A Universal Signature in Whole Genomes

Li-Ching Hsieh[1,2], Ta-Yuan Chen[1], Chang-Heng Chang[1], and H.C. Lee[1−3*]

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

*To whom correspondence should be addressed E-mail: hclee@phy.ncu.edu.tw

Genomes are replete with duplicated sequences in the form of paralogs, transposons, pseudogenes, simple repeats, and others. To understand the origin of this phenomenon we did a systematic study of occurrence frequencies of short words in all extant complete genomes and found a common pattern of duplications in complete genomes so clear and pronounced that it allows all the genomes except one to be placed in a single class expressed by an extremely simple formula. Our analysis including extensive computer simulation in growth of DNA sequences shows that the formation of the class may be attributed to a universal genome growth mechanism in which maximally stochastic segmental duplication is the major mode of growth.

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? We try to answer this question by studying the occurrence frequencies of short words in all the publicly available complete genome sequences.

Occurrence frequencies of k-nucleotide words (k-mers) in a genome have been used, for instance, in studies in biological lexicon (*10-14*) and phylogeny and evolution (*12,15*). These studies are based on the well-founded assumption that a markedly overrepresented or underrepresented k-mer indicates potential biological significance. Here we do not pay attention to the abundance of individual k-mers. Rather we focus our attention on the pattern of the distribution of occurrence frequencies of an *entire* set of k-mers in a complete genome and compare it to that of a "random match" of the genome, namely, a random sequence having the same profile - length and base composition - as the genome.

From each genome (or chromosome) and for each k, we extract the set $\{n_f | f = 1, 2, \ldots\}_k$, where n_f is the number of k-mers occurring with frequency f in the genome. We call $\{n_f\}_k$, more specifically the plot n_f versus f a *k-spectrum*, in view of its analogy to a normal spectrum, and define the its relative spectral width (RlSW) as ratio of its half-width (i.e. standard deviation in frequency) to mean frequency. It will be seen that the genomic RlSW relative to that of its random match is highly sensitive to the amount of duplications in the genome. We examined the k-spectra of all complete genome sequences available from the GenBank (*16*) - 155 complete prokaryotic genomes (prokaryotes) and 127 complete chromosomes of eukaryotic genomes (eukaryotes) - for k=2 to 10. The complete sequences are highly diverse in profile (Fig. S1). 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 are susceptible to obliteration through later mutations (except those protected by being biologically functional), and shorter duplications have better chances of escaping such obliteration than longer ones.

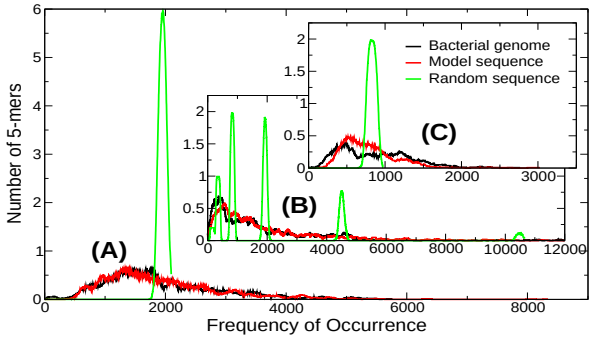

Figure 1: The 5-spectra (n_f *vs.* f), normalized to sequence length of 2 Mb, from two prokaryotes. For display (only) fluctuations in the spectra have been reduced by forward and backward average, which explains why n_f in the plots need not be an integer. (A) *Archaeoglobus fulgidus* (with (A+T) content p=0.5) and (B) *Clostridium acetobutylicum* (p=0.7). The black, green and orange curves are from the genome, its random match and a model sequence having the profile of the genome, respectively. See text for description. (C) focuses on the m=2 subspectra from (B).

