Deposited research article

Universality in large-scale structure of complete genomes

Li-Ching Hsieh[1,2], Ta-Yuan Chen[1], Chang-Heng Chang[1], Wen-Lang Fan[1] and Hoong-Chien Lee[1,2,3]

Addresses: [1]Department of Physics, [2]Department of Life Sciences and [3]Center for Complex Systems, National Central University, Chungli, Taiwan 320.

Correspondence: Hoong-Chien Lee. Email: hclee@phy.ncu.edu.tw

→ .deposited research

Universality in Large-Scale Structure of Complete Genomes

Li-Ching Hsieh[1,2]**, Ta-Yuan Chen**[1]**,Chang-Heng Chang**[1]**, Wen-Lang Fan**[1] **and Hoong-Chien Lee**[1−3]

Addresses: [1]Department of Physics and [2]Department of Life Sciences and [3]Center for Complex Systems, National Central University, Chungli, Taiwan 320

Correspondence: Hoong-Chien Lee. Email: hclee@phy.ncu.edu.tw

Abstract

The abundance of duplications in genomes in the form of paralogs, pseudogenes and a variety of repeats suggests that genomes may have used duplications as one mode for their growth. However a systematic knowledge on all possible duplications in whole genomes is still lacking. This paper reports the results of a detailed study of occurrence frequencies of short oligonucleotides in all extant complete genomes. We found a systematic pattern of repeats of short oligonucleotides that places all the complete genomes except *Plasmodium* in a single universality class expressed by an extremely simple formula. Our analysis of the data combined with computer simulation of genome growth models suggest a simple coarse-grain representation of genome growth: the ancestors of the genomes began to grow when they were no greater than 300 b in length via a mechanism whose main components were neutral stochastic segmental replicative translocations and random small mutations.

Key words: Complete genomes - Oligonucleotide frequency - Statistical analysis - Universality class - Genome growth model - Segmental duplication - Molecular evolution - Neutral theory of evolution

Introduction

There is abundant evidence suggesting that genomes used duplications as one mode for their growth: the existence of transposable elements and replicative translocation as a duplication mechanism; the large amounts of repeats in both prokaryotes [1] and eukaryotes [2, 3]; the preponderance of paralogs (genes) and pseudogenes in all life forms [4, 5]; chromosome segment exchanges that seem to characterize mammalian [6] and plant [7] radiations. There is also evidence suggesting that such a growth strategy may have the effect of enhancing the rate of evolution [8] and increasing the robustness of organisms [9]. This motivates the question: How pervasive were duplications in the formation of whole genomes? Because short oligonucleotides are least susceptible to alteration by mutations, we made a detailed study of occurrence frequencies of short oligonucleotides in all extant complete genomes in search for possible traces of duplication.

Occurrence frequencies of k-nucleotide words (k-mers) in a genome have been used for a variety of purposes including searches for statistical patterns in the distributions of short words [11, 12], attempts to construct a dictionary for biologically meaningful words [13, 14] and applications to studies in phylogeny and evolution [15, 16]. An enhanced abundance of frequencies far above or below the mean indicates potential biological significance. For each k we determine the frequencies by sliding a k-nucleotide-wide window one nucleotide at a time across the genome. We then use the complete set of occurrence frequencies to generate the set $\{n_f\}$, a *k-spectrum* with discrete frequencies, where n_f is the number of k-mers with frequency f. Here we focus on a particular k-spectrum property, namely its *reduced spectral width*, which measures (see below) its generalized spectral width relative to that of a corresponding random sequence and which is highly sensitive to the repeat content and randomness in the (genomic) sequence. The scope of our study covers all complete prokaryotic genomes (as of April 2003) and all chromosomes from complete eukaryotic genomes (July 2003) [17], for k=2 to 10. We stop at k=10 for two main reasons. One is statistical. The average occurrence frequency of 10-mers in a microbial genome of typical size (2 Mb; some eukaryotic chromosomes are much longer) is two, barely adequate for a study of variation in abundance. The other is biological. Duplications made at one time by whatever means - and not biologically functional - are susceptible to obliteration through later mutations, and shorter duplications have better chances of escaping such obliteration than longer ones.

