Algebraic length distribution of sequence duplications in whole genomes

Eddy Taillefer, and Jonathan Miller

***Abstract*—The field of comparative genomics relies upon inference of neutrality or selection from sequence conservation. Recent studies of exactly-conserved sequences have revealed an anomalous, algebraic distribution of conserved sequence lengths that is inconsistent with standard models of neutral evolution based solely on local mutation. It has been proposed that linkage contributes to the shape of this anomalous distribution. Here we identify, for a variety of species, all 'maximal' repeats, direct or reverse-complement, within a chromosomal or whole-genome sequence of a single genome. For a set of maximal repeats of a given nucleotide length L, we report that the number of elements in the set $F(L)$ typically exhibits an algebraic tail. We propose a method based on a cost function that allows us to analyze this distribution and estimate the range over what the distribution is most likely to be well-approximated a power law. We find that the range is proportional to the genome size and that although the power-law exponent differs between species, it falls chiefly within a relatively narrow range of values. A sharp cut-off in the power-law regime is observed for some genomes that turns out to coincide with a peak in contig lengths and therefore can be attributed to artifacts of genome assembly, leading to a prediction that the extent of the power-law regime will increase as assemblies are improved. The typical algebraic behavior of length-frequency distribution is the most remarkable observation emerging from our analysis. The algebraic form of the empirical distribution of duplication lengths characterized here suggests that recombination events might as a general rule involve transfer of chunks of sequence with an algebraic length distribution. It also places strong constraints on any model of genome evolution. The observation of an algebraic distribution of exactly-duplicated sequence lengths within a genome is a direct demonstration of the net impact of linkage on genome evolution, and is consistent with the proposal that linkage contributes to the anomalous distribution of strongly-conserved sequence lengths.**

***Index Terms*—Genome, Chromosome, Ultraduplication, Frequency, length, distribution, power-law.**

I. Introduction

ULTRA-CONSERVATION has been defined as the sharing of sufficiently long identical subsequences of DNA among the genomes of three or more species [1]. In the field of comparative genomics, sequence conservation is interpreted as an indication of whether or not a sequence is under selection, and in this context, ultra-conservation has been taken as evidence of ultra-selection [2]; however, the extent and degree of sequence conservation depends on the number of species compared, the evolutionary distances among them, and the size of the genomes, among other factors. These factors can in principle be accounted for by phylogeny, which assigns appropriately reduced weights to similarities that emerge simply because the genomes compared did not evolve independently. Furthermore, existing methods in comparative genomics largely assume that the impact of recombination and repair processes can be safely neglected. Consequently, the proper interpretation of ultra-conservation depends on a set of calculations that, although worthwhile, has so far not been carried out.

Partly for this reason, we began in 2005 a systematic study of pairwise comparisons between genomes, in which we counted the numbers of contiguous sequences shared identically by two species. Among the first steps in investigating the origin of such sequences should be characterization of their length distribution and assessment of any differences between that distribution and what is expected based upon a suitable 'null model.' This assessment underlies — in principle — all existing methods of comparative genomics, yielding typically an 'E-' or 'p-' value indicating the likelihood that the sequence identity arose 'by chance,' by which is meant implicitly 'within the null model' or 'within a model for neutral evolution.'

When the length distributions of perfectly-conserved sequences across a small number (two or three) of diverse genomes were first computed, strong deviations from an exponential form were observed [3], [4]. Although smooth, the tails of the distributions turned out to be 'heavy' or 'stretched.' This observation is potentially important because existing models for genome sequence evolution (or null models for comparative genomics) entail an exponential (or 'geometric') distribution, as they are based on an assumption that correlations of conserved bases ('correlations of conservation') are strictly local — that is, based upon an assumption that neutral base substitutions at distinct sites in the genome are essentially independent of one another. The heavy tails, on the other hand, imply that base substitutions between human and mouse, for example, must be correlated with one another over distances on the order of a thousand bases. Nearest-neighbor or next-nearest-neighbor coupling can't rescue these models.

The form of the tail is smooth and homogeneous (algebraic, in fact) over many decades in scale, and homogeneous over regions of homologous sequence as short as a few million bases. Analogous tails are observed for pairwise conservation among bacterial species.

'Correlations of conservation' are familiar to biologists as a manifestation of 'linkage disequilibrium.' Although we originally interpreted them as reflecting negative selection, more recently we described a number of lines of evidence that they are in part reflections of linkage and recombination. Since recombination is largely neglected by existing approaches to

E. Taillefer and J. Miller are with Physics and Biology Unit, Okinawa Institute of Science and Technology, Okinawa, 1919-1 Tancha, Onna-son, Kunigami-gun, Japan 904-0412 (e-mail: {etaillefer,jm}@oist.jp)

comparative genomics and genome sequence evolution, such a conclusion would — if valid — have a major impact on the interpretation of sequence conservation.