Fig. 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 (length normalized to 2 Mb) and plots in green show the 5-spectra of corresponding random matches. In Fig. 1 (A), where the fractional (A+T) content, or p, of the sequences is 0.5, the shape of the random spectrum is the expected Poisson distribution with mean value equal to the mean frequency, in this case $2\times10^6/4^5$=1953, and half-width equal to the square-root of the mean, or 44. In great contrast, the half-width of the genomic

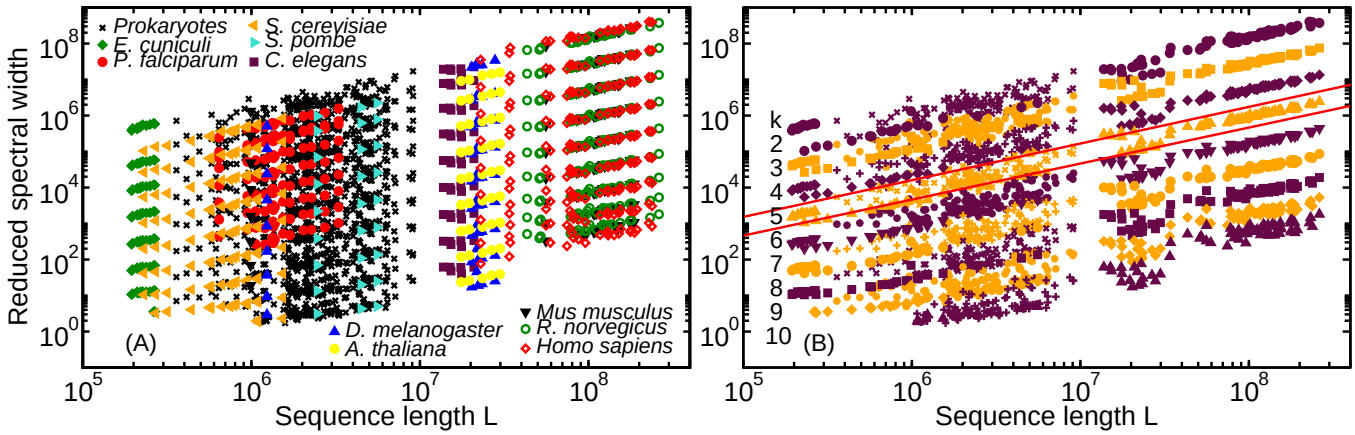

Figure 2: Reduced spectral widths, $\mathcal{M}_\sigma$, versus sequence length L (in units of b), from 155 prokaryotes and and 127 chromosomes of eukaryotes. Each symbol is the $\mathcal{M}_\sigma$ value of one k-spectrum from one complete sequence. (A) $\mathcal{M}_\sigma$ color-coded by organism; (B) $\mathcal{M}_\sigma$ color-coded by k to show that they form "k-bands" (k is the length of words the distribution of whose occurrence frequency is under study). Most bands contain 268 pieces of data from 155 prokaryotes (+ and ×) and 113 eukaryotes (solid symbols; *P. falciparum* excluded). Data have been multiplied by a factor of 2^{10-k} to delineate the k-bands for better viewing; those for which $L<4^k$ have been discarded. Straight red lines in (B) are $\mathcal{M}_\sigma \propto L$ lines.

5-spectrum is 1007, about 23 times the half-width of its random match. Heuristically, a genome with a broader k-spectrum has more over- and underrepresented k-mers and therefore a higher capacity for carrying more coded information. Thus the figures in Fig. 1 is consistent with the general expectation that a genome sequence would have vastly more information than its random match.

When a genome has a strong (A+T) bias the spectrum for the random match becomes visibly more complicated because subsets of k-mers each having a different fixed number of m (A+T)'s have different mean occurrence frequencies (*17,18*). Fig. 1 (B) shows the 5-spectra of sequences with p=0.7. The six peaks in the spectrum of the random sequence (green curve) correspond to the mean frequencies of the six m-sets, m= 0 to 5. The broadening of these narrow peaks in the genomic spectrum is seen in Fig. 1 (C), in which the subspectra corresponding to the m=2 set are isolated. Because of this broadening the six genomic subspectra overlap sufficiently to make an apparently unimodal overall 5-spectrum (black in Fig. 1 (B)). In such cases the crucial difference between a genome and its random match lies in the differences in the widths of the m-sets from the two sequences, but not in the overall spectral widths of the two sequence.