Method

Figs. 1 (A) and (B) give a general overview of k-spectra for short words. Plots in black are the 5-spectra of two representative complete genomes with different percentages of (A+T) content or p, and plots in green show the 5-spectra of corresponding random sequences obtained by thoroughly scrambling the genomes (the spectra in orange are from sequences generated in a growth model, see below). Depending on the value of p, the random spectra are composed of one or more very narrow peaks while each of the genomic spectra essentially comprises a single, much wider distribution. The difference between (A) at $p=0.5$ and (B) at $p=0.7$ reflects the fact that the k-spectrum of an approximately compositionally self-complementary sequence (most genomes are such) is the superposition of $k+1$ subspectra. Each subspectrum corresponds to a subset (an m-set) of k-mers with m (A+T)'s, $m= 0$ to k, with mean frequency $\bar{f}_m(p)=\bar{f}2^k p^m(1-p)^{k-m}$, where $\bar{f}=4^{-k}L$ is the overall mean frequency of the k-spectrum(L is the sequence length). The subspectra are either narrow or broad and overlapping, respectively, if the sequence is random or genomic. They coalesce into a unimodal spectrum when $p=0.5$. Fig. 1 (C) focuses on detail from (B) showing only the subspectra of the $m=2$ set.
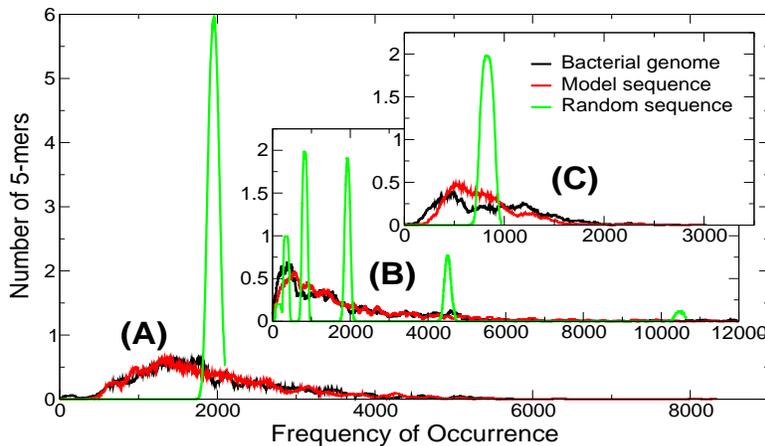

Figure 1: Frequency occurrence distributions of 5-mers, or 5-spectra, per 2 Mb length from two prokaryotes: (A) *A. fulgidus* (with (A+T) content $p=0.5$) and (B) *C. acetobutylicum* ($p=0.7$). Abscissa give occurrence frequency and ordinates give number of 5-mers; non-integer numbers occur as a result of averaging over a small span of frequencies to reduce fluctuation for better viewing. The black, green and orange curves represent distributions of the complete genomes, the randomized genome sequences and sequences generated in a model, respectively. See text for description of model. (C) focuses on details of the $m=2$ subspectra from (B).