Linkage reflects recombination inasmuch as sequences are fixed during the process of evolution as contiguous blocks, or 'chunks.' It is not easy to see how an algebraic tail could arise except if the lengths of these chunks themselves take on an algebraic distribution. In this manuscript, we report that the distributions of length of *duplicated* sequence within a *single* genome are typically algebraic, as first observed in chromosomal and whole-genome self-alignments [5]. This phenomenon, which we have called 'ultraduplication,' is seen in all clades with the possible exception of viruses.

Because alignment methods are heuristic, it was at first a significant concern that our observations might be artifacts of alignment. In contrast, the k-mer computations reported here involve rigorous, exhaustive sequence matching. Furthermore, k-mer methods turn out to be much faster than alignment, and we have been able to investigate and exhibit here comparisons for a far broader and more diverse set of genomes than is feasible by alignment. Perhaps most importantly, exact duplicates identified by alignment can only be a subset of those identified by exhaustive all-on-all sequence matching, whose outcome is first described here.

The field of sequence duplication dates back at least to the early twentieth century [6]–[8]; indeed observations of gene duplications among certain subsets of duplicated sequences have been characterized intensively [9]. Length distributions of duplications of or within genes have also been studied in the past — but one drawback to these studies is that the criteria for a sequence to be considered a gene have changed dramatically over recent years; the estimate of what fraction of a eukaryotic genome is functional jumped in the conventional wisdom from 2% to 6% in 2001 [10] and its future trajectory is unpredictable. In contrast to previous studies, our distributions were derived from entire genome sequences directly, without any preprocessing and irregardless of specific annotation. We did not restrict ourselves to genes, whose definition varies in time.

Rather than studying a ranked list, or Zipf-type frequency ranking such as the number of sequences with a given number of occurrences in a text, we studied the number of sequences duplicated within a single genome as a function of length. Length is a geometrical quantity, with a natural metric interpretation in terms of physical distance, for example in nanometres; as Mandelbrot observed in the 1950s, this geometric content distinguishes them fundamentally from ranked lists [11].

A set of conceptual tools for analyzing and understanding such geometry-based distributions was developed in the physical sciences starting in the middle of the twentieth century [12]. Recent popular guides to characterizing the form of distributions steer clear of any examples that are geometry-based, focusing instead entirely on ranked lists; consequently many of their recommendations for inferring the algebraic form for a distribution, are inapplicable in our context [13], [14]. In particular, physical sciences concepts stress that any algebraic form applies strictly only in the limit of large system size (here, genome length) — e.g. asymptotically in a thermodynamic or continuum limit. For finite system sizes that apply to real-world phenomena, any purely algebraic form is expected to represent an approximation at best; ultraviolet (short length, high energy) and infrared (large scale, low energy) corrections are inevitable. Nevertheless, because linearity on a log-log plot is necessary to infer algebraic form, but not sufficient [13], [14], our log-log plots always contain a semi-log inset of the same data.

In this paper, we propose and apply a characterization of the full distribution of duplication lengths within a single whole-genome sequence. The algebraic behavior exhibited by this distribution is investigated by an exponent/extent parameterization that is meant to quantify the range of lengths over which the algebraic form applies.

II. Results

Our computation identifies all 'maximal' repeats, forward or reverse-complement, within a chromosome or whole-genome sequence. A 'maximal' repeat is defined as a set of duplicated contiguous sub-sequences that is not contained within any other maximal repeat. For a set of sub-sequences of a given length L in nucleotides, the number of elements in the set, F, is reported as frequency as a function of L. For convenience this function, essentially the duplication length distribution, is referred to as the "length-frequency distribution, '$F(L)$, which denotes 'frequency as a function of length'. Note that $F(L)$ is a histogram and not a cumulative distribution.

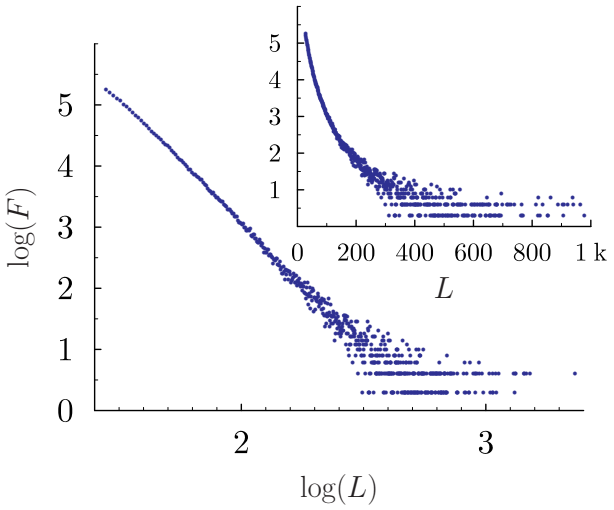


Fig. 1. Length-frequency distribution of the *Equus caballus* chromosome 3.

We examined $F(L)$ for different genomes and observed algebraic behavior over a large range of many of their distributions. As demonstrated in Figure 1 for a single chromosome of horse, the distribution behaves like a straight line on the log-log plot and exhibits strong curvature on the semi-log plot. (In this paper, the value $\log(x)$ stands for the base 10 logarithm of x.)

