OPEN ACCESS Freely available online

PLoS one

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

Hong-Da Chen[1,2], Wen-Lang Fan[2,3], Sing-Guan Kong[1,2], Hoong-Chien Lee[1,2,4,5]*

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 Sinaca, 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_es of 865 complete chromosomes and found that they have nearly universal but (k-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 $(1\sim5)\times10^{-3}$/site/Mya previously determined by sequence comparison for the human and *E. coli* genomes.

Citation: Chen H-D, Fan W-L, Kong S-G, Lee H-C (2010) Universal Global Imprints of Genome Growth and Evolution – Equivalent Length and Cumulative Mutation Density. PLoS ONE 5(4): e9844. doi:10.1371/journal.pone.0009844

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

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

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: hclee@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-fluctuaing 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_e's; one for each k) of a sequence and discovered a universality in these quantities. We then identify these L_e's and their small values, as a clear and distinct global imprints of genome growth and evolution. (The L_e of a sequence is inversely proportional to the FFD part of the variance and is defined such that the L_e of a random sequence is its own true length. Therefore, a sequence whose equivalent length is L_e has the characteristic randomness of a random sequence of length L_e.) We computed the L_e 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_e's is well represented by the universal formula $L_e^{\{uc\}}(k) = L_{e2}e^{a_0(k-2)}$ where $L_{e2} \sim 310^{+290}_{-150}$ b (base pair) and $a_0 = 0.92$. The formula means that, for the smaller k's, the universal genomic L_e is only a small fraction of the genome length even for the shortest genomes. Another unexpected result is the small difference between the L_e'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 $|p-0.5| \lesssim 0.1$, 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, ${CV_{nf}}^2$ and ${CV_{fl}}^2$, 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 ${CV_{fl}}^2$ for genomes and random sequences are drastically different ((blue) bullets, Fig. 1(c,d)). The genomic ${CV_{fl}}^2$ span a narrow band ranging from 0.01 to 0.1, while the random ${CV_{fl}}^2$ are several orders of magnitude smaller. In fact for random sequences the value of ${CV_{fl}}^2$ 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.

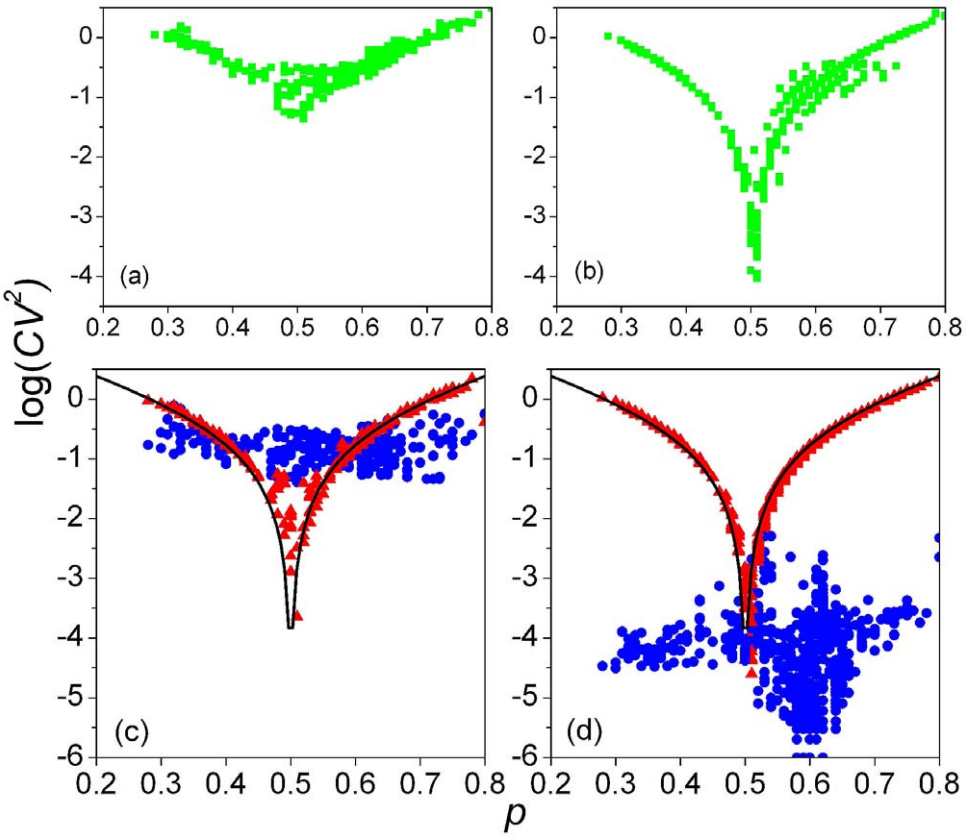

Figure 1. Fluctuating and non-fluctuating parts of variance. (a) Variances of 2-mer frequency distribution of 865 complete sequences. (b) Same as (a) but for for 865 matching random sequences. Bottom: same data as in top plots, but with each variance split into non-fluctuating (triangles) and fluctuating (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 l_e is approximately a constant of sequence length

Throughout this paper we use l_e to denote generically the equivalent length of any sequence (Eq. (14), Methods), and reserve L_e for denoting entire sequences such as a complete chromosomes. Fig. 2 shows l_e versus segment length l_s 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 l_s is at least four times 4^k, since for shorter lengths the systematic error becomes too large. It is seen that whereas the l_e of random sequences closely tracks l_s, as expected, the l_e of genomic sequences quickly levels off to a saturation value $L_e(k)$. These results for $l_s \gtrsim 5$ kb may be summarized in terms of the scaling relation $l_e \propto (l_s)^\gamma$. Then we have the two distinct classes $\gamma \approx 1$ for random sequences and $\gamma \approx 0$ 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 L_e is seen not to depend strongly on organism. For small k, $L_e(k)$ is diminutive relative to genome length: ~ 0.35 and ~ 1.0 kb when $k = 2$ and 4, respectively, growing to $\lesssim 600$ kb when $k = 10$. Within a genome, the apparent invariance of CV (not CV_{fl}) with respect to segment length was noted in [47–49] and the relation between Shannon information and a quantity similar to CV_{fl} was discussed in [50].

Whole chromosomes have nearly universal $L_e(k)$

A list of the 865 complete chromosomes studied here is given in Table S1, and a list of $L_e(k)$'s, $k = 2$ to 10, for the chromosomes is give in Table S2. Fig. 3 shows $L_e(k)$, 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 $L_e(k)$, $k = 2$ to 10, of chromosomes of seven model organisms. It is seen that $L_e(k)$ has a clear dependence on k, is essentially independent of sequence length, and has a weak dependence on p. Fig. 4 gives $L_e(k)$ 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 L_e of a systematic dependence on chromosome length (similarly for other k's). In the $k = 3$ and 7 plots L_e's are given separately for the whole chromosome, and genic (*gn*), and intergenic (*ig*), exon (*ex*) and intron (*in*, when applicable) concatenates (Methods). The unicellulars are seen to have the largest variation in L_e, especially for the *ig* and *in* 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 L_e 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 L_e's, and the fungus *S. pombe* with relatively large L_e'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 L_e and true chromosome lengths. Table 2 lists $L_e(k)$, $k = 2$, 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) $L_e(k)$ increases with k. (b) For given k, L_e has no systematic dependence on L and has a weak dependence on p. (c) For given k, L_e for different organisms are of the same order of magnitude. (d) Within a genome, L_e 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 $L_e(k)$'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.

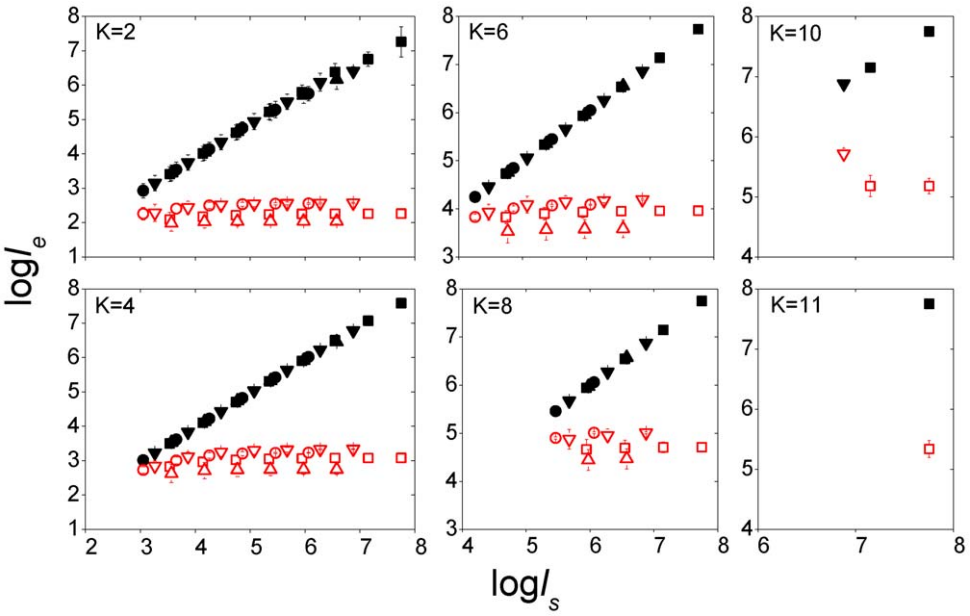

Figure 2. Segmental equivalent lengths from four model organisms. Equivalent length l_e versus sequence length l_s for genomic (hollow symbols) and matching random (solid symbols) sequences. Genomic segments are from *E. coli* (○), worm (*C. elegans* (*chromosome*) *I*, Δ), mustard (*A. thaliana I*, ▽), and human (*H. sapiens I*, □). Each l_e in the form of mean ± SD is averaged over the maximum number of non-overlapping segments (of length l_s) in the chromosome or, if the chromosome is longer than $20l_s$, 20 randomly selected segments.
doi:10.1371/journal.pone.0009844.g002