We quantify the broadening of genomic subspectra relative to their random-sequence counterparts as follows. For each k-spectrum of a complete genome we define a reduced spectral width, $\mathcal{M}_\sigma$, as the weighted average - weighted by the number of k-mers in the m-set - of $(\Delta_m/\Delta'_m)^2$, where Δ_m and Δ'_m, respectively, are the half-width (standard deviation) and expected half-width of the m-set subspectra of the genome and corresponding random sequences. The subspectra of
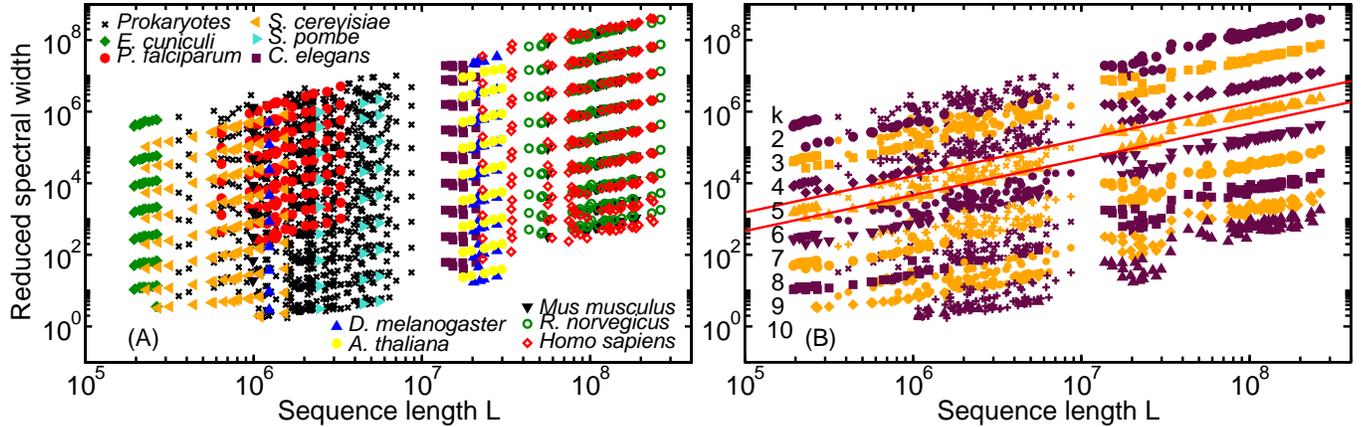
3

Figure 2: Reduced spectral widths $\mathcal{M}_\sigma$ versus sequence length L (in units of b), from 108 complete microbial genomes and 127 chromosomes of complete eukaryotic genomes. Each symbol is the $\mathcal{M}_\sigma$ value of one k-spectrum from one complete sequence. (A) $\mathcal{M}_\sigma$ color-coded by organism; (B) $\mathcal{M}_\sigma$ color-coded by k, where each "k-band" contains 221 pieces of data from 108 prokaryotic ($+$ and $\times$) and 113 eukaryotic (solid symbols; *Plasmodium* excluded) complete sequences. Data have been multiplied by a factor of 2^{10-k} to delineate the k-bands for better viewing and those for which $4^k > L$, when $\mathcal{M}_\sigma \approx 1$ regardless of sequence content, have been discarded. Straight red lines in the plots are $\mathcal{M}_\sigma \propto L$ lines.

the latter are given by known, slightly modified Poisson distributions [18]. When p approaches 0.5 (Fig. 1 (A)) $\mathcal{M}_\sigma$ simplifies to the square of the ratio of the half-widths of the genomic and random k-spectra.

Results

Fig. 2 shows the log-log plots of $\mathcal{M}_\sigma$ versus L for the k-spectrum, $k=2$ to 10, of 108 complete prokaryotic genomes (the prokaryotes) and 127 complete chromosomes from ten eukaryotic genomes (the eukaryotes). Each datum gives the $\mathcal{M}_\sigma$ value of one k-spectrum from a sequence. In Fig. 2 (A) the data are color-coded by organisms. Data from the three mammals, human (orange $\diamond$), mouse ($\blacktriangledown$) and rat (green $\circ$) are practically supervenient, showing that the present analysis is insensitive to whatever mutations, from large chromosomal segment exchanges to gene-modifying point mutations, which may have caused closely related organisms to diverge. Data from *Plasmodium* (red squares) are the exception in being more compact than all others in the vertical direction.