In order to quantify this observation, we characterize $F(L)$ for a given genome by two parameters under the assumption of an algebraic form for the distribution: (i) the *exponent* α

TABLE I
INFORMATION ON ANALYZED GENOMES

Taxonomy	Version	Coverage [×]
A. thaliana	ath1	n/a
C. elegans	ce2	n/a
C. milii	eshark1.4x	1.4
E. caballus	equCab2	6.8
H. sapiens	hg19	n/a
M. truncatula	Mt1.0	20
S. purpuratus	strPur2	6
T. guttata	taeGut1	6
T. truncatus	Ttru1.0	2.5

of a power-law function; and (ii) the *extent* E. The range of the distribution that best fits a power-law is determined by maximizing a specific cost function (details in section III). The exponent α and the extent E are calculated over this 'best fit' range.

Length-frequency distributions for whole genome sequences (WGS) or single chromosomes of *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Callorhinchus milii*, *Equus caballus*, *Homo sapiens*, *Medicago truncatula*, *Strongylocentrotus purpuratus*, *Taeniopygia guttata*, and *Tursiops truncatus*, were parameterized by α and E. Whenever they were unambiguously labeled in the genome assembly, *Un*placed sequences were deleted from the genome sequences used in the computation. Table I shows the assembly version and the reported coverage of the selected genomes.

The outcome is summarized in Figure 2, where the estimated exponent is plotted versus computed extent for WGS and separately for each chromosome. Inferred exponent/extent parameters are given in tables II, III, IV, V, VI, VII, and VIII, and, in figures 3, 4, 5, 6, and 7, where $F(L)$ plots are displayed for selected sequences together with fitted power-laws over estimated optimal ranges.

For human, horse, worm, barrel clover, and thale cress, both WGS and individual chromosomes are available, but only WGS distributions are displayed because the fit and the $F(L)$ of the chromosomes are nearly indistinguishable from what is obtained for their respective WGS. For purple sea urchin (Figure 5b), elephant shark (Figure 6a), and dolphin (Figure 6c), only WGS was available.

Some comments can be made based on these observations:

The extent of the fit: For most species, WGS yields a larger extent than chromosomes because the range of power-law behavior evidently increases with sequence length; however, there are some exceptions, such as zebra finch. A close look at $F(L)$ of zebra finch WGS sequences, in Figure 6b, and zebra finch chromosomes, in Figure 7, shows that for those sequences the fits were obtained only over a relatively small range.

The exponent of the fit: The plot shows that the exponent depends on the species, but is never less than around 2, and rarely greater than around 4.5

The shape of the distribution: As Figure 6 shows, in some cases oscillation (dolphin in 6c), folding (elephant shark in 6a), bending (zebra finch in 6b), cut off the straight-line regime in $F(L)$ for sufficiently large L. Often, the deviation is sharp, and begins very close to $L = 1000$. Indeed, inspection of the distribution of the contiguous runs of unambiguously defined bases (i.e. of A,G,C, or T but excluding N or X) within each genome assembly often reveals a sharp peak near $L = 1000$ followed by a rapid decay (E. Taillefer, J. Miller, manuscript in preparation). Such peaks are almost certainly artifacts of assembly. Obviously, exact duplications can only be subsequences of these contiguous runs.

Thus, the deviation from the power-law at large L can often be plausibly attributed to a cut-off intrinsic to the assembly. This cut-off may arise from *Un*placed or misplaced sequences; alternatively, unknown or poor-quality bases may break the genome sequence into a set of sparse short contigs instead of the single contiguous sequence that is believed to comprise each chromosome.

A prediction is that as the assemblies improve, the range of the good fit to a straight line on the log-log plot will increase indefinitely.

TABLE II
ESTIMATED α AND E FOR *Arabidopsis thaliana*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	3.130	4.59	3	3.09	3.99
1	3.07	3.91	4	3.01	4.04
2	3.15	3.77	5	2.95	4.12

TABLE III
ESTIMATED α AND E FOR *Caenorhabditis elegans*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	2.836	5.00	IV	2.70	3.70
I	3.09	3.95	V	2.87	3.94
II	2.62	3.44	X	2.76	3.56
III	2.85	3.48			

TABLE IV
ESTIMATED α AND E FOR *Equus caballus*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	3.639	6.60	17	4.01	4.62
1	4.031	4.90	18	3.878	4.94
2	3.955	5.10	19	4.03	4.80
3	3.937	5.17	20	4.05	4.89
4	3.831	4.76	21	4.00	4.78
5	3.814	5.03	22	4.19	4.18
6	4.06	4.91	23	3.3	1.95
7	3.790	5.07	24	4.11	4.28
8	3.866	5.00	25	3.90	4.44
9	3.974	5.00	26	4.23	4.24
10	3.840	5.01	27	4.22	4.35
11	4.02	4.55	28	4.20	4.22
12	3.86	4.48	29	4.20	4.40
13	3.88	4.44	30	4.11	4.41
14	3.870	5.11	31	4.31	3.93
15	3.95	5.20	X	3.878	5.36
16	3.903	4.95			