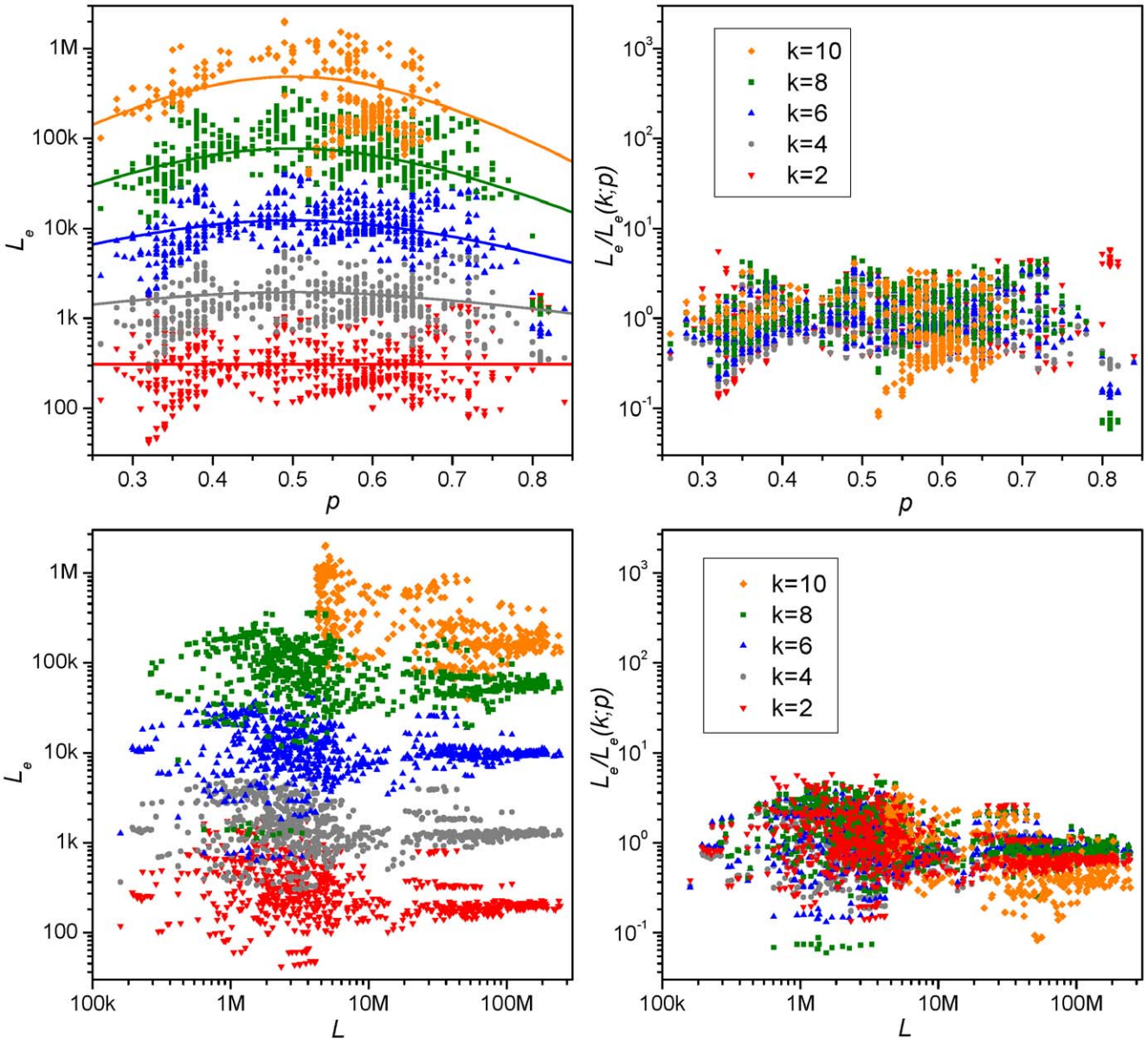

Figure 3. Chromosomal equivalent length (L_e) versus p and L. Top panels: L_e versus p; bottom panels: L_e versus L. Each piece of data gives the L_e from a complete chromosome: ∇ (red), $k=2$; ○ (gray), $k=4$; Δ (blue), $k=6$, □ (green), $k=8$, ◇ (orange), $k=10$. Lines in top-left panel represent the "universality class" $L_e^{\{uc\}}(k;p)$ (Eq. (1)). The right panels show the collapse of genomic data to around unity when the genomic $L_e(k)$ is divided by $L_e^{\{uc\}}(k;p)$.
doi:10.1371/journal.pone.0009844.g003

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-chromome 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 η_{sim} ($\lesssim 0.2$, black-blue) indicates high degree of similarity, and a large value ($\gtrsim 1$, 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 phyogenetic distance and the shades (colors) of the seven inter-chromosome plots. Fig. 7 gives the mean η_{sim} for the plots

Table 1. Genomic equivalent lengths for model organisms.

	L_e (kb)[d]								
Organism \k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
H. sapiens (24)[a]	.188±.021	.448±.046	1.22±.13	3.39±.41	9.34±1.36	23.8±4.4	53.9±12.6	103±29	170±54
H. sapiens (*gn*; 43.2%)[b]	.185±.022	.440±.048	1.20±.14	3.31±.42	9.02±1.33	22.4±4.0	49.2±10.9	90.5±23.7	144±42
H. sapiens (*ig*; 63.6%)[b]	.190±.021	.452±.045	1.24±.13	3.44±.41	9.51±1.36	24.5±4.5	56.6±13.4	111±32	186±61
H. sapiens (*ex*; 2.1%)[b,c]	.171±.019	.412±.042	1.12±.12	3.07±.39	8.21±1.26	19.9±3.8	41.9±10.3	72.2±21.6	117±22
H. sapiens (*in*; 37%)[b,c]	.182±.020	.434±.043	1.18±.13	3.26±.40	8.84±1.34	21.9±4.2	47.7±11.5	87.2±24.9	139±45
A. thaliana (5)[a]	.373±.005	.871±.013	2.20±.04	5.89±.10	16.0±.3	42.1±.8	109±2	273±7	642±20
A. thaliana (*gn*; 55.8%)[b]	.333±.004	.822±.011	2.06±.03	5.57±.08	15.9±.2	44.9±.7	129±2	367±6	981±22
A. thaliana (*ig*; 44.1%)[b]	.394±.007	.798±.014	1.94±.04	4.95±.10	12.3±.2	28.9±.6	66.1±1.5	144±4	296±12
A. thaliana (*ex*; 32.9%)[b,c]	.288±.003	.715±.007	1.75±.02	4.72±.05	13.6±.1	38.9±.4	113±2	326±7	865±35
A. thaliana (*in*; 16.1%)[b,c]	.350±.003	.752±.006	1.80±.02	4.42±.04	11.1±.1	27.3±.4	68.1±1.0	167±3	400±1
Drosophila (4)[a]	.409±.142	.957±.213	2.54±.46	6.90±1.17	18.7±3.2	48.2±9.5	117±31	268±102	676±294
Drosophila (*gn*; 56.4%)[b]	.432±.108	1.02±.15	2.71±.30	7.35±.85	20.0±2.8	51.6±9.9	127±35	326±120	756±321
Drosophila (*ig*; 43.5%)[b]	.392±.194	.882±.305	2.30±.66	6.15±1.57	16.1±3.3	39.4±7.5	90.0±28.1	235±87	536±231
Drosophila (*ex*; 23.9%)[b,c]	.478±.023	1.16±.09	2.82±.41	7.55±1.39	21.0±4.2	55.6±10.7	140±29	377±111	907±324
Drosophila (*in*; 34.8%)[b,c]	.378±.145	.833±.168	2.15±.30	5.65±.73	14.8±2.3	36.2±7.9	84.0±26.2	207±79	458±198
C. elegans (6)[a]	.119±.012	.258±.032	.624±.089	1.63±.26	4.46±.78	12.6±2.3	35.5±6.9	98.8±21.0	264±63
C. elegans (*gn*; 58.6%)[b]	.126±.017	.284±.047	.697±.135	1.83±.40	5.06±1.21	14.3±3.7	40.8±11.1	114±34	306±99
C. elegans (*ig*; 41.3%)[b]	.109±.009	.226±.022	.539±.061	1.39±.18	3.78±.51	10.5±1.5	29.3±4.5	79.5±13.6	202±41
C. elegans (*ex*; 27.5%)[b,c]	.184±.010	.483±.025	1.28±.07	3.64±.23	10.9±.7	33.2±2.4	102±8	306±25	822±58
C. elegans (*in*; 32.3%)[b,c]	.085±.015	.169±.037	.382±.096	.939±.265	2.44±.73	6.52±1.99	17.4±5.3	45.4±14.1	113±37
S. pombe (3)[a]	.362±.010	.894±.030	2.41±.09	6.74±.28	19.2±.9	54.6±3.0	153±11	402±39	1013±39
S. pombe (*gn*; 57.8%)[b]	.339±.002	.880±.006	2.38±.01	6.82±.05	20.2±.2	59.6±.8	173±6	455±42	—
S. pombe (*ig*; 42.1%)[b]	.364±.019	.812±.045	2.08±.12	5.31±.32	13.5±.8	33.6±2.1	81.7±5.8	187±16	—
S. pombe (*ex*; 53.9%)[b,c]	.357±.007	.889±.018	2.40±.06	6.73±.18	19.2±.6	54.4±2.3	149±10	374±42	—
S. pombe (*in*; 3%)[b,c]	.361±.007	.898±.017	2.41±.06	6.53±.14	17.0±.4	38.2±3.1	—	—	—
Plasmodium (14)[a]	1.40±.20	.287±.019	.376±.023	.512±.036	.729±.059	.998±.089	1.34±.13	1.73±.19	—
Plasmodium (*gn*; 56%)[b]	.595±.118	.659±.085	1.02±.12	1.86±.29	3.59±.74	6.73±1.86	12.3±4.3	16.3±10.4	—
Plasmodium (*ig*; 44%)[b]	.665±.108	.111±.017	.130±.017	.162±.022	.212±.031	.276±.042	.357±.057	.398±.032	—
Plasmodium (*ex*; 53%)[b,c]	.515±.058	.717±.060	1.12±.07	2.10±.11	4.21±.23	8.30±.56	16.0±1.3	32.0±1.6	—
Plasmodium (*in*; 5.7%)[b,c]	.163±.019	.052±.002	.064±.003	.076±.003	.095±.004	.116±.003	—	—	—
E. coli (1)[a]	.373	.729	1.74	4.52	12.6	37.0	111	328	879
E. coli (*gn*; 88.7%)[b]	.346	.656	1.56	4.05	11.3	33.0	98.9	292	—
E. coli (*ig*; 11.2%)[b]	.553	1.22	2.60	6.33	16.0	39.3	83.9	—	—