In Fig. 2 (B) where the 14 *Plasmodium* chromosomes are excluded, the data are color-coded after k. In spite of great disparity in length (0.2 Mb to 0.3 Bb) and base composition (p= 0.2 to 0.8) of the sequences, data for a given k from the 221 genome (108 prokaryotes and 113 eukaryotes) sequences form a narrow k-band that falls on a straight line indicating $\mathcal{M}_\sigma$ is proportional to L. For instance, data from the mammalian chromosomes and those from the thousand-fold shorter chromosomes of the single-celled parasite *E. cuniculi* are virtually collinear. Vertically the bands are about equally spaced indicating $\mathcal{M}_\sigma$ increases approximately geometrically with decreasing

k. On average the reduced spectral width of a 2-spectrum is about 1700 times greater than its 10-spectrum counterpart.
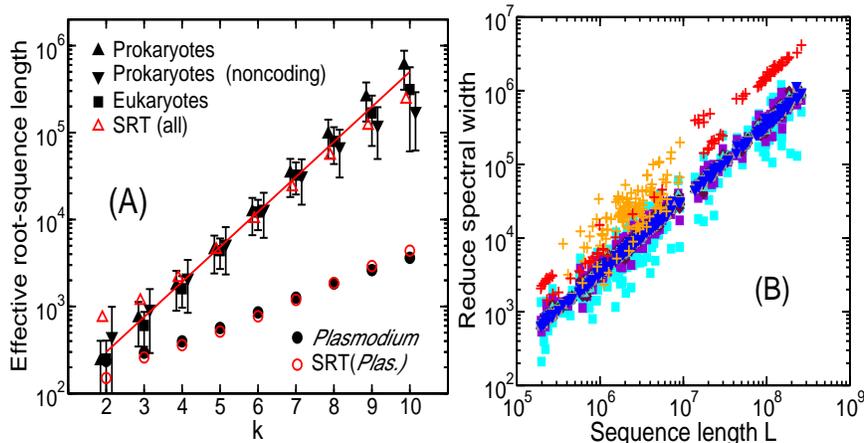

Figure 3: (A) Effective root-sequence lengths L_r versus k. Each piece of data gives $L/\mathcal{M}_\sigma$ averaged over a k-band (Fig. 2 (B)). Black symbols show genomic data (some with deliberate horizontal offsets for clarity): ▲, prokaryotes; ▼, noncoding regions in prokaryotes; ■ eukaryotes; ● *Plasmodium*. The red line gives the mean of the relation Eq. (1). Red symbols show results obtained from the SRT model (see text for description): △, for all organisms except *Plasmodium* (L_0=30, $1 \le l \le 1000$, R=78); ○, for *Plasmodium* (L_0=40, $1 \le l \le 1000$, R=40). (B) $\mathcal{M}_\sigma$ versus L for k-spectra, k=2 to 10, of 221 sequences in the control set (squares; k=2, cyan; k=3, purple; k=4-10, blue) and for 2-spectra (multiplied by factor 3) of complete prokaryotes (orange +) and eukaryoyes (red +).

The linearity of the k-bands implies that for given k the quantity $L_r(k)=L/\mathcal{M}_\sigma$, an *effective root-sequence length*, is an approximately sequence-independent universal constant. In Fig. 3 (A) the black symbols give values for $L_r(k)$ averaged over k-bands. The formula

$$\log L_r(k) = ak + B; \quad 2 \le k \le 10 \tag{1}$$

with lengths taken in units of b, summarizes the combined prokaryote (▲ in Fig. 3 (B)) and eukaryote (■) data and reduces the 1989 pieces of data in Fig. 2 (B) to two universal constants. The constants a and B have the smallest errors when k=6 is taken as the origin and the right-hand-side of Eq. (1) is re-written as $a(k-6) + B'$, then $B' \equiv \log L_r(6)$=4.05±0.19 ($L_r(6)$=12,200±5,000) and a=0.410±0.030. We refer to Eq. (1) (mean given by straight line in Fig. 3 (A)) as a universality class. The relation predicts the half-width Δ_m of an m-set of the k-spectrum of *any* genome of length L with base composition p belonging to the universality class to be $\Delta_m \approx 0.154 \times 1.60^{-k}\sqrt{L\bar{f}_m(p)}$, with an average error of about 30%, where $\bar{f}_m(p)$ is the mean frequency of the m-set. This yields, for instance, predicted half-widths of 910±270 and 590±180 for the genomic spectra in Fig. 1 (A) and (C), respectively, as compared to the measured half-widths of 1007 and 452, and of 44 and 29 for the half-widths from random sequences. Data for the fourteen chromosomes of *Plasmodium* (● in Fig. 3 (A)), the sole exceptions to this class, is also given by Eq. (1) but with B'=2.95±0.05 and a=0.146±0.012. The two classes in fact have a common $L_r(2)$: 240±160 b for the main class and