III. METHODS

Our analysis consists of the following sequence of computations:

1) Append to the downloaded sequence its reverse complement;
2) Compute the length-frequency distribution;

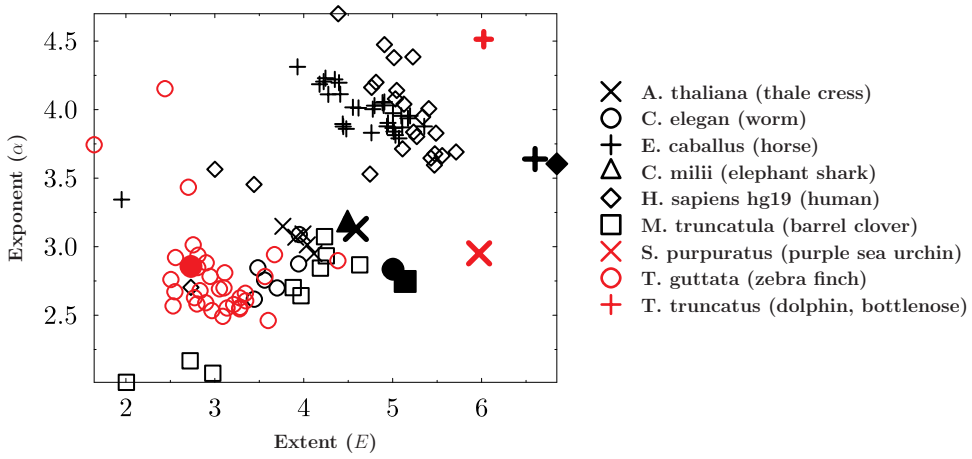
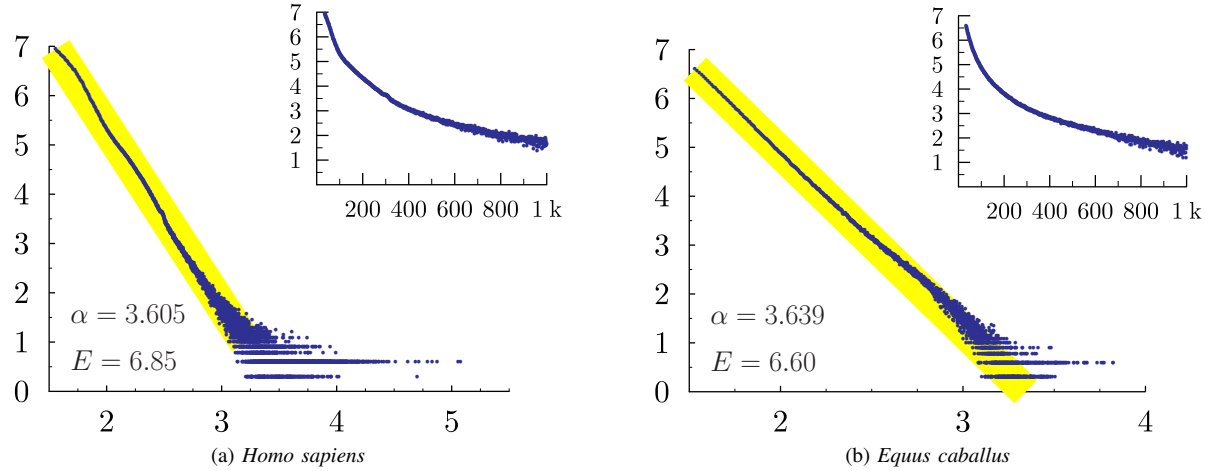
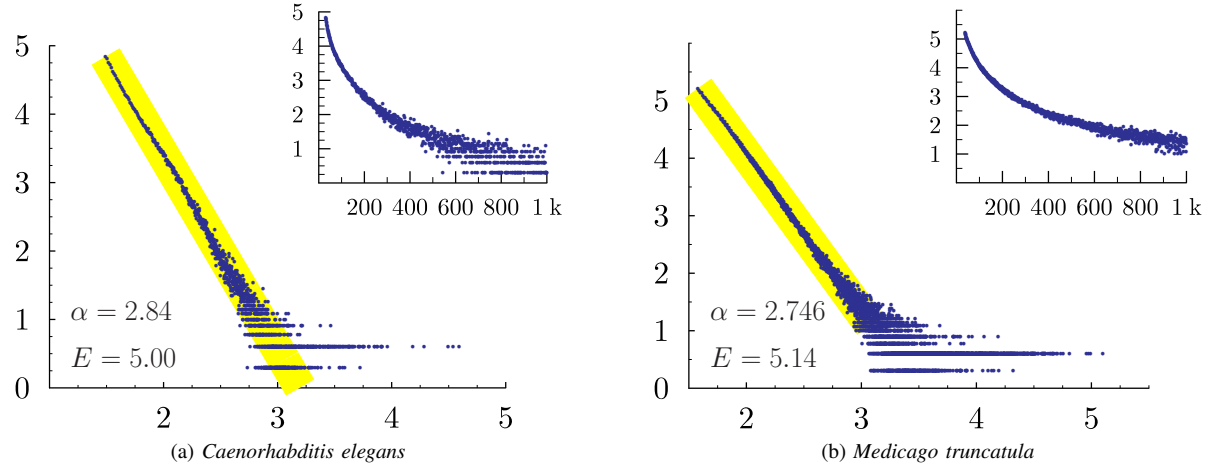