$L_e(k)$, $k = 2$ to 10, of chromosomes of model organisms. The L_e's given are mean ± SD averaged over chromosomes of the organism, except for the single chromosome *E. coli*. See Table S2 for list of all computed $L_e(k)$'s. (*a*) Number in parentheses indicates total number of complete chromosomes in organism. (*b*) Abbreviations: *gn*, gene; *gn*, intergenic; *ex*, exon; *in*, intron. Percentage given indicates portion of complete sequence. "N-runs" or gaps in sequences are not counted. (*c*) *Ex* and *in* segments selected as given by Genbank; sum of percentages for *ex* and *in* may be less than or exceed that of *gn* due to incomplete or duplicated segments. (*d*) $L_e(k)$ computed only if category has more than one sequence whose length exceeds 4^{k+1}.

doi:10.1371/journal.pone.0009844.t001

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_e 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 *ig* 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-genic horizontal/lateral transfer [53,54] and suggests that similarity plots may be useful for locating such events.

A universal formula for L_e

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

$$L_e^{\{uc\}}(k;p) = L_{e2}\exp\left((k-2)a(p)\right);\ (2 \le k \le 10) \tag{1}$$

$$a(p) = \frac{a_0}{1+\epsilon\tan\left[\left(p^2+(1-p)^2-0.5\right)\pi\right]} \tag{2}$$

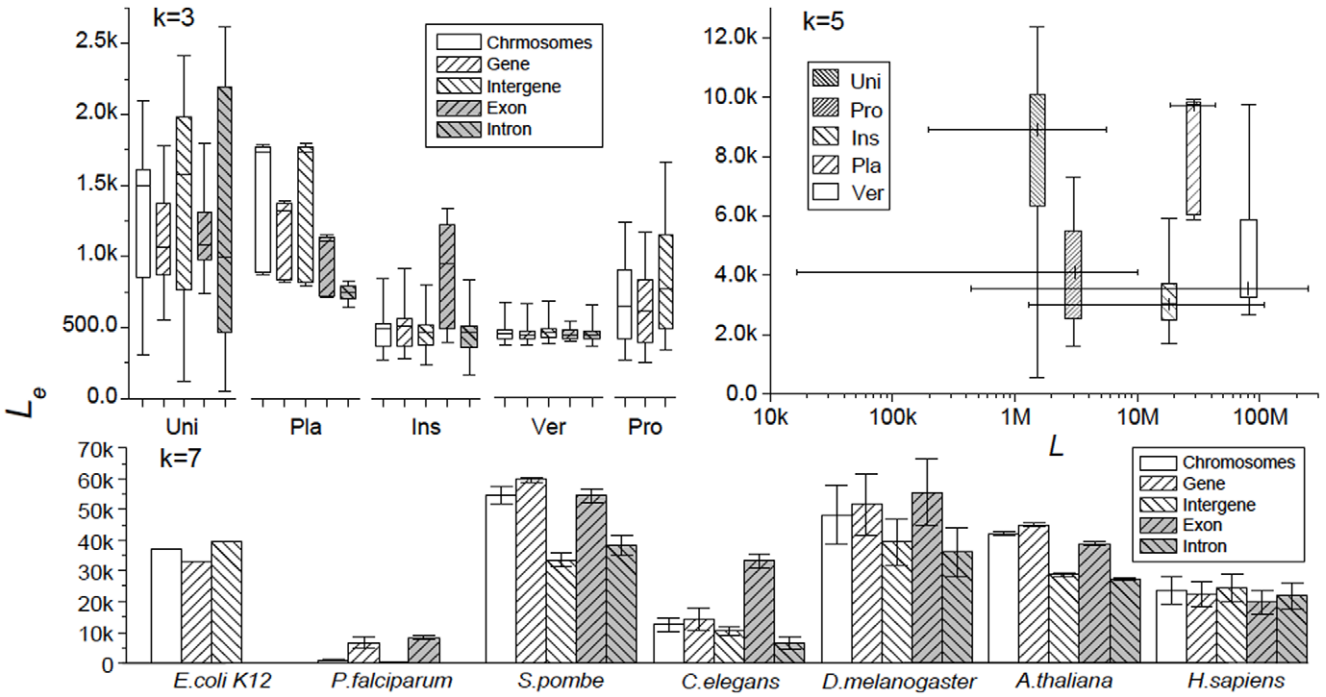

Figure 4. Averaged equivalent lengths for complete chromosomes and concatenates. The concatenates are: "gene" (*gn* in main text), coding regions; "intergene" (*ig*), non-coding or intergenic regions; "exon" (*ex*), exons in *gn* (for eukaryotes); "intron" (*in*), introns in *gn*. Top left, L_e ($k=3$) averaged over phylogenetic categories (Uni, unicellulars; Pla, plants; Ins, insects; Ver, vertebrayes; Pro, prokaryotes); top right, L_e ($k=5$) versus chromosome length average over categories; bottom, L_e ($k=7$) for seven model organisms averaged over chromosomes. Boxes indicate data in the 10, 25, 50, 75 and 90% range.
doi:10.1371/journal.pone.0009844.g004

where $a_0=0.92$, $L_{e2}=310^{+290}_{-150}$ b, and $\epsilon=0.50\pm0.05$. The central values of the formula are shown as solid lines in Fig. 3 and listed as the entries in the row labeled $L_e^{\{uc\}}$ 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 $|p-0.5|$ (Table S4, *SI*). For the vast majority of genomic L_e's, $\chi^2\equiv\ln^2(L_e(k)/L_e^{\{uc\}}(k;p))$ (Text S1) is less than 1 (Fig. S2) and, averaged over the 7360 pieces of data in the "All" set, $\langle\chi^2\rangle=0.43$. This means that on average the genomic L_e is within a factor of two of $L_e^{\{uc\}}(k;p)$. In recognizing that genomes as a category exhibit such a non-trivial common feature which is itself the manifest of an underlying but yet undetermined 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.

Table 2. Average genomic equivalent lengths.

