, which are orders of magnitudes shorter than even the smallest chromosomes. To summarize, if one wants to compare the statistical properties in the frequency distributions of \(k\)-mers in the genomic and random sequence, one must use \(CV_{fl}\), not \(CV\).

Two examples: \(E. coli\) and \(C. acetobutylicum\)

We explain the formulation presented in the last two sections by presenting results of distributions, or spectra, of frequency of \(5\)-mers (as an example), and values of quantities such as \(f_m\), \(\sigma_{nf}^2\), and \(CV^2\) for two genomes with very different base compositions: \(E. coli\) \((p = 0.492)\) and \(C. acetobutylicum\) \((p = 0.691)\). Here, a spectrum is the number of \(k\)-mers plotted against occurrence frequency. The spectra for the two genomes are shown as black curves in panels (A) and (B) of Fig. 9. The solid green curves characterized by narrow peaks are the spectra for random sequences obtained by scrambling the genomes. (The red curves are for sequences generated in the RSD model, see text.) In (A) the mean frequency of both spectra is \(f = 2 \times 10^6 / 4^5 = 1953\). However, the genomic spectrum is seen to be much broader than the random-sequence spectrum, indicating that whereas in the random sequence frequencies \(f_m\) of individual \(5\)-mers deviate little from the mean \(f\), in the genomic sequence that is not the case; frequencies of individual \(5\)-mers fluctuate widely around the mean. Drastically different from (A), the overall widths of genome and random-sequence spectra in (B) are similar. Instead of having a single peak, the random-sequence spectrum is composed of six widely spread narrow subspectra whose peaks are near the theoretical mean frequencies (for \(p = 0.7\)) of the \(m\)-sets, \(f_m^{(x)} \approx 152, 354, 827, 1930, 4500, \text{and} 10500\), for \(m = 0\) to 5, respectively. Eq. (6) shows that these mean values are determined by \(m\) and the base composition of the sequence, or \(p\), and does not depend on the
fluctuation of frequencies of \(m\)-specific 5-mers. (B) and (C) in Fig. 9 show that in the random sequence frequency fluctuation within an \(m\)-set is again small. In contrast, and just as in (A), frequency fluctuations of \(m\) specific 5-mers in the genomic sequence are large (Fig. 9 (C) and Fig. 10 [70]).

Table 6 shows that \(\phi_m\) gives a very accurate estimate of \(f_m\) for random sequences and a fair one for genomic sequences. In the \(p = 0.492\) case, the relation \(f_m \approx \phi_m\) for all the \(m\)'s explains the narrowness of the random spectrum in Fig. 9 (A); like its counterpart in (B), it is also composed of six subspectra, but unlike (B) whose subspectra are spread widely, now the subspectra are superimposed. Table 7 highlight important aspects of our formulation: (i) \(CV_{nf}\) has a strong dependence on \(p\) but not on whether a sequence is genomic or random; (ii) \(CV_{p(x)}\) gives an excellent estimate of \(CV_{nf}\) for random sequences, and a fair estimate for genomes; (iii) \(CV_p\) depends weakly on \(p\) but strongly on whether a sequence is genomic (relative large value) or random (several orders of magnitude smaller, and much smaller than \(CV_{nf}\) except when \(p \approx 0.5\)); (iv) For random sequences Eq. (13) is a fairly accurate relation.

**Equivalent length**

The \(k\)-mers equivalent length of a sequence is defined as

\[
l_e = b_k \tau / CV_p^2 \quad \text{for} \quad k \geq 2
\]

where \(CV_p^2\) is given by the frequency distribution of \(k\)-mers. Recalling that for a random sequence \(CV_p^2\) is inversely proportional to average sequence length (Eq. (15)), we see that \(l_e\) is the length of a random sequence whose \(CV_p^2\) has the same value as that of the genome. The empirical factor \(b_k = 1 - 2^{-k+1}\), instead of the theoretical binomial factor \(1 - \tau^{-1}\), is used to ensure that for a random sequence, regardless of base composition, \(l_e\) approximates the true sequence length with a high degree of accuracy. With the signal term \(CV_p^2\) included but the strongly \(p\)-dependence background term \(CV_{nf}\) excluded in its definition, \(l_e\) is expected to have at most a weak \(p\)-dependence. That is, \(l_e\) is a quantity with which we can compare on the same footing genomes with widely disparate base compositions.