Fig. 2. Plot of exponent versus extent. Chromosomes are symbolized by thin lines or unfilled polygons; WGS by thick lines or filled polygons.



(a) *Homo sapiens* (b) *Equus caballus*

Fig. 3. Log-log plot of $F(L)$ for human and horse WGS. The power-law fit to the optimized range is plotted in thick yellow.



(a) *Caenorhabditis elegans* (b) *Medicago truncatula*

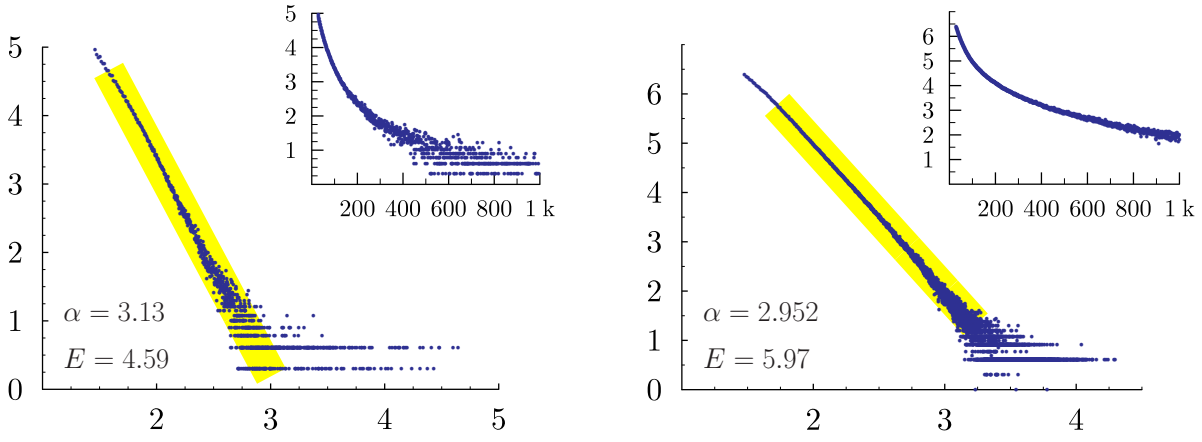
Fig. 4. Log-log plot of $F(L)$ for worm and barrel clover WGS.

3) Determine the length at which the 'non-random' region of $F(L)$ begins;
4) Estimate exponent and extent by minimizing a cost function.

More detail is given below.

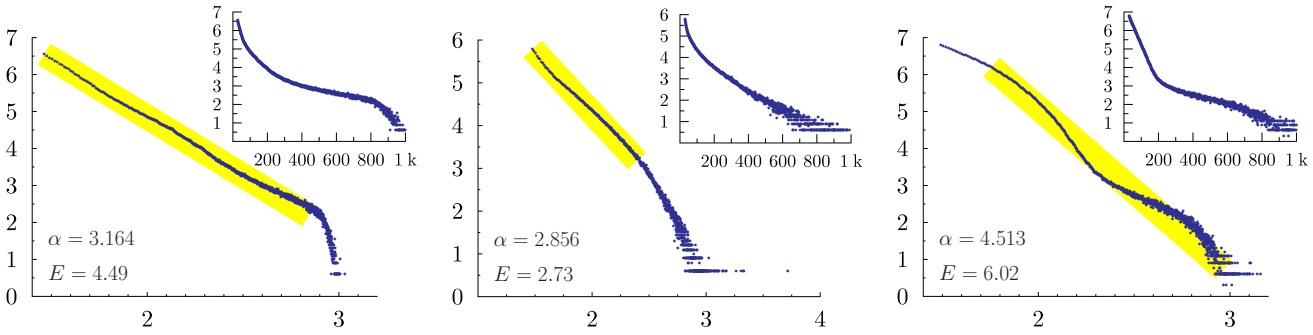
A. Compute the length-frequency distribution

We wrote a software package, SEQANALYSIS (E. Taillefer, J. Miller, submitted) based on an enhanced suffix array [15], to yield all maximal repeats in the DNA sequence and to compute $F(L)$.