We quantify the broadening of the subspectra of a genome by defining, for each k-spectrum, a reduced spectral width (RdSW), $\mathcal{M}_\sigma$, a weighted average over m=0 to k of $(\sigma_m/\sigma'_m)^2$, where σ_m and σ'_m are the RlSWs of the m-set subspectra of the genome and those expected of its random match, respectively (*19*). This definition does not rely on the property of any actually simulated random sequence, but it does imply that $\mathcal{M}_\sigma \approx 1$ for any random sequence, which we have verified to be true with numerous simulations (see also below). Fig. 2 shows the log-log plots of $\mathcal{M}_\sigma$ versus sequence length L for the k-spectra, k=2 to 10, of the 282 complete sequences. Each datum gives the $\mathcal{M}_\sigma$ value for one k-spectrum from a genome/chromosome. In Fig. 2 (A) the data are color-coded by organisms. The prokaryotes are shown as black crosses. Data from the three mammals, human (orange ⋄), mouse (▼) and rat (green ○) are practically superimposed, 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 falciparum* (red bullets) are the exception in being more compact than all others in the vertical direction. In what follows we refer to the 14 *P. falciparum* chromosomes as PF, the prokaryotes as PK, the eukaryotes minus PF as EK and the combined PK and EK as CB.

In Fig. 2 (B) data for CB are color-coded after k. In spite of great diversity in length (0.2 to 300 Mb) and base composition (p= 0.28 to 0.78) of the sequences, for each k the 268 pieces of data form a narrow *k-band* indicative of a linear relation between $\mathcal{M}_\sigma$ and L (red lines in (B)). For instance, data from the mammalian chromosomes and those from the thousand-fold shorter chromosomes of the single-celled parasite *Encephalitozoon cuniculi* are virtually collinear. Vertically the k-bands are about equally spaced. On average the RdSW of a 2-spectrum is about 1500 times that of a 10-spectrum. The eukaryote PF is a parasite with an atypically compartmentalized genome (*20*). At p=0.81±0.01 its base composition is far more biased than the other genomes in EK, which have p=0.51 to 0.65. However, in CB there are many eukaryotes and prokaryotes that are either parasitic (or symbiotic) or compositionally extremely biased, or both (Table S3).

The linearity of the k-bands implies that for given k the quantity $L_r(k) \equiv L/\mathcal{M}_\sigma$ is an approximately genome-independent universal constant. In Fig. 3 (A) the black symbols give values for the $L_r(k)$'s, each averaged over a k-band (Table S1). The ▲ and ■ symbols are results for the PK and EK, respectively. 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) give the $L_r(k)$'s for sequences obtained by concatenating the noncoding segments of genomes in PK. These data show that no significant difference in $L_r(k)$ obtains either between coding and noncoding regions in PK or between PK and EK. The CB data, shown as ▲'s in Fig. 3 (B), are accurately given by

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

where L_r is in units of b, a=0.398±0.038 and B= 1.64±0.11. This simple formula reduces the more than 2400 pieces of data in Fig. 2 (B) to two universal constants. Because both the constants a and B are independent of genome sequences, in particular its profile L and p, we refer to Eq. (1) as a *universality class*. The mean $L_r(k)$ for the main class is given by the straight line in Fig. 3 (A). Recall that $\mathcal{M}_\sigma$ is essentially defined as the square of RlSW of the genome relative to that of its random match. Since the latter decreases linearly with sequence length, the constancy of $L_r(k)$ implies that the RlSW of genomes are universal constants independent of genome profile. We call $L_r(k)$ an *effective root-sequence length* (ERSL) because it is the length of a random sequence whose RlSW is equal to the universal genomic value. It follows from the multiplication law of probability that, if a random sequence of length λ=$L_r(k)$ is replicated L/λ times to yield a sequence, or a *replica*, of length L, then $\mathcal{M}_\sigma(k)\approx L/\lambda$ for all $k<<\lambda$ (and $4^k<L$). Note however that $L_r(k)$ has a strong k-dependence. In this context, as far as $\mathcal{M}_\sigma(2)$ is concerned, all genomes studied, PF excluded, are equivalent to replicas of ERSLs of length ∼300 b (Fig. 3 (A)); for $\mathcal{M}_\sigma(3)$ they are equivalent to replicas of matching ERSLs of length ∼680 b; and so on. Data for PF (● in Fig. 3 (A)), the sole exceptions to the main class, is also given by Eq. (1) but with a=0.146±0.012 and B=2.14±0.05. The reason for the anomaly, that $L_r(2)$ at ∼1800 is much greater than the value ∼270 dictated by the formula, is not completely understood (*19*).