	L_e (kb)			
Category	**($k=$) 2**	**5**	**7**	**10**
All	$.359^{+.333}_{-.172}$	$4.56^{+3.60}_{-2.01}$	$33.7^{+30.0}_{-15.9}$	388^{+524}_{-223}
gn (41.8%)	$.317^{+.253}_{-.141}$	$4.21^{+2.82}_{-1.67}$	$31.2^{+23.7}_{-13.4}$	337^{+396}_{-186}
ig (59.6%)	$.462^{+.879}_{-.302}$	$4.99^{+4.49}_{-2.36}$	$31.6^{+26.9}_{-14.5}$	213^{+170}_{-95}
ex (3.3%)	$.292^{+.215}_{-.122}$	$4.40^{+2.55}_{-1.62}$	$35.3^{+20.8}_{-13.1}$	620^{+298}_{-201}
in (31.8%)	$.348^{+.679}_{-.230}$	$3.65^{+2.55}_{-1.50}$	$23.5^{+13.9}_{-8.7}$	213^{+206}_{-105}
$L_e^{\{uc\}}$ ($p=0.5$)	$.310^{+.290}_{-.150}$	$4.90^{+4.58}_{-2.24}$	$30.1^{+28.1}_{-13.8}$	487^{+455}_{-235}
RSD model	$.597^{+.756}_{-.351}$	$4.79^{+0.82}_{-0.70}$	$32.0^{+7.0}_{-5.8}$	510^{+211}_{-149}

$L_e(k)$, $k=2$, 5, 7 and 10, averaged over 865 chromosomes. Total sequences length is about 2.2×10^{10} bases. Abbreviations: All, complete chromosome; *gn*, genes; *ig*, intergenic; *ex*, exons; *in*, introns. Percentage given indicates portion of complete sequence. $L_e^{\{uc\}}$ is defined in Eq. (1) and RSD results are averaged over 200 model sequences. See Table S4 for $L_e(k)$ of other k values.
doi:10.1371/journal.pone.0009844.t002

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_{fl} 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:

$$\bar{\sigma}_{fl}(k)\approx0.14^{+0.05}_{-0.04}\;10^{-k/2}L, \tag{3}$$

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 $p\approx0.5$, $\bar{\sigma}^2_{fl}$ reduces to the usual variance. Note that for random sequences $\sigma_{fl}(k)\sim L^{1/2}4^{-k/2}$. Since L is large, genomic $\bar{\sigma}_{fl}$ can be orders of magnitude greater than its random counterpart. For instance, for the 4.6 Mb chromosome, the $k=4$ values for $\bar{\sigma}_{fl}$ 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

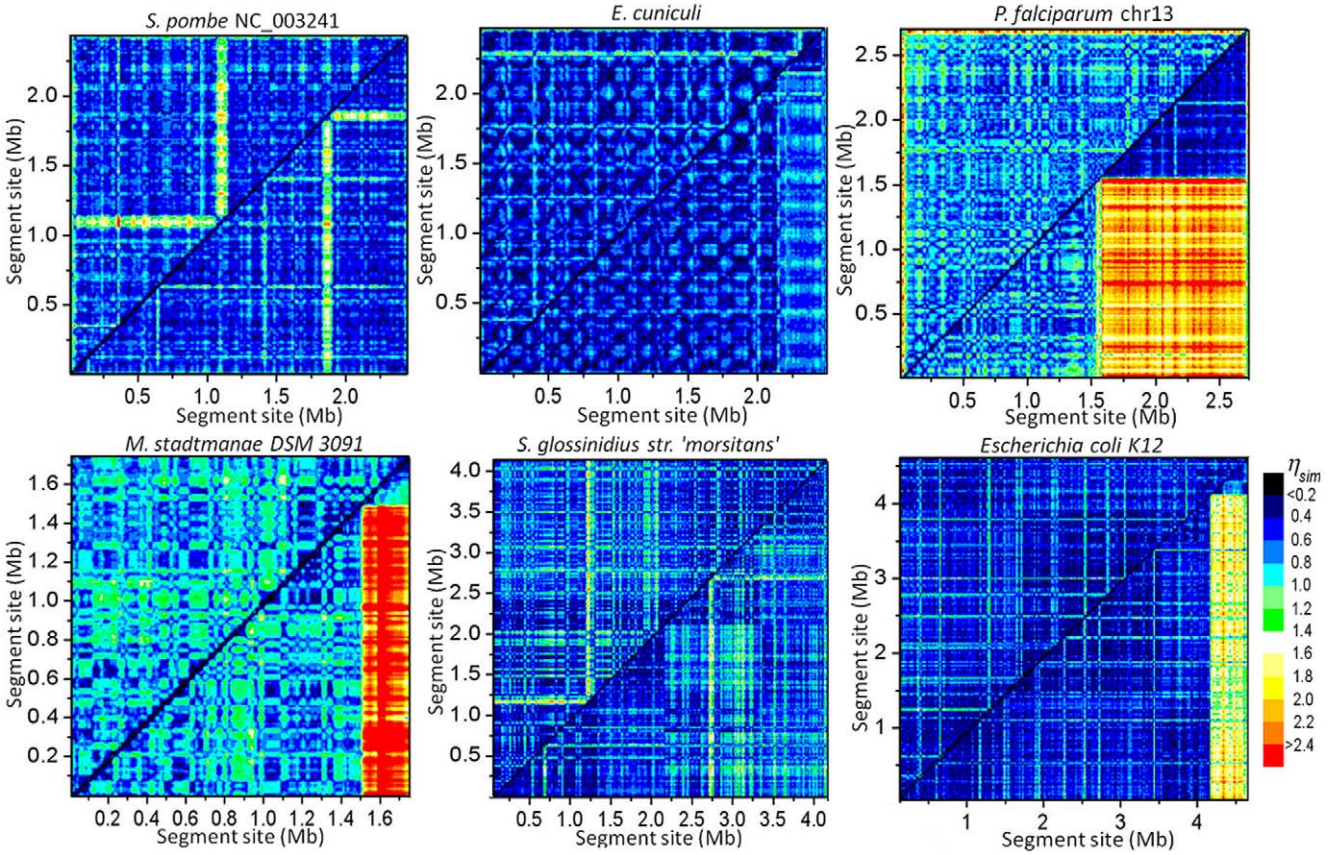

Figure 5. Intra-chromosomes similarity plots. Plots are for $k = 2$ (Methods). Sliding window has width 25 kb and slide 10 kb; pixel size is 10 kb by 10 kb. In each plot, the coordinates for the upper-left triangle are sites along the chromosome (*chr*), and those for the lower-right triangle are along a concatenate composed of gene (*gn*, left side) and intergene (*ig*, right side) parts. In effect, the upper-left triangle shows *chr-chr* similarity, and the lower-right triangle shows *gn-gn* (lower-left sub-triangle), *ig-ig* (upper-right sub-triangle), and *gn-ig* (rectangular) similarities in three separate regions. The lengths of the *gn* and *ig* parts are given in Table 3.
doi:10.1371/journal.pone.0009844.g005

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, and 8, respectively. Because $\bar{\sigma}_{fl}(k)/\sigma_{fl}^{\{ran\}}(k) \propto L^{1/2}(0.4)^{k/2}$, 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. coli* chromosome), fourteen of the sixteen 2-mers have $P \lesssim P_8$ ($= 1.3 \times 10^{-15}$), 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_e

We now discuss probable causes for the formation of the universality class. We first list some general properties of the ratio ρ of l_e to the sequence length l: if the sequence is (nearly) random then $\rho (= l_e/l) \approx 1$; 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

Table 3. Intra-chromosome similarity indexes.

	Length (Mb)/*p*			Average η_{sim}			
Organism	*chr*	*gn*	*ig*	*chr-chr*	*gn-gn*	*ig-ig*	*gn-ig*
S. pombe Chr. 1	2.45/0.64	1.40/0.61	1.05/0.69	0.648	0.569	0.615	0.647
E. cuniculi (genome)	2.50/0.53	2.15/0.53	0.35/0.55	0.527	0.481	0.450	0.666
P. falciparum Chr. 13	2.73/0.82	1.55/0.79	1.18/0.87	0.801	0.742	0.641	2.11
M. stadtmanae	1.77/0.73	1.51/0.71	0.26/0.83	0.805	0.782	0.757	2.52
S. glossinidius morsitans	4.17/0.46	2.15/0.44	2.02/0.47	0.638	0.510	0.635	0.729
E. coli K12	4.64/0.50	4.12/0.49	0.52/0.58	0.517	0.481	0.548	1.63