**Genic, non-genic, exon, and intron concatenates**

These various concatenates are formed by splicing corresponding sections from a single strand of the DNA sequence and then stitching the sections together in the order and orientation they appear in the sequence. In particular, the genic and exon concatenates include genetic codes in positive and negative orientations.

**Similarity index and similarity matrix**

Given a pair of equal-length sequences \(\alpha\) and \(\beta\), the similarity index \(\eta_{\alpha\beta}(\alpha, \beta)\) for the pair is defined as

\[
\eta_{\alpha\beta}^2(\alpha, \beta) = \frac{1}{k+1} \sum_m \frac{1}{2s_m} \sum_{m \in S_m} \frac{(f_m(\alpha) - f_m(\beta))^2}{\sigma_m^2(\alpha) \sigma_m(\beta)}
\]

where \(S_m\) is an \(m\)-set and \(\sigma_m^2\) is the variance of the frequency of the \(k\)-mers in \(S_m\). The pair are similar (in \(m\)-content) when \(\eta_{\alpha\beta} \ll 1\), are (considered to be) identical when \(\eta_{\alpha\beta} = 0\), and are highly dissimilar when \(\eta_{\alpha\beta} \gg 1\). If we divide \(\alpha\) and \(\beta\) into (possibly overlapping) segments \(\{a_1, a_2, \ldots\}\) and \(\{b_1, b_2, \ldots\}\), respectively, then we call the matrix whose element \((i, j)\) is valued \(\eta_{\alpha\beta}(a_i, b_j)\) a similarity matrix. In Fig. 6, similarity matrices are displayed as similarity plots by color coding elements of similarity matrices.

**Minimum RSD model for genome growth**

We denote by \(L\) the designated length of a sequence and \(p\) the designated AT-fraction of the sequence. We call the pair \((L, p)\) the profile of a sequence; in our model, the two profiles \((L, p)\) and \((L, 1-p)\) are mathematically equivalent. By a growth model we mean a computer algorithm for generating, from an initial sequence, a target sequence that has a given profile and other specific genome-like attributes. Ours is a model of random
segmental duplication (RSD) [49] in which the three main steps are: (i) randomly select a site from the sequence, (ii) from that site cull a segment of random length (but from a given length distribution) for duplication; (iii) reinsert the duplicated segment into the sequence at a (second) randomly selected site. The model has three explicit parameters: $L_0$, the initial sequence length; $d$, the average length of duplicated segments; $r$, the cumulative point mutation density (replacement only), or number of mutations per site. The generation of a model sequence involves three steps: selection of initial sequence, growth by RSD, point mutations. An initial sequence (of length $L_0$) is chosen such that it has a target value $p$ but is otherwise random. The lengths $l$ of the duplicated segments are selected with uniform probability within the range 1 to $2d$, unless the current length of the genome $L$ is less than $2d$, in which case $l$ is selected from within the range 1 to $L$. Growth is stopped when the length of the sequence exceeds the target length for the first time. Point mutations have a base bias defined by $p$ and are administered after the growth is complete. That is, the administration of point mutations on the sequence is not meant to emulate point mutations suffered by a genome during its growth. Rather, $r$ is meant to indicate the average cumulative number of point mutations per site experience by the genome throughout its life. Because RSD causes drifts in base composition, the profile of the generated sequence will have a profile that is a close approximation of, but not exactly equal to, the target profile.