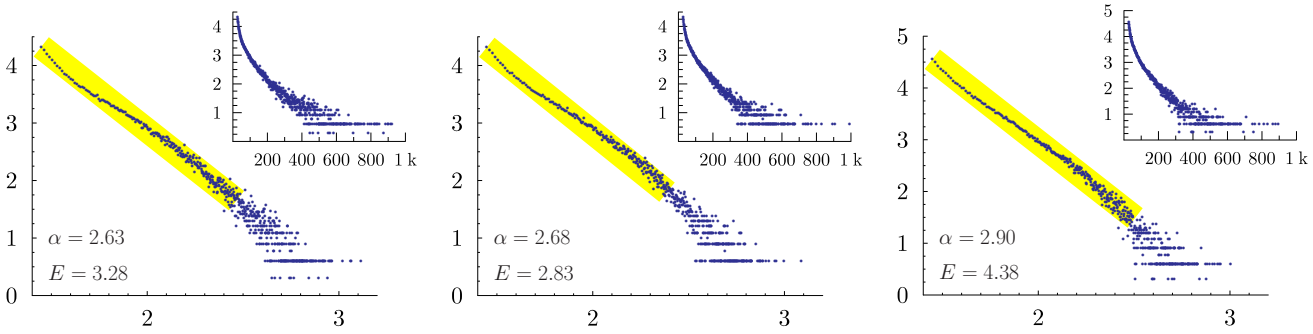
(a) *Arabidopsis thaliana* (b) *Strongylocentrotus purpuratus*

Fig. 5. Log-log plot of $F(L)$ for thale cress and purple sea urchin WGS.



(a) *Callorhinchus milii* (b) *Taeniopygia guttata* (c) *Tursiops truncatus*

Fig. 6. Log-log plot of $F(L)$ for elephant shark, zebra finch, and bottlenose dolphin WGS.



(a) Chr. 4 (b) Chr. 5 (c) Chr. Z

Fig. 7. Log-log plot of $F(L)$ of selected chromosomes of zebra finch.

B. Approximating the beginning of the 'non-random' region.

As shown in figures 8 and 9, on a log-log plot, comparison of the distributions for chromosome and for a 'random' sequence generated by a Markov model exhibits a regime for small L where the two distributions superpose, and a near-straight line regime for large L, where the distributions diverge from one another. In this regime, the distribution for the random sequence falls off too sharply for its curvature show up on a log-log plot, but zoomed in on a semi-log plot its exponential form is readily apparent. The beginning of the 'non-random' region, D_{start}, is approximated by the length at which the contribution of this exponential becomes negligible.

C. Estimation of the exponent and extent parameters.

The points of length-frequency distribution used for estimation are the ordered pairs (L_i, F_i) from the 'non-random' distribution region, where $i = D_{\text{start}}, D_{\text{start}} + 1, \ldots, n$, and, $L_{D_{\text{start}}} < L_{D_{\text{start}}+1} < \cdots < L_n$.

To estimate exponent and extent parameters, we maximize the cost function $\mathrm{C_f}$ over all possible ranges $[k, m]$, $D_{\text{start}} \leq k < m \leq n$. The cost function is defined as

$$\mathrm{C_f} = w^2 R^2 + \overline{E}^2, \tag{1}$$

TABLE V
ESTIMATED α AND E FOR *Homo sapiens*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	3.6047	6.85	13	4.20	4.81
1	3.691	5.71	14	4.040	5.13
2	3.828	5.49	15	3.530	4.74
3	4.007	5.41	16	3.595	5.47
4	3.804	5.27	17	3.947	5.34
5	3.646	5.43	18	4.16	4.76
6	4.140	5.05	19	4.385	5.23
7	3.664	5.56	20	4.47	4.91
8	3.836	5.24	21	4.70	4.39
9	3.454	3.44	22	2.70	2.73
10	3.678	5.47	X	3.714	5.11
11	4.078	5.03	Y	3.564	3.00
12	4.379	5.02			

TABLE VI
ESTIMATED α AND E FOR *Medicago truncatula*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	2.746	5.14	4	2.17	2.72
0	2.64	3.97	5	3.07	4.23
1	2.08	2.98	6	2.70	3.88
2	2.84	4.19	7	2.93	4.25
3	2.87	4.63	8	2.01	2.01

TABLE VII
ESTIMATED α AND E FOR *Taeniopygia guttata*.

Chr./Seq.	α	E	Chr./Seq.	α	E
WG	2.856	2.73	14	2.78	3.56
1	4.2	2.44	15	2.78	2.95
1A	2.70	3.10	16	3.7	1.64
1B	2.67	2.55	17	2.92	2.56
2	3.43	2.70	18	2.59	2.90
3	2.94	3.67	19	2.94	2.81
4	2.63	3.28	20	3.01	2.76
4A	2.53	2.97	21	2.56	3.28
5	2.68	2.83	22	2.81	3.11
6	2.85	2.81	23	2.55	3.14
7	2.63	2.77	24	2.57	2.53
8	2.58	2.80	25	2.76	2.51
9	2.88	2.90	26	2.66	3.34
10	2.58	3.21	27	2.49	3.09
11	2.69	3.05	28	2.61	3.35
12	2.46	3.60	Z	2.90	4.38
13	2.55	3.28			

TABLE VIII
ESTIMATED α AND E FOR *Callorhinchus milii*, *Strongylocentrotus purpuratus*, AND *Tursiops truncatus*.