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

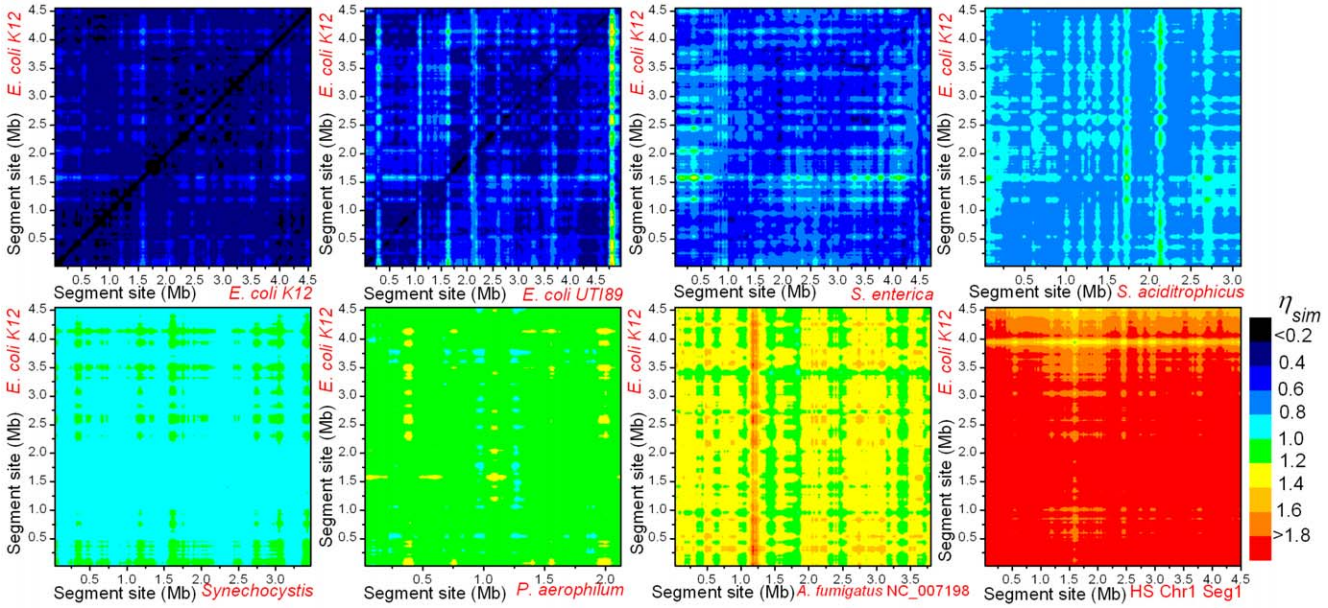

Figure 6. Intra- *E. coli* and inter-chromosome similarity plots. The plots are those of *E. coli* chromosome *vs.* the chromosomes of, left to right and top to bottom, *E. coli*, *E. coli UT189*, *Salmonella*, the delta-proteobacteria *S. aciditrophicus*, the cyanobacteria *Synechocystis*, the archaea *P. aerophilum*, chromosome 5 of the fungus *A. fumigatus*, and the first 4.5 Mb segment from chromosome 1 of *H. sapiens*. Coordinates are sites along the sequence. Sliding window width is 100 kb and slide is 25 kb, pixel size is 25 kb by 25 kb.
doi:10.1371/journal.pone.0009844.g006

the n-fold replication of a random sequence, then $\rho \approx 1/n$. We illustrate how segmental duplication can cause a sequence to have ρ 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, $l_e(\mathrm{X}) \approx l_e(\mathrm{Y})$ and concatenating X and Y is similar to doubling X by replication, hence $l_e(\mathrm{XY})$ will be nearly equal to $l_e(\mathrm{X})$.

In general, if the k-mer-contents of X and Y are similar, then (provided the sequences are sufficiently long) we expect $l_e(\mathrm{XY}) \approx l_e(\mathrm{X}) \approx l_e(\mathrm{Y})$. Conversely, if the k-mer-contents of X and Y are significantly different, then we expect $l_e(\mathrm{XY}) > \min(l_e(\mathrm{X}), l_e(\mathrm{Y}))$ (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). The first two rows of results in Table 5 verify that for random sequence ρ is always close to one, or $l_e \approx l$. 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_i have k-mer-contents relatively more similar to A (Figs. 4 and 5), therefore $l_e(\mathrm{AC}_i) \approx l_e(\mathrm{AA}') \approx l_e(\mathrm{A})$ as expected. Chromosomes labeled D_i and B have k-mer-contents more dissimilar to A, therefore $l_e(\mathrm{AX}) > \min(l_e(\mathrm{A}), l_e(\mathrm{X}))$. The case of AD_4, where D_4 is *H. sapiens chr. 1*, is not an exception to the rule even for $k = 2$, because $l_e(\mathrm{D}_4) < l_e(\mathrm{A})$. In the bottom portion of Table 5 the approximate relation $l_e \approx n^2 l_{e0}$ (Table S5; l_{e0} is the equivalent length of the genomic portion and n is the ratio of the

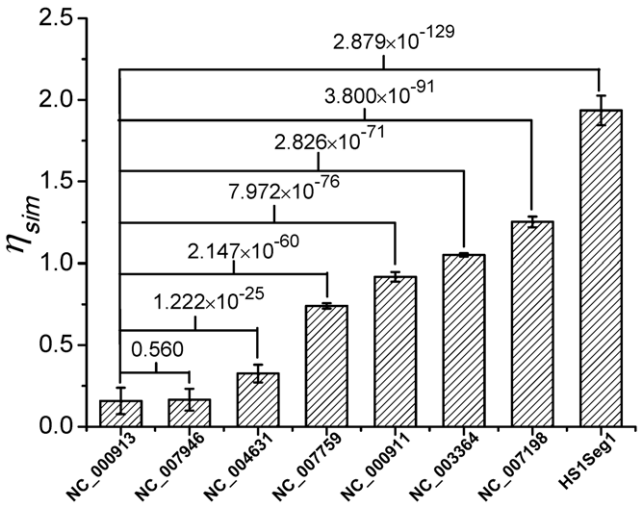

Figure 7. Comparison of inter-chromosome similarity matrices. Mean values and SD of the eight η_{sim}-plots (of η_{sim}-matrices) shown in Fig. 6 and P-values for the null assumption that the 2nd to 7th cases are the same as the 1st case.
doi:10.1371/journal.pone.0009844.g007

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

Fraction of k-mers whose P-value is less than P_3, P_6, or P_8						
	$k=2$ ($L_e=310$ b)			$k=9$ ($L_e=194$ kb)		
Length (Mb)	$P<P_3$	$P<P_6$	$P<P_8$	$P<P_3$	$P<P_6$	$P<P_8$
0.8	0.953	0.906	0.875	0.139	0.0031	0.0001
4.6	0.980	0.960	0.955	0.538	0.418	0.100
30	0.992	0.985	0.979	0.809	0.628	0.519
226	0.997	0.994	0.992	0.930	0.860	0.815

P-values for k-mer distribution given by Eq. (1) (at $p=0.5$). Null theory assumes genomes are random sequences. The P-values $P_3 = 2.7 \times 10^{-3}$, $P_6 = 2.0 \times 10^{-9}$, and $P_8 = 1.3 \times 10^{-15}$ correspond to z-values of three, six and eight, respectively.
doi:10.1371/journal.pone.0009844.t004

Table 5. Equivalent lengths of composite sequences.

	l_e			
	$k=2$		$k=6$	
Sequence	$l=50$	$l=200$	$l=50$	$l=200$
R	47.5±28.2	154±126	48.6±1.5	192±5
RR′	37.0±16.2	124±46	48.2±1.2	197±5
A	.348±.037	.360±.033	9.55±.69	11.7±.7
AA′	.357±.046	.352±.023	9.88±1.07	11.1±.7
AC_1	.351±.061	.361±.021	9.37±1.01	11.5±.6
AC_2	.354±.043	.384±.045	9.18±.83	11.6±.9
AC_3	.359±.051	.371±.034	11.0±.9	14.2±1.5
AD_1	.411±.044	.423±.024	11.8±.9	14.3±.6
AD_2	.942±.275	1.05±.09	14.9±1.4	20.4±1.1
AD_3	.598±.104	.613±.052	17.9±1.6	24.0±1.6
AD_4	.324±.052	.383±.055	11.2±1.9	16.9±1.9
B	.124±.029	.166±.099	5.17±.68	6.54±2.00
BB′	.232±.155	.258±.183	6.16±1.94	7.54±2.30
AB	.463±.241	.502±.263	11.2±1.9	15.2±3.5
RA	1.19±.09	1.34±.20	22.6±1.2	38.5±3.0
RB	.575±.321	.754±.637	15.6±4.2	23.3±8.5
RAB	.873±.424	1.10±.49	18.4±3.2	31.3±6.0
RR′A	2.63±.66	3.16±.30	31.5±2.1	72.2±6.8
RR′B	1.03±.62	1.37±.70	22.9±4.5	44.7±14.3

Equivalent lengths l_e of composite sequences of total length l (in kb). The composite XY is the concatenation of two equal-length components X and Y. Similarly for the composite XYZ. A and A′ are segments from *E. coli*, and B and B′ are from *C. tetani* (2.80 Mb, $p=0.70$). $C_{1,2,3}$ and $D_{1,2,3,4}$, are the seven "other" chromosomes in Fig. 6, in the order given there. R and R′ are $p=0.5$ random sequences. Results are averaged over 10 samples in all cases.
doi:10.1371/journal.pone.0009844.t005

length of the concatenate to the that of the genomic portion) is seen to hold: $l_e(\mathrm{RX}) \approx 4l_e(\mathrm{X})$ (X being A or B), $l_e(\mathrm{RAB}) \approx 2.3l_e(\mathrm{AB})$, and $l_e(\mathrm{RR'X}) \approx 9l_e(\mathrm{X})$.