270±120 b for *Plasmodium*. Hence $L_r(2) \approx 250$ b may be referred to as a universal root-sequence length.

The genomes in the universality class span large ranges in sequence length (L) and percentage (A+T) content (p). In size the eukaryotes are more varied (0.2 to 300 Mb) than the prokaryotes (0.4 to 7 Mb). Compositionally the prokaryotes are more heterogeneous (p=28%-75%) than the eukaryotes (p=53%-64%). *Plasmodium* at p=81±1% is compositionally the exceptional eukaryote[10]. This alone would not explain why it should form its own universality class since all compositionally extremely biased prokaryotes - *U. urealyticum* and *B. aphidicola* at p=75% and *S. coelicolor* at 28% - are found in the dominant class.

On average about 85% of a prokaryote is comprised of coding regions, whereas most of an eukaryotic chromosome is noncoding (coding regions make up less than 2% of the human genome [2, 3]). The ▼ symbols in Fig. 3 (A) show the $L_r(k)$ for sequences obtained by concatenating the noncoding segments in prokaryotes. Both these and the eukaryote data (■) display an apparent leveling-off beginning at k=9. For the eukaryotes this probably reflects a real effect while for the concatenated sequences it is at least partly caused by their being shorter than 4^9. In any case, from the data shown in Fig. 3 (A), one may infer when the issue of sequence length is moot (*i.e.*, when $k \leq 8$) that no essential difference between coding and noncoding regions obtains. (Codons must be read in the open reading frame - one of six possible frames - and they seem not to make a major impact on the 3-spectra under study. When, however, a coding sequence is made by concatenating all the positively oriented protein-coding genes in a genome, and 3-spectra are constructed from the sequence by using a sliding window that moves *three* letters at a time - there are three such distinct spectra - then the $\mathcal{M}_\sigma$ of such a 3-spectrum is measurably greater than that of a 3-spectrum belonging to the universality class. This is expected since presumably genes have been subjected to more selective pressure, are conserved to a greater degree, and hence should be even less random than noncoding regions.)

Discussion

We now interpret the data. First we define an "n-replica" as the n-fold replication of a random root-sequence - it does not matter whether the replication is made algebraically, geometrically, or a combination of both. Because replication lengthens a sequence without increasing its randomness, we have a most noteworthy property of an n-replica: its reduced spectral width $\mathcal{M}_\sigma$ is n irrespective

of k. This allows one to understand the significance of the effective root-sequence length $L_r(k)$ as follows. The $\mathcal{M}_\sigma$ value of a k-spectrum of a genome of length L and base composition p is the same as that of an n-replica, $n=\mathcal{M}_\sigma$, of a *random* sequence of length $L_r(k)=L/\mathcal{M}_\sigma$ and base composition p.

We have constructed a control set of 221 sequences generated by replicating 300 b (approximate universal root-sequence length) root-sequences such that, corresponding to each complete sequence in the genome set, there is a control sequence having the same final length and base composition as the genome. The $\mathcal{M}_\sigma$ of the k-spectra, $k=2$ to 10, for all sequences in the control set were computed. The results shown in Fig. 3 (B) (squares) confirm expectation: for each sequence $\mathcal{M}_\sigma \approx L/300$ independent of p and k. Since it is very difficult to greatly increase the value of $\mathcal{M}_\sigma$ by any means other than replication/duplication and any randomizing action on the sequence would reduce $\mathcal{M}_\sigma$, a large value for $\mathcal{M}_\sigma$ suggests a correspondingly large amount of replication in the sequence. As well, for a set of sequences with different lengths, the relation $\mathcal{M}_\sigma \propto L$ suggests a common root-sequence length for the set.