Species	α	E
Callorhinchus milii	3.1643	4.49
Strongylocentrotus purpuratus	2.952	5.97
Tursiops truncatus	4.513	6.02

where R is the coefficient-of-determination (CoD) computed from the complementary cumulative distribution function (CCDF) of $F(L)$, $\overline{E}$ is the extent normalized over the non-random region, and the coefficient w controls the weight given to the CoD. To select the best CoD among the large extents, we took $w > 1$. Moreover, to avoid selecting trivial ranges, we also defined a minimum threshold (here 0.90) on the CoD.

The CCDF curve (L_i, Y_i) is obtained from the length-frequency distribution using the formula

$$Y_i = \sum_{j=i}^{n} F_j \Big/ \sum_{j=1}^{n} F_j \tag{2}$$

Because the CCDF features a smoother variation than the original distribution [13], [14], it is used to evaluate the CoD.

The CoD (R), which allows us to assess how well the range $[k, m]$ may be fit by a power-law, is computed on the CCDF data as follows

$$R = 1 - \left(\sum_{i=k}^{m} (y_i - f_i)^2 \Big/ \sum_{i=k}^{m} (y_i - \overline{y})^2 \right) \tag{3}$$

where

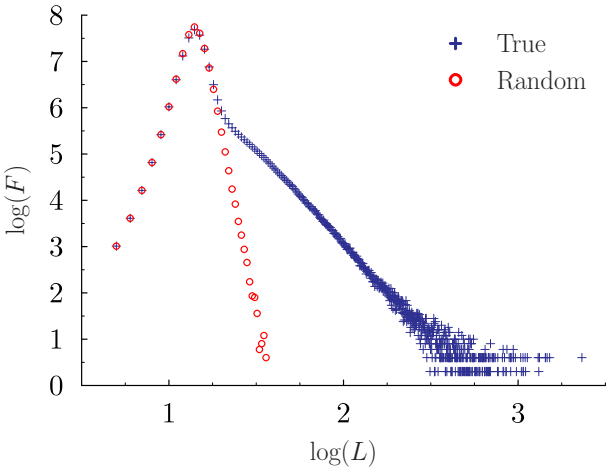


Fig. 8. Comparison of the $F(L)$ distribution of *Equus caballus* chromosome 3 with the $F(L)$ distribution of a 'random' sequence of the same length generated by a first-order Markov model with transition probabilities computed from *Equus caballus* chromosome 3.

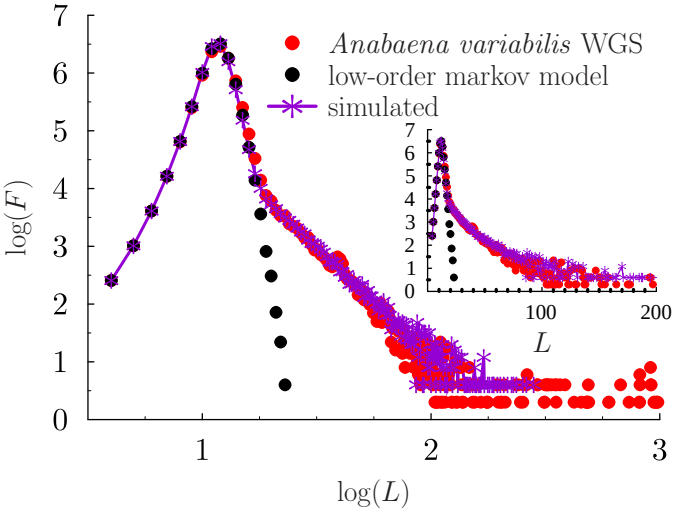


Fig. 9. Comparison between, the $F(L)$ distribution of *Anabaena variabilis*, the $F(L)$ distribution of a 'random' sequence of the same length generated by a first-order Markov model with transition probabilities computed from the cyanobacterium *Anabaena variabilis*, and the steady-state of a model of sequence duplication (M. Koroteev and J. Miller, submitted).

$$\overline{y} = \frac{1}{m-k+1}\sum_{i=k}^{m} y_i \qquad \begin{matrix} x_i = \log(L_i) \\ y_i = \log(Y_i) \end{matrix}$$

$$f_i = -(\widehat{\alpha}-1)(x_i - x_k)$$

and the estimate $\widehat{\alpha}$ of the exponent is obtained using the maximum likelihood estimator, expressed as

$$\widehat{\alpha} \simeq 1 + \left(\sum_{i=k}^{m} F_i\right)\left[\sum_{i=k}^{m} F_i \ln \frac{L_i}{L_k - \frac{1}{2}}\right]^{-1}. \qquad (4)$$

In (4), the function $\ln\{x\}$ is the natural logarithm (logarithm to the base e) of x.