The observed universality classes are unexpected. The existence of two classes is already proof that the existence of the observed universality classes is nontrivial. Namely, that genomes need not be in the same class. To further demonstrate the non-triviality of the observed universality classes we show, in green symbols in Fig. 3 (A), the $L_r(k)$'s of several sets of artificially generated sequences (Table S2). Because the $L_r(k)$ of a random sequence is approximately equal to its length, a set of random sequences of different lengths does *not* form a universality class. The ▽'s give $L_r(k)$'s computed from a set of 155 random sequences matching PK in profile. The large deviations in the data reflect the range of the lengths - 0.4 to 9.0 Mb - of the sequences in PK. A trivial universality class with a=0 and B=log λ is obtained by separately replicating to arbitrary lengths random sequences of a common (root-sequence) length λ. The □'s give $L_r(k)$'s computed from such a *replica set* of CB-matching sequences generated with λ=300 b. As expected, $L_r\approx$300 for all k's from every replica.

Although whole-genome replication has been suggested as a mode for genome growth (*21*), available evidence is against that mode being dominant (*22*). We devised a more biologically motivated genome growth model by substituting simple replication with stochastic segmental duplication (also know as replicative translocation), whereby a randomly selected segment of random length is duplicated and reinserted into the (growing) sequence at a randomly selected site. In the model the initial (random) sequence length is L_0, the average length of the duplicated segments is $l_x/2$ and, after being grown to full length, the sequence is subjected to random point substitutes at a rate of r mutations per base (*18*). Two main properties of the model are that L_0 is less than the smallest $L_r(k)$ and r controls the k-dependence of $L_r(k)$. Having the mutations all occur after the completion of growth does not necessarily reflect the actual workings of Nature in detail. The scheme chosen here is just the simplest representation for an infinite number of possible chronological combinations of duplication and mutation events. The green △ symbols in Fig. 3 (A), which give the results from a set of PK-matching model sequences generated with L_0=1000, l_x=500 and r=0.33, is an example of a nontrivial universality class that does not agree with data.

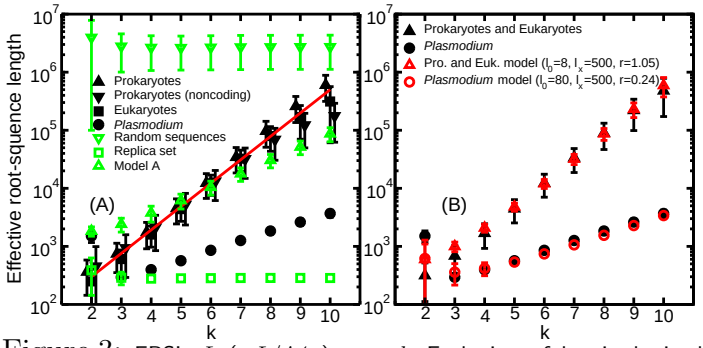

Figure 3: ERSLs $L_r(\equiv L/\mathcal{M}_\sigma)$ versus k. Each piece of data is obtained by averaging over a k-band. Black symbols give genome data, some with deliberate horizontal offsets for clarity. (A) Green symbols are results from artificially generated sequences: ▽, set of 155 random sequences matching PK in profile; □, set of 258 CB-matching replicas generated from 300 b random root-sequences; △ (model A), a non-genome-like universality class of CB-matching sequences. The red line gives the mean of the relation Eq. (1). (B) Red symbols give data from model sequence sets that emulate genome sequences: △, 255 CB-matching sequences generated with L_0=8, l_x=500, r=1.05; ○ 14 PF-matching sequences generated with L_0=80, l_x=500, r=0.24.