Also shown in Fig. 3 (B) in orange (prokaryotes) and red (eukaryotes) "+" symbols is the $k=2$ band of genomic data seen in Fig. 2 (here boosted by a factor of 3 to separate it from the control-set band). The similarity between the two bands suggests a possible similarity in the large-scale structure of the two sets of sequences; however, the presence of a strong k-dependence in the genome data (Fig. 2 (B)) and its absence in the control-set data rule out the possibility that the complete genomes are simple replicas.

An obvious candidate that may account for the observed k-dependence are small mutations that partially randomize the sequences. For simplicity, we focus on random point replacement and consider its effect on a k-spectra of an n-replica. Let d be the average distance between two adjacent mutation sites. If the total number of mutations is very small, then with d much greater than the maximum k, the total effect of the mutations on the k-spectra will be neglegible. If d is of the order of k, then the mutation will affect the $(k+1)$-spectrum more than the k-spectrum, in both cases by reducing their respective $\mathcal{M}_\sigma$'s or, equivalently, increasing their L_r's. In the limit of a very large number of mutations, all traces of replication in the n-replica will be obliterated and $\mathcal{M}_\sigma$ will approach unity regardless of k as befits a random sequence. Presumably, given an n-replica, there may be an appropriate number of mutations whose effect is to generate a k-dependence in $\mathcal{M}_\sigma$ in

a way similar to that seen in Fig. 2.

Based on the above considerations and using maximum stochasticity and simplicity as guidelines we devised a minimal model for genome growth - the stochastic replicative transposition (SRT) model - in which an initial random sequence of length L_0 is grown to full length via duplications of randomly selected segments that are reinserted into the sequence at randomly selected sites[20]. The initial random sequence is given the base composition of the target genome sequence to be modelled and the lengths l of the duplicated segments have a simple prescribed distribution $g(l)$. After full growth the sequence is subjected to random point replacements - with a letter-bias reflecting the base composition of the target sequence - at a frequency of R mutations per 100 b sequence length. Having the mutations all occurring after the completion of growth does not necessarily reflect the actual workings of Nature in detail, rather is adopted in this paper to limit the complexity of the model. The genomic data do not uniquely determine the model parameters but impose constraints on them. The most important constraints are: (i) a small initial length satisfying $L_0 < L_r(2) \approx 300$ b to ensure $\mathcal{M}_\sigma$ for the smaller k's is sufficiently large and (ii) an appropriate amount of randomness in the sequence - generated by both the stochastic replication process and the mutations - to give $L_r(k)$ the correct k dependence.

Red symbols in Fig. 3 (A) show results in the form of L_r's computed from model sequences generated with $L_0=30$, $R=78$ and a $g(l)$ yielding equal non-zero probability to $1 \leq l \leq l_x$, where $l_x=1000$ if the (growing) sequence length is less than 2 Mb, and $l_x=5000$ otherwise; they represent the best results from the model after a non-exhaustive search in model space [18]. The model *Plasmodium* chromosomes are similarly generated except that $L_0=40$ and $R=40$ and the results are shown as red circles in Fig. 3 (A). That R for *Plasmodium* is significantly less than for genomes in the main class suggests that during its evolution *Plasmodium*, by comparison to all the other genomes studied, experienced as much duplication but significantly fewer point (or small) mutations per length. Among the eukaryotes studied *Arabidopsis*, which belongs to the main class, is phylogenetically the least remote from *Plasmodium* [10, 19]. It will be interesting to see how closer taxonomic relatives [19] of *Plasmodium* are classified by $\mathcal{M}_\sigma$. A general property of sequences generated by the model is - with regards to k-spectra - that a correct value for $\mathcal{M}_\sigma$ of a k-spectrum guarantees a correct shape for that spectrum (but not necessarily for the spectra of other k's). For instance, the good agreement between 5-spectra of model (orange curves) and genome (black) sequences seen in Fig. 1