**Mutation rates**

We derive formulas for computing the rate density, or per site rate, of duplication events, $\mu_{SD}$, and the rate density of “point mutation” – including small deletion and insertion but excluding SD – events, $\mu_p$. If the genome grows from time $t_1$ to time $t_2$ at a rate proportional to its length $l$, that is, $\Delta l = \lambda \Delta t$ where $\lambda$ is the event rate (number of events per unit of time), then

$$\lambda = (t_2 - t_1)^{-1} \ln (l_2 / l_1),$$

If the grow is purely by SD and the average length of the duplicated segment is $d$, then

$$\mu_{SD} = \lambda / d.$$  

If $n_p$ is the cumulative number of point mutations, then $\Delta n_p = \mu_p \Delta t$. In SD dominated growth, the effect of point mutation on the overall length of a genome is negligible, so integrating the relation yields

$$n_p(l_2) - n_p(l_1) = \frac{\mu_p (l_2 - l_1)}{\lambda},$$  

(18)

For any $l$ such that $l \gg l_1$, $n_p = \mu_p l / \lambda$. The cumulative mutation sites is greater than $n_p$ because mutation sites are copied during SD. The number of copied mutation sites satisfy $\Delta n_p = n_p \Delta l / l \approx \mu_p \Delta l / l$ (for large $l$). Therefore, $n_p \approx n_p$, that is, the cumulative number of mutated sites is twice $n_p$. At full genome length $L$, this number is $rL$, hence

$$\mu_p \approx r \lambda / 2.$$  

(19)

**Supporting Information**

**Figure S1** Category $L_c$ for coding and non-coding parts. Averages of $p$ (fractional A/T-content) and $L_c$ for $k = 7$ (situations for other $k$s are similar) for the coding parts (solid symbols; $\alpha$ for eukaryotes and $gn$ for prokaryotes) and non-coding parts (hollow symbols; $in$ for eukaryotes and $ig$ for prokaryotes) of chromosomes. Symbols for categories are: vertebrates, red (square); unicellulars, blue (triangle-up); insects, orange (triangle-down); plants, green; prokaryotes, gray (bullet/circle). Numerical indicates number of chromosomes in each category. The curve represents $L_c$ for the universality class: $L_{uc}^{(\alpha)}(k; p)$.

Found at: doi:10.1371/journal.pone.0009844.s001 (0.26 MB TIF)

**Figure S2** Distributions of $\chi^2$ versus $L$ and $p$. Each symbol gives the $\chi^2$ for one chromosomal $L_c$. Top panels, for genomic ($gn$) and exon ($ex$) concatenates. Bottom panels, for intergenic ($ig$) and intron ($in$) concatenates. Symbols, with color, number of data in group, and number of data whose $\chi^2$ is less than $10^{-3}$ given in brackets, stand for: diamond, $gn$ (blue; 7100; 229); square, $\alpha$ (red; 2844, 95); triangle-down, $ig$ (green; 6377, 270); triangle-up, $in$ (orange; 2960, 104).

Found at: doi:10.1371/journal.pone.0009844.s002 (0.77 MB TIF)

**Figure S3** Results from minimal RSD model. Top-left: Equi-$\chi^2$ contour as function of $r$ and $d$, with $L_0 = 64$ (bases); length ($L$) of generated model sequence is 2 Mb and only $L(k)$ results for $k = 7$ are used. Top-right: $L(k)$, $k = 2, 4, 6, 8, 10$ from 200 model sequences generated using the “best” parameters $L_0 = 64$, $<d^2> = 1000$ (b) and $r = 0.73$ (cumulative point mutations per base). The lines are $L_{uc}^{(\alpha)}(k; p)$ that represent the universality class.
given in the main text. The \( p^2 \) for the model sequences is 0.18. Bottom-left: \( p^2 \) versus \( L_0 \) (otherwise best parameters); model sequences have \( L_0 = 2 \) Mb and \( p = 0.5 \). Bottom-right: \( L_0 \) versus \( L_e \) for a \( p = 0.5 \) model sequence generated using the best parameters.

**Table S1** List of complete sequences included in the study (20 pp).

| Found at: | doi:10.1371/journal.pone.0009844.s004 (0.13 MB PDF) |

**Table S2** Equivalent lengths of complete sequences (100 pp).

| Found at: | doi:10.1371/journal.pone.0009844.s005 (0.36 MB PDF) |

**Table S3** \( L_0(k) \), \( k = 2 \) to 10, averaged over categories of organisms.

| Found at: | doi:10.1371/journal.pone.0009844.s006 (0.06 MB PDF) |

**References**