We have done χ^2 searches to find the optimum model parameters (Fig. S2). Red symbols in Fig. 3 (B) show results from a set of 258 CB-matching model sequences (△'s; L_0=8, l_x=500, r=1.05) and from a PF-matching set (○'s; L_0=80, l_x=500, r=0.24) (Table S4). Within the context of our model L_0 is constrained to be less than $L_r(2)$ and the optimum values of l_x and r are weakly anticorrelated and are constrained by the slope of $L_r(k)$ as a function of k. Within a limited range of r, a larger r or smaller l_x leads to a steeper slope. If r is too large then all model sequences are reduced to random sequences (green ▽'s in Fig. 3 (A)). The fact that model generates smaller deviations in the L_r's (red △'s in Fig. 3 (B)) then the genome data (black △'s in

Fig. 3 (B)) implies we could have generated the model sequence set using a range of parameter values centered around the optimum values without lowering the degree of agreement with empirical data. The maximally stochastic nature of the model renders it extremely robust and the results on $L_r(k)$ (but the model sequences themselves) highly reproduceable.

Concerning the k-spectra a general property of the model is that a correct value for L_r or relative spectral width guarantees a correct shape for that spectrum (*18*). The good agreement between 5-spectra of model (orange curves) and genome (black) sequences seen in Fig. 1 is typical of k-spectra for all k's and genomes.

Because the universal signature reported here is shared by coding and noncoding regions alike, the inference is that the majority of the individual fixed duplications and substitutes during genome growth were selectively neutral, which may be taken as an independent corroboration of Kimura's neutral theory of molecular evolution (*23,24*). The parameters in our genome growth model can be converted to nucleotide substitution (R_S) and duplication event (R_D) rates provided a growth period is assigned. For the purpose of this discussion we shall consider the case of *H. sapiens*, whose genome appears to have been still growing in the last 50 million years (My) (*25*), and take four billion years (By) to be its full life span. Then the model parameter r=1.0 translates to R_S=0.25/site/By averaged over the entire growth history. With the full length of the *H. sapiens* genome being 3 Bb, an average duplicated segment length of 600 b (value chosen for consistency with R_S, see below) translates into R_D=0.43/Mb/My.

Up to now all other estimates of molecular mutation rates have been extracted from the degree of nucleotide divergence of homologous sequences. Estimated silent site substitute rates for plants and animals range from 1 to 16 (/site/By) (*26*) and for humans it is R'_S=2±1/site/By (*25,27*). The animal gene duplication rate is estimated to be 0.01 (with a range of 0.002 to 0.02) per gene per My (*27*) which for *H. sapiens* (assuming coding region is 3% of genome) translates to 3.9/Mb/My. Human retrotransposition event rate is estimated to be about 2.8/Mb/My (average of 1.25, 1.29 and 6.0) (*25*). For comparison we will use the average value R'_D=3.4/Mb/My.

The rates R_S and R_D extracted from the growth model are both about 8 times less than the sequence alignment-based rates (R'_S and R'_D). However, the former pair of rates are averaged over the entire life span (T) of the human genome, while the latter are *terminal* values averaged over very recent times (about the last $\Delta T \approx 60$ My). Since the rates are given as number of events averaged over the entire length of a genome that grew in time, average rates are expected to be less than terminal rates. If we assume the rates per unit length per unit time are constant whereas the genome length grew exponentially with time as $L(t) = L_0 \exp(t/\tau)$, then with $R/R' \approx 1/8 << 1$ and $\Delta T/T \approx 0$, the terminal rate is essentially the constant rate and we have $\tau/T \approx R/R'$, which yields τ=0.50 By and L_0=1.0 Mb (*19*). These parameters imply the genome acquired the last 11% of its current length in the last 50 My, in reasonable agreement with the findings of Liu *et al.* (*25*). The value L_0=1.0 Mb should not be taken to mean literally that the genome was 1 Mb at the beginning of time, but should be taken to imply that at a time much earlier than 0.5 By after its birth the genome had already acquired a size of the order of 1 Mb.