Artificial sequences generated by RSD growth model exhibit universal L_e

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_e's very close to the universal $L_e^{\{uc\}}$ 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 effect L_e, for simplicity they are not given special treatment in this study.) The three parameters of the model are L_0 (initial length), $\bar{d}$ (average duplicated segment length), and r (cumulative point mutation per-base density) (Methods. L_e generated by the model is insensitive to sequence length provided it is longer than 0.5 Mb, allows a generous range in $\bar{d}$ and a tighter range in r, and is highly sensitive to L_0 (Fig. S3, *SI*). (Because RSD will at least initially cause L_e to be longer than L_0 and because L_e $(k=2) \approx 300$ 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 $\bar{d}=120$ to 5000 and $r=0.65$ to 0.80. L_e's of model sequences obtained using the "best set" of parameters $L_0=64$, $\bar{d}=1000$, and $r=0.73$ are shown in the right panel in Fig. 8, where the lines represent the universality class $L_e^{\{uc\}}$ (Eq. (1)). The $\langle\chi^2\rangle$ for these L_e's is 0.18 and implies that on average, the model L_e and $L_e^{\{uc\}}$ agree to within a factor of 1.6. This small χ^2 can easily be increased to match that of the genomic data ($\langle\chi^2\rangle=0.43$) by using model parameters that cover suitable ranges of values centered around the best values.

The range of $\bar{d}$ 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 $\bar{d}=30$, $r=0.3$ 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 χ^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 λ, we have $\lambda=(t_2-t_1)^{-1}\ln(L_2/L_1)$ where L_i 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 (10^6 years) [39]. The ancestors of eubacteria and archaea-eukaria diverged $\sim$3.4 Gya (10^9 years) ago [59–61]), and before that proto-genomes 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_1 to be from 0.05 to 0.2 Mb and $L_2=3$ Gb. Then $\lambda_{hs}=2.7\sim3.7$/Mya. These rates imply the human genome grew $14\sim20$% in the last 50 Mya, in agreement with [39]. If we assume the growth is purely SD and take the length of duplicated segment $\bar{d}$ to be 500 b to 2 kb, then the rate of SD events is $\mu_{SD,hs}=\lambda_{hs}/\bar{d}=1.4\sim7.4$/Mb/Mya. These values are comparable to the estimates of 3.9/Mb/Mya (from animal gene duplication rate of $\sim$0.01 per gene per Mya [6] and human coding region $\sim$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 μ_p of "point mutations" – including small deletion and insertion but excluding SD – by $\mu_p \approx r\lambda/2$ (Eq. (19), Methods). If we take the best value $r=0.73$ from the RSD model then $\mu_{p,hs}=0.98\sim1.4\times10^{-3}$/site/Mya. This agrees well with the value $\mu_{sc,hs}\sim1\times10^{-3}$/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-eukaria 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 $\lambda_{ec}=5.4\sim11$/Mya, and $\mu_{p,ec}=2.0\sim4.0\times10^{-3}$/site/Mya. Fortuitously or perhaps this range of rates represent an equilibrium value, it is compatible with the sequence-comparison *E. coli* rate of

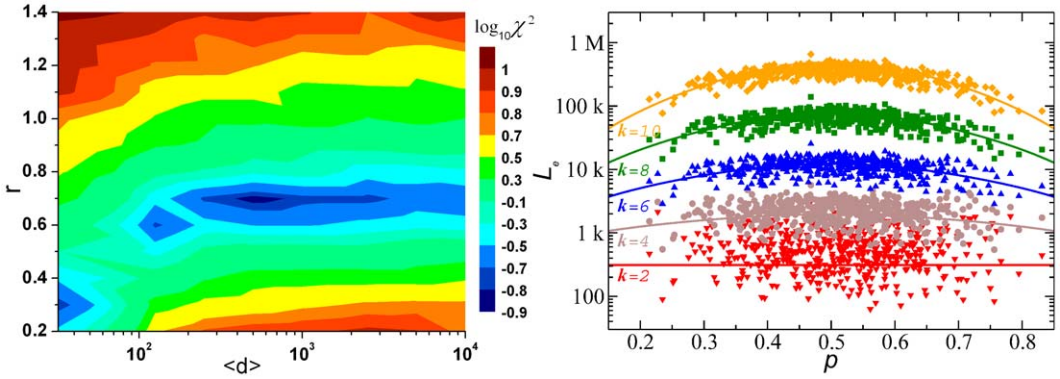

Figure 8. Results from minimal RSD model. Left: Equi-χ^2 contour on the r-$\bar{d}$ plane, with $L_0 = 64$ (bases). Right: $L_e(k)$, $k = 2, 4, 6, 8, 10$ from 200 model sequences of length 2 Mb generated using the "best set" of parameters $L_0 = 64$, $\bar{d} = 1000$ (b) and $r = 0.73$ (b^{-1}). Lines in right panel are $L_e^{\{uc\}}(k;p)$ (Eq. (1)).
doi:10.1371/journal.pone.0009844.g008

$\mu_{sc,ec} \sim 5 \times 10^{-3}$/site/Mya based on mutations that (putatively) occurred in the last 0.5 Gya or less [37,38]. There is some 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* [65], *C. elegans* [65,66], and *Drosophila* [65,67] tend to be two or three orders of magnitudes higher than the characteristic rates of ~ 0.001/site/Mya of wild types.

Presumably the same selective force is what causes the L_e'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 865 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 archaea) 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. cuniculi* (11), *E. gossypii* (7), *Kluyveromyces lactis* (6), *S. cerevisiae* (16), *S. pombe* (3), *Y. lipolytica* (6), *P. falciparum* (14)), 5 insects (*A. gambiae* (3), *A. mellifera* (16), *C. elegans* (6), *D. melanogaster* (4), *T. casteneum* (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. multatta* (21), *M. musculus* (21), *P. troglodytes* (25), *R. norvegicus* (21)). The complete list of sequences, their 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 $\tau \equiv 4^k$ types of k-mers by S. Given a sequence, we count the frequency of occurrence (or frequency) f_u of each k-mer-type u in S using an overlapping sliding window of width k and slide one [36]. Then the sum of the frequencies is $\sum_{u \in S} f_u = L - k + 1$, here approximate by L, and the mean frequency is $\bar{f} = L/\tau$. Let the fractional AT- and CG-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 {AA AT, TA, TT}. The 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)

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

Let $L_m = \sum_{u \in S_m} f_u$ 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 $\bar{f}_m = L_m/\tau_m$. The large-L limit of $\bar{f}_m$ for a random sequence, $\bar{f}_m^{\{ran\}}$, is obtained from the binomial expansion

$$L = \bar{f}\tau = \bar{f} 4^k (p+q)^k = \sum_{m=0}^{k} \left(2^k \binom{k}{m}\right)\left(2^k p^m q^{k-m} \bar{f}\right) = \sum_{m=0}^{k} \tau_m \bar{f}_m^{\{\infty\}}. \qquad (5)$$

That is,

$$\bar{f}_m^{\{\infty\}} \equiv \lim_{L \to \infty} \bar{f}_m^{\{ran\}} = 2^k p^m q^{k-m} \bar{f}. \qquad (6)$$

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

Fluctuation in occurrence frequency

The coefficient of variation of the frequency distribution is $CV = \sigma/\bar{f}$, where σ 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 $\sigma^2 = \bar{f}$, hence $CV^2 = \bar{f}^{-1} = \tau/L$, 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

Table 6. Average frequency of occurrence ($\bar{f}_m$) of 5-mers in $p \approx 0.5$ and $p \approx 0.7$ sequence.

	f_m					
Sequence	**(m=) 0**	**1**	**2**	**3**	**4**	**5**
$p = 0.492$						
E. coli	2509	2245	1877	1760	1944	2656
Random	2101	2044	1987	1922	1857	1795
$\lim_{L\to\infty}$ Random*	2114	2048	1983	1920	1860	1801
$p = 0.691$						
C. acetobutylicum	154	397	918	1951	4272	10300
Random	176	394	882	1970	4400	9832
$\lim_{L\to\infty}$ Random*	176	393	880	1968	4402	9845

All sequences normalized to a length of 2 Mb; $\bar{f} = 2\times10^6/4^5 = 1953$. Random means matching random sequence, or sequence obtained by scrambling the genome. *Values of $\bar{f}_m^{\{\infty\}}$ given by Eq. (6).

doi:10.1371/journal.pone.0009844.t006

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:

$$\sigma^2 = \tau^{-1}\sum_{u\in S}(f_u-\bar{f})^2 = \tau^{-1}\sum_{m=0}^{k}\sum_{u\in S_m}(f_u-\bar{f}_m+\bar{f}_m-\bar{f})^2$$
$$= \tau^{-1}\sum_{m=0}^{k}\left(\tau_m(\bar{f}_m-\bar{f})^2+2(\bar{f}_m-\bar{f})\sum_{u\in S_m}(f_u-\bar{f}_m)+\sum_{u\in S_m}(f_u-\bar{f}_m)^2\right). \quad (7)$$

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

$$\sigma^2 \equiv {\sigma_{nf}}^2 + {\sigma_{fl}}^2, \quad (8)$$

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

$${\sigma_{nf}}^2 = \sum_{m=0}^{k}\frac{\tau_m}{\tau}(\bar{f}_m-\bar{f})^2, \quad (9)$$

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

$${\sigma_{fl}}^2 = \sum_{m=0}^{k}\sum_{u\in S_m}\frac{(f_u-\bar{f}_m)^2}{\tau} \equiv \sum_{m=0}^{k}\frac{\tau_m}{\tau}\,\sigma_{m,fl}^2. \quad (10)$$