8

is typical of the general case. That the SRT model can generate results that concord with data on $\mathcal{M}_\sigma$ systematically as well as with individual genomic k-spectra in detail gives us reason to believe that it may have captured the essence of genome growth. The simplicity of the model and the stochastic nature of the growth mechanisms may underlie the robustness of the results and explain the emergence of the universality classes. It is understood in our model that key elements, point mutation and replicative transposition, are representations of real mutation events - expected to be much more complex - and are subject to the usual rule of natural selection: only non-deleterious events can become fixations.

Our results suggest the following as a coarse-grain description of genome growth and evolution: very early on, when they were less than 300 b long, genomes started to grow mainly by stochastic replicative translocations followed by (or admixed with) small mutations at an accumulated frequency of slightly less than one replacement per base. Our study was applied to whole genomes and not just to coding regions. Whereas the complete genomes studied varied greatly in coding regions as a percentage of the whole genome (from 85% in microbes to less than 2% in the mammals), the universal genome property reported here seems not to depend on that percentage. If we assume that coded words other than genes such as binding sites, regulatory signals, and microRNA's [21] collectively do not occupy a dominant portion of the noncoding regions in eukaryotes, then our findings appear to imply that the majority of the individual fixed duplications and replacements during genome growth were selectively neutral. This notion of selective neutralism, based as it is on the present study's whole-genome analysis that is not sensitive to whatever effect selection had on the evolution of individual genes, seems to independently corroborates Kimura's neutral theory of molecular evolution [22, 23], a theory that was based on the investigation of polymorphisms of genes.

Acknowledgment

References

1. Jensen, L.J. *et al.* (1999) Three views of microbial genomes. Res. Microbiol. 150:773-777.

2. Lander E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

3. Venter J.C. *et al.* (2001) The sequence of the human genome. Science 291:1304-1351.

4. Otto S. & Yong P. (2001) The evolution of gene duplicates. Adv. Genetics 46:451-483.

5. Meyer A. (2003) Duplication, duplication. Nature 421:31-32.

6. O'Brien S.J. *et al.* (1999) The Promise of Comparative Genomics in Mammals. Science 286:458-481.

7. Grant D. *et al.* (2000) Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. PNAS 97:4168-4173.

8. Zhang Y-X. *et al.* (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature 415:644-646.

9. Gu Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

10. Gardner M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498-511.

11. Karlin S. *et al.* (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucl. Acids Res. 20:1363-1370.

12. Smith H.O. *et al.* (1995) Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. Science 269:538-540.

13. van Helden J., Andre B. & Collado-Vides J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281:827-842.

14. Bussemaker H.J., Li H. & Siggia E.D. (2000) Building A Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis. PNAS 97:10096-10100.

15. Karlin S. & Burge C. (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics 11:283-290.

16. Qian J., Luscombe N.M. & Gerstein M. (2001) Protein family and fold occurrence in genomes: power-law behavior and evolutionary. J. Mol. Biol. 313:673-681.

17. All complete sequences are taken from GenBank. Prokaryotes: www.ncbi.nlm.nih.gov /genomes/Complete.html (April 2003); Eukaryotes: .../genomes/static/euk_g.html (July 2003)

18. Hsieh L.C. *et al.*, in preparation.

19. Baldauf S.L. *et al.* (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-977.

20. Hsieh L.C. *et al.* (2003) Minimal model for genome evolution and growth. Phys. Rev. Lett. 90:018101-018104.

21. Ambros V. (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. Cell 113:673-676.

22. Kimura M. (1968) Evolutionary rate at the molecular level. Nature 217:624-626.

23. Kimura M. The neutral theory of molecular evolution. (Cambridge Univ. Press, 1983).