The rate estimates given above are rough and are not supposed to be universal, but they suggest the possibility of a coherent picture of genome growth and evolution accommodating both results from the whole-genome analysis reported here and those from earlier alignment-based comparisons of sequence similarity. For a discussion of rates in prokaryotes our approach may need to be refined. Owing to their much shorter lengths it may be necessary to assume that the growth of prokaryotes had reached an equilibrium state in some time past. If this is so then a mechanism for keeping stable-sized genomes within the universality class will be needed. Whereas such genomes are robust against balanced combinations of random and sufficiently large segmental duplications and deletions, the characteristics of point mutations - substitutes, insertions and deletions - against which such genomes are robust remain to be determined.

References and Notes

1. Jensen, L.J. *et al.* Three views of microbial genomes. *Res. Microbiol.* **150**, 773-777 (1999).
2. Lander E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
3. Venter J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
4. Otto S. & Yong P. The evolution of gene duplicates. *Adv. Genetics* **46**, 451-483 (2001).
5. Meyer A. Duplication, duplication. *Nature* **421**, 31-32 (2003).
6. O'Brien S.J. *et al.* The Promise of Comparative Genomics in Mammals. *Science* **286**, 458-481 (1999).
7. Grant D. *et al.* Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *PNAS* **97**, 4168-4173 (2000).
8. Zhang Y-X. *et al.* Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415**, 644-646 (2002).
9. Gu Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-66 (2003).
10. Karlin S. & Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* **11** 283-290 (1995).
11. Smith H.O. *et al.* Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538-540 (1995).
12. Karlin S., Campbell, A.M. & Mrazek, J. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32** 185-225 (1998).
13. van Helden J., Andre B. & Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827-842 (1998).
14. Bussemaker H.J., Li H. & Siggia E.D. Building A Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis. *PNAS* **97**, 10096-10100 (2000).
15. Qian J., Luscombe N.M. & Gerstein M. Protein family and fold occurrence in genomes: power-law behavior and evolutionary. *J. Mol. Biol.* **313**, 673-681 (2001).
16. http://www.ncbi.nlm.nih.gov/genomes/Complete.html (2003/12/11) for the 155 prokaryotes, and .../genomes/static/euk_g.html (2003/05/31) for the 127 chromosomes of 10 eukaryotes.

17. Xie H.M. and Hao B.L. Visualization of K-Tuple Distribution in Procaryote Complete Genomes and Their Randomized Counterparts. *IEEE Proc. Comp. Sys. Bioinformatics*, 31-42 (2002).
18. Hsieh L.C. *et al.* Minimal model for genome evolution and growth. *Phys. Rev. Lett.* **90** 018101-018104 (2003).
19. See Supporting Online Material for detail.
20. Gardner M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).
21. Ohno S. Evolution by gene duplication. (Springer, New York, 1970).
22. Hughes A.L., da Silva J. & Friedman R. Ancient genome duplications did not structure the Human Hox-bearing chromosomes. *Genome Res.* **11**, 771-780 (2001).
23. Kimura M. Evolutionary rate at the molecular level. *Nature* **217**, 624-626 (1968).
24. Kimura M. The neutral theory of molecular evolution. (Cambridge Univ. Press, 1983).
25. Liu G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Gen. Res.* **13**, 358-368 (2003).
26. Li W.H. *Molecular Evolution.* (Sinauer Associates, 1997).
27. Lynch M. and Conery J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1152-1155 (2000).
28. This work is supported in part by grant no. 92-2119-M-008-012 from the National Science Council (ROC).

Supporting Online Material

Material and Methods
Figs. S1, S2
Tables S1 to S4
References
http://sansan.phy.ncu.edu.tw/∼hclee/ppr/hsieh_online04.pdf

June 16, 2004