The extent E is calculated as follows:

$$E = \sqrt{\log\left(\frac{L_m}{L_k}\right)^2 + \log\left(\frac{F_k}{F_m}\right)^2}, \qquad (5)$$

is the length, on logarithmic scale, of the line segment spanning the interval range $[k, m]$. The normalized extent $\overline{E}$ is obtained by

$$\overline{E} = \frac{E}{\sqrt{\log\left(\frac{L_n}{L_{D_{\text{start}}}}\right)^2 + \log\left(\frac{F_{D_{\text{start}}}}{F_n}\right)^2}}. \qquad (6)$$

IV. Conclusions

The consistent algebraic behavior of length-frequency distribution is a remarkable observation emerging from our computation, which is complementary to our self-alignment studies [5]. The k-mer methods applied here permit rigorous and exhaustive identification of all exact duplications, irrespective of whether they are identified by intrinsically heuristic self-alignment methods. Alignments, on the other hand, permit access to regimes of inexact matching that are out of reach of k-mer methods. Nevertheless, because alignments are heuristic, ad-hoc algorithms, their outcome can be *a priori* suspect unless it is confirmed by an independent route. This paper demonstrates that the distributions yielded by alignment are not artifacts of the alignment algorithm, and greatly extends the range and variety of genomes for which exact duplication length distributions have been computed.

Our analysis addressed the question: Over what range is length-frequency distribution most likely to be close to a power law? We proposed and implemented a cost-function method allowing selection of the 'best' extent.

The analysis naturally raises other questions: How does the quality of the assembly affect the distribution? Could the shape of length-frequency distribution represent and effective measure of assembly quality? What is the rate per cell division at which these duplications are created or destroyed? What evolutionary forces shape the distribution?

It seems readily apparent that duplication involves the copying of DNA, in contiguous blocks or chunks of linked sequence, from one location on a chromosome to another location not necessarily on the same chromosome. This process is a form of (not necessarily homologous) recombination, where we have taken the term 'recombination' to encompass processes of gene conversion and DNA repair as well.

The algebraic form of the empirical distribution of duplication lengths characterized here suggests that recombination events might as a general rule involve transfer of chunks of sequence with an algebraic length distribution, and that such events contribute significantly to neutral evolution of natural genomes. Their rates and origins remain to be determined.

References

[1] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, J. W. Kent, J. S. Mattick, and D. Haussler, "Ultraconserved elements in the human genome," *Science*, vol. 304, no. 5675, pp. 1321–1325, May 2004.

[2] S. Katzman, A. D. Kern, G. Bejerano, G. Fewell, L. Fulton, R. K. Wilson, S. R. Salama, and D. Haussler, "Human genome ultraconserved elements are ultraselected," *Science*, vol. 317, no. 5840, p. 915, Aug. 2007.

[3] W. Salerno, P. Havlak, and J. Miller, "Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments," *Proceedings of the National Academy of Sciences*, vol. 103, no. 35, pp. 13 121–13 125, Aug. 2006.

[4] J. Miller, "Colossal ultraconservation and super-colossal ultraconservation," *IPSJ SIG Technical Report*, vol. 2009-BIO-17, no. 7, pp. 1–8, May 2009.

[5] K. Gao and J. Miller, "Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments," *PLoS ONE*, in-press.

[6] A. H. Sturtevant, "Genetic factors affecting the strength of linkage in drosophila," *Proceedings of the National Academy of Sciences*, vol. 3, no. 9, pp. 555–558, Sep. 1917.

[7] S. Ohno, *Evolution by gene duplication*. Berlin: Springer-Verlag, 1970, ISBN 0-04-575015-7.

[8] T. Ohta, "Further simulation studies on evolution by gene duplication," *Evolution*, vol. 42, no. 2, pp. 375–386, Mar. 1988.

[9] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler, "Recent segmental duplications in the human genome," *Science*, vol. 297, no. 5583, pp. 1003–1007, Aug. 2002.

[10] R. H. Waterston, K. Lindblad-Toh, E. B. Birney, J. Rogers *et al.*, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–562, dec. 2002.

[11] B. B. Mandelbrot, *The Fractal Geometry of Nature*, 1st ed. New York: W. H. Freeman and Co., 1983.

[12] G. I. Barenblatt, *Scaling, self-similarity, and intermediate asymptotics*. Cambridge University Press, Dec. 1996, ISBN-13: 978-0-521-43522-2.

[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *Society for Industrial and Applied Mathematics Review*, vol. 51, no. 4, pp. 661–703, 2009.

[14] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, May 2005.

[15] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch, "Replacing suffix trees with enhanced suffix arrays," in *Journal of Discrete Algorithms*, ser. International Symposium on String Processing and Information Retrieval. Amsterdam, The Netherlands: Elsevier Science, Mar. 2004, vol. 2, pp. 53–84.