Thus,

$$CV^2 = (\sigma/\bar{f})^2 = (\sigma_{nf}/\bar{f})^2 + (\sigma_{fl}/\bar{f})^2 \equiv {CV_{fl}}^2 + {CV_{nf}}^2. \quad (11)$$

The non-fluctuating, or ''non-statistical'', part, CV_{nf}, has a well-defined value in the large-L limit, obtained by replacing $\bar{f}_m$ by $\bar{f}_m^{\{\infty\}}$ in Eq. (9):

$$\left(CV^{\{\infty\}}\right)^2 \equiv \lim_{L\to\infty} {CV_{nf}}^2 = \sum_{m=0}^{k} 2^{-k}\binom{k}{m}\left(2^k p^m q^{k-m}-1\right)^2$$
$$= 2^k(p^2+q^2)^k - 1, \quad (12)$$

which has a strong dependence on p and vanishes $p = 0.5$. Because genomes are large, $CV^{\{\infty\}}$ gives an accurate description of CV_{nf} 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 $\sigma_{m,fl}^2 \to \bar{f}_m$ in the large-L limit. Therefore, from Eq. (10),

$$\lim_{L\to\infty} {CV_{fl}}^2 = \frac{1}{\bar{f}^2}\lim_{L\to\infty}\sigma_{fl}^2 = \frac{1}{\bar{f}^2}\sum_m\frac{\tau_m}{\tau}\bar{f}_m$$
$$= \frac{1}{\bar{f}} = \frac{\tau}{L} \qquad \text{(random sequence)}, \quad (13)$$

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 ${CV_{nf}}^2$ becomes the leading term in CV^2 (when $p \neq 0.5$). When $p = 0.7$, this crossover value is 42, 316 and 2851 (bases) for $k = 2$, 4, 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 $\bar{f}_m$, $\sigma_{m,fl}^2$, and CV_{fl}^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 $\bar{f} = 2\times10^6/4^5 = 1953$. However, the genomic spectrum is seen to be much broader then the random-sequence spectrum, indicating that whereas in the random sequence frequencies (f_u) of individual 5-mers deviate little from the mean ($\bar{f}$), 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, $\bar{f}_m^{\{\infty\}} \approx 152$, 354, 827, 1930, 4500, 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

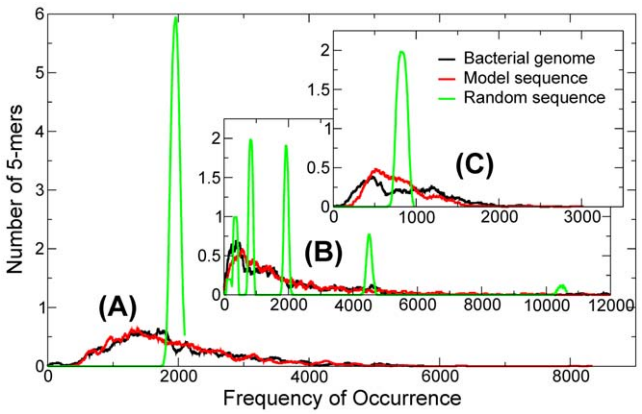

Figure 9. Frequency distributions of 5-mers. Frequency occurrence distributions, or spectra, of 5-mers from the genomes of two prokaryotes, (A) *E. coli* (with (A+T) content $p \approx 0.5$) and (B) *C. acetobutylicum* ($p \approx 0.7$), normalized to a sequence length of 2 Mb. Abscissa give occurrence frequency and ordinates give number of 5-mers averaged, for better viewing, over a range of 21 frequencies to reduce fluctuation. The black, green and red curves represent spectra of the complete genomes, the randomized genome sequences and sequences generated in a model (see text), respectively. (C) Details of the m = 2 subspectra from (B).
doi:10.1371/journal.pone.0009844.g009

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 $\bar{f}_m^{\{\infty\}}$ gives a very accurate estimate of $\bar{f}_m$ for random sequences and a fair one for genomic sequences. In the $p = 0.492$ case, the relation $\bar{f}_m \approx \bar{f}$ 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

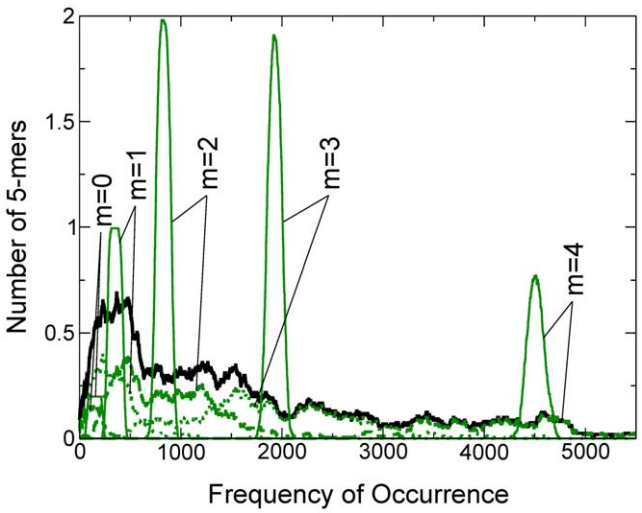

Figure 10. Frequency distributions of 5-mers in *m*-sets. Details of $k = 5$, m-specific subspectra from the *C. acetobutylicum* genome (broken green curves) and matching random sequence (solid green curves); black curve is the same as in (B) Fig. 9. The five narrow subspectra peak (approximately) at $\bar{f}_m^{\{\infty\}}$, $m = 0$ to 4, or at 152, 354, 827, 1939, 4500, respectively; the $m = 5$ peak at 10500 is off scale (see Fig. 9 (B)).
doi:10.1371/journal.pone.0009844.g010

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) $CV^{\{\infty\}}$ gives an excellent estimate of CV_{nf} for random sequences, and a fair estimate for genomes; (iii) CV_{fl} 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

$$l_e = b_k \tau / CV_{fl}{}^2 \qquad (\text{for } k \geq 2) \tag{14}$$

where $CV_{fl}{}^2$ is given by the frequency distribution of k-mers. Recalling that for a random sequence $CV_{fl}{}^2$ is inversely proportional sequence length (Eq. (13)), we see that l_e is the length of a random sequence whose $CV_{fl}{}^2$ 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_{fl} 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 them 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 α and β, the similarity index $\eta_{sim}(\alpha,\beta)$ for the pair is defined as

$$\eta_{sim}^2(\alpha,\beta) = \frac{1}{k+1} \sum_m \frac{1}{2\tau_m} \sum_{u \in S_m} \frac{(f_u^{\{\alpha\}} - f_u^{\{\beta\}})^2}{\sigma_m^{\{\alpha\}} \sigma_m^{\{\beta\}}} \tag{15}$$

where S_m is an m-set and σ_m^2 is the variance of the frequency of the k-mers in S_m. The pair are similar (in k-mer-content) when $\eta_{sim} \ll 1$, are (considered to be) identical when $\eta_{sim} = 0$, and are highly dissimilar when $\eta_{sim} \gtrsim 1$. If we divide α and β into (possibly overlapping) segments $\{\alpha_1,\alpha_2,\cdots\}$ and $\{\beta_1,\beta_2,\cdots\}$, respectively, then we call the matrix whose element (i,j) is valued $\eta_{sim}(\alpha_i,\beta_j)$ 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

Table 7. Values of σ's from 5-mers in $p \approx 0.5$ and $p \approx 0.7$ sequences.

	$\sigma^2_{m,fl}$ (in units of 10^4)						CV^2_{fl}	CV^2_{nf}	$(CV^{\{\infty\}})^2$
Sequence	**(m=) 0**	**1**	**2**	**3**	**4**	**5**			
$p = 0.492$									
E. coli	144	141	74.2	58.4	66.4	83.7	0.212	0.013	– –
Random	.174	0.203	0.185	0.177	0.144	0.110	4.6×10^{-4}	0.0012	0.0013
$p = 0.691$									
C. acetobutylicum	0.60	6.95	26.1	65.4	97.1	336	0.145	1.00	– –
Random	0.011	0.038	0.102	0.218	0.500	1.24	5.8×10^{-4}	0.969	0.976

All sequences normalized to a length of 2 Mb; for $k = 5$, $\bar{f} = 1953$, $\tau = 1024$, and $\tau_m = 32, 160, 320, 160, 32$, for $m = 0$ to 5.
doi:10.1371/journal.pone.0009844.t007

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; $\bar{d}$, 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 $2\bar{d}$, unless the current length of the genome L' is less than $2\bar{d}$, 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, μ_{SD}, and the rate density of "point mutation" – including small deletion and insertion but excluding SD – events, μ_p. If the genome grows from time t_1 to time t_2 at a rate proportional to its length l, that is, $\Delta l = \lambda l \Delta t$ where λ is the event rate (number of events per unit of time), then

$$\lambda = (t_2 - t_1)^{-1} \ln (l_2/l_1), \tag{16}$$

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

$$\mu_{SD} = \lambda/\bar{d}. \tag{17}$$

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

$$n_p(l_2) - n_p(l_1) = \mu_p (l_2 - l_1)/\lambda, \tag{18}$$

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 $\Delta n_c = n_p \Delta l/l \approx \mu_p l \Delta t$ (for large l). Therefore $n_c \approx n_p$, that is, the cumulative number of mutated sites is twice n_p. At full genome length L, this number is rL, hence

$$\mu_p \approx r\lambda/2. \tag{19}$$

Supporting Information

Figure S1 Category L_e for coding and non-coding parts. Averages of p (fractional A/T-content) and L_e for $k = 7$ (situations for other ks are similar) for the coding parts (solid symbols; *ex* 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). Numeral indicates number of chromosomes in each category. The curve represents L_e for the universality class: $L_e^{\{uc\}}(k; p)$.
Found at: doi:10.1371/journal.pone.0009844.s001 (0.26 MB TIF)

Figure S2 Distributions of χ^2 versus L and p. Each symbol gives the χ^2 for one chromosomal L_e. Top panels, for genic (*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 χ^2 is less than 10^{-3} given in brackets, stand for: diamond, *gn* (blue; 7100; 229); square, *ex* (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-χ^2 contour as function of r and d, with $L_0 = 64$ (bases); length (L) of generated model sequence is 2 Mb and only $L_e(k)$ results for $k = 7$ are used. Top-right: $L_e(k)$, $k = 2$, 4, 6, 8, 10 from 200 model sequences generated using the "best" parameters $L_0 = 64$, $\langle d \rangle = 1000$ (b) and $r = 0.73$ (cumulative point mutations per base). The lines are $L_e^{\{uc\}}(k; p)$ that represent the universality class

given in the main text. The χ^2 for the model sequences is 0.18. Bottom-left: χ^2 versus L_0 (otherwise best parameters); model sequences have $L = 2$ Mb and $p = 0.5$. Bottom-right: L_e versus L, for a $p = 0.5$ model sequence generated using the best parameters.
Found at: doi:10.1371/journal.pone.0009844.s003 (1.17 MB TIF)

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 $L_e(k)$, $k = 2$ to 10, averaged over categories of organisms.
Found at: doi:10.1371/journal.pone.0009844.s006 (0.06 MB PDF)

Table S4 L_e of sequences with highly biased compositions.
Found at: doi:10.1371/journal.pone.0009844.s007 (0.06 MB PDF)

Table S5 Effect of replication and segmental duplication on l_e.
Found at: doi:10.1371/journal.pone.0009844.s008 (0.04 MB PDF)

Text S1
Found at: doi:10.1371/journal.pone.0009844.s009 (0.07 MB PDF)

Author Contributions

Conceived and designed the experiments: HCL. Performed the experiments: HDC WLF SGK. Analyzed the data: HDC WLF SGK HCL. Wrote the paper: HCL.

References

1. Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76: 5269–5273.
2. Li WH (1997) Molecular Evolution Sinauer Associates: Sunderland, MA.
3. Ohno S (1970) Evolution by Gene Duplication Springer-Verlag: Berlin.
4. Hansche PE, Beres V, Lange P (1978) Gene duplication in *Saccharomyces cerevisiae*. Genetics 88: 673–687.
5. Yamanaka K, Fang L, Inouye M (1998) The CSPA family in *Escherichia coli*: multiple gene duplication for stress adaptation. Mol Microbiol 27(2): 247–255.
6. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
7. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
8. Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18(6): 292–298.
9. Lewin B (2000) Genes VII Oxford Univ Press. pp 89–115.
10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
12. Kleckner N (1981) Transposable elements in prokaryotes. Ann Rev Genet 15: 341–404.
13. Castilho BA, Olfson P, Casadaban MJ (1984) Plasmid insertion mutagenesis and lac gene fusion with mini-mu bacteriophage transposons. J Bacteriol 158(2): 488–495.
14. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell 75(6): 1083–1093.
15. Li WH, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292: 237–239.
16. Vanin EF (1985) Processed pseudogenes: Characteristics and evolution. Annu Rev Genet 19: 253–272.
17. Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons: genes, pseudogenes, and trans- posable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55: 631–661.
18. Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol 16(6): 314–321.
19. McGrath JM, Jancso MM, Pichersky E (1993) Duplicate sequences with a similarity to expressed genes in the genome of *Arabidopsis thaliana*. Theor Appl Genet 86: 880–888.
20. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: Organization and impact within the current human genome project assembly. Genome Res 11: 1005–1017.
21. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007.
22. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Human Genet 77: 78–88.
23. Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A 94: 6809–6814.
24. Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci U S A 95: 1971–1974.
25. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. Cell 110: 689–699.
26. Coghlan A, Wolfe KH (2002) Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res 12: 857–867.
27. Pevzner P, Tesler G (2003) Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. Genome Res 13: 37–45.
28. Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc Natl Acad Sci U S A 97: 4168–4173.
29. Spring J (2002) Genome duplication strikes back. Nat Genet 31: 128–129.
30. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome dupli- cation in the yeast *Saccharomyces cerevisiae*. Nature 428: 617–624.
31. Peng CK, Buldyrev SV, Goldberg AL, Havlin S, Simons M, et al. (1993) Finite-size effects on long-range correlations: Implications for analyzing DNA sequences. Phys Rev E 47: 3730–3733.
32. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, et al. (1994) Linguistic features of noncoding DNA sequences. Phys Rev Lett 73: 3169–3172.
33. Forsdyke D (1995) Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41: 573–581.
34. Karlin S, Mrazek J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci U S A 94: 10227–10232.
35. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classiffication of species assessed by chaos game representation of sequences. Mol Biol Evol 16(10): 1391–1399.
36. Hao BL, Lee HC, Zhang SY (2000) Fractals related to long DNA sequences and complete genomes. Chaos, Solitons and Fractals 11: 825–836.
37. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96: 12638–12643.
38. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.
39. Liu G, Program NCS, Zhao S, Bailey JA, Sahinalp SC, et al. (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. Genome Res 13: 358–368.
40. Voss RF (1996) Comment on "Linguistic features of noncoding DNA sequences". Phys Rev Lett 76: 1978.
41. Bonhoeffer S, Herz AV, Boerlijst MC, Nee S, Nowak MA, et al. (1996) No signs of hidden language in noncoding DNA. Phys Rev Lett 76: 1977.
42. Israeloff NE, Kagalenko M, Chan K (1996) Can Zipf distinguish language from noise in noncoding DNA? Phys Rev Lett 76: 1976.
43. Mantegna RN, Buldyrev SV, Goldberger AL, Halvin S, Peng CK, et al. (1996) Mantegna et al. reply:. Phys Rev Lett 76: 1979–1981.
44. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, et al. (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49: 1685–1689.
45. Bernaola-Galvffan P, Carpena P, Roman-Roldan R, Oliver JL (2002) Study of statistical correlations in DNA sequences. Gene 300: 105–115.
46. Messer PW, Arndt PF, Lassig M (2005) Solvable sequence evolution models and genomic correlations. Phys Rev Lett 94: 138103.
47. Fickett JW, Torney DC, Wolf DR (1992) Base compositional structure of genomes. Genomics 13: 1056–1064.
48. Xie HM, Hao BL (2002) Visualization of k-tuple distribution in procaryote complete genomes and their randomized counterparts. In: Proceedings of the IEEE Computer Society Bioinformatics Conference. pp 31–42.
49. Hsieh LC, Luo LF, Lee HC (2003) Genomes are large systems with small-system statistics: Seg- mental duplication in the growth of microbial chromosomes. AAPPS Bulletin 13: 22–27.
50. Chen TY, Hsieh LC, Lee HC (2005) Shannon information and self-similarity in complete chromosomes. Comput Phys Commun 169: 218–221.
51. Zhou F, Olman V, Xu Y (2008) Barcodes for genomes and applications. BMC Bioinformatics 9: 546.
52. Kong SG, Chen HD, Fan WL, Wigger J, Torda AE, et al. (2009) Quantitative measure of random- ness and order for complete genomes. Phys Rev E 79: 061911.

53. Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol 12: 406–411.
54. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361–375.
55. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401–1404.
56. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet 21: 673–682.
57. Messer PW, Bundschuh R, Vingron M, Arndt PF (2007) Effects of long-range correlations in DNA on sequence alignment score statistics. J Comput Biol 14: 655–668.
58. Bartel DP (2004) Micrornas: Genomics, biogenesis, mechanism, and function. Bioinformatics 116: 281–297.
59. Doolittle WF (1997) Fun with genealogy. Proc Natl Acad Sci U S A 94: 12751–12753.
60. Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: Update and reevaluation. Proc Natl Acad Sci U S A 94: 13028–13033.
61. Hedges SB (2002) The origin and evolution of model organisms. Nat Rev Genet 3: 838–849.
62. Woese CR (1998) The universal ancestor. Proc Natl Acad Sci U S A 95: 6854–6859.
63. Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci U S A 99: 8742–8747.
64. Glansdorff N, Xu Y, Labedan B (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol Direct 3: 29.
65. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667–1686.
66. Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature 430: 679–682.
67. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, et al. (2007) Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature 445: 82–85.
68. GenBank (2009) The genbank genome database. URL http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome.
69. Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. iii. direct analysis. Proc Natl Acad Sci U S A 60: 921–922.
70. Chen HD, Chang CH, Hsieh LC, Lee HC (2005) Divergence and Shannon information in genomes. Phys Rev Lett 94: 178103.