The Origins of Genome Complexity
Michael Lynch, *et al.*
Science **302**, 1401 (2003);
DOI: 10.1126/science.1089370

17. Materials and methods are available as supporting material on *Science* Online and are derived from protocols at http://singerlab.aecom.yu.edu/protocols/insitu_yeast.htm.
18. J. P. O'Connor, C. L. Peebles, *Mol. Cell. Biol.* **11**, 425 (1991).
19. S. Sarkar, A. K. Hopper, *Mol. Biol. Cell* **11**, 3041 (1998).
20. F. Hediger, F. R. Neumann, G. Van Houwe, K. Dubrana, S. M. Gasser, *Curr. Biol.* **12**, 2076 (2002).
21. J. M. Huibregtse, D. R. Engelke, *Mol. Cell. Biol.* **4**, 3244 (1989).
22. P. Liljelund, S. Sariotte, J. M. Buhler, A. Sentenac, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9302 (1992).
23. M. R. Paule, R. J. White, *Nucleic Acids Res.* **28**, 1283 (2000).
24. P. L. Deininger, M. A. Batzer, *Genome Res.* **12**, 1455 (2002).
25. J. S. Smith, J. D. Boeke, *Genes Dev.* **11**, 241 (1997).
26. D. Donze, C. R. Adams, J. Rine, R. T. Kamakaka, *Genes Dev.* **13**, 698 (1999).
27. D. Donze, R. T. Kamakaka, *EMBO J.* **20**, 520 (2001).
28. The *Saccharomyces* Genome Database is available at http://genome-www.stanford.edu/Saccharomyces/.
29. We thank A. Hopper and members of the Engelke lab for experimental suggestions, and D. Thiele, E. Phisicky, and J. Abelson for helpful comments on the manuscript. Supported by research grant GM63142 from NIH.

Supporting Online Material
www.sciencemag.org/cgi/content/full/302/5649/1399/DC1
Materials and Methods
Figs. S1 to S3
References

30 July 2003; accepted 10 October 2003

The Origins of Genome Complexity

Michael Lynch[1]* and John S. Conery[2]

Complete genomic sequences from diverse phylogenetic lineages reveal notable increases in genome complexity from prokaryotes to multicellular eukaryotes. The changes include gradual increases in gene number, resulting from the retention of duplicate genes, and more abrupt increases in the abundance of spliceosomal introns and mobile genetic elements. We argue that many of these modifications emerged passively in response to the long-term population-size reductions that accompanied increases in organism size. According to this model, much of the restructuring of eukaryotic genomes was initiated by nonadaptive processes, and this in turn provided novel substrates for the secondary evolution of phenotypic complexity by natural selection. The enormous long-term effective population sizes of prokaryotes may impose a substantial barrier to the evolution of complex genomes and morphologies.

The ~100 fully sequenced eubacterial and archaeal genomes contain between 350 and 6000 genes, packed into 0.6 to 7.6 megabases (Mb) (*1*). Whereas some unicellular eukaryotes have genomes well within the range of these prokaryotes (such as 2000 genes in 2.9 Mb for the parasitic microsporidian *Encephalitozoon cuniculi*), all well-characterized genomes of multicellular animals and plants contain more than 13,000 genes in at least 100 Mb. The amount of DNA associated with just 30 human genes is equivalent to the entire genome size of an average prokaryote. Accompanying the increase in gene number in multicellular species is an expansion in the size and number of intragenic spacers (introns) and a dramatic proliferation of mobile genetic elements.

It remains unclear whether the expansions of genome size and complexity during eukaryotic evolution were essential for adaptive phenotypic diversification. After all, there are many ways to generate multiple functions from individual genes, such as tissue-specific gene regulation, alternative splicing, and RNA editing. In addition, the millions of mobile elements in the human genome and the massive increase in the average intron size in some multicellular eukaryotes have no obvious advantages. Finally, given that some prokaryotes are capable of cell differentiation, have linear chromosomes, and in rare cases have nuclear membranes, it is unclear whether the relatively simple genomes of microbes are merely reflections of unusual physiological constraints. Any general theory of genomic architecture evolution must account for the peculiar molecular attributes of various genetic elements, in addition to being compatible with the principles of population genetics. We argue here that the transitions from prokaryotes to unicellular eukaryotes to multicellular eukaryotes are associated with orders-of-magnitude reductions in population size; by magnifying the power of random genetic drift, reduced population size provides a permissive environment for the proliferation of various genomic features that would otherwise be eliminated by purifying selection.

Direct counts from multicellular and unicellular eukaryotes consistently show an inverse relationship between population density per unit of area and average individual body mass within a species (*2*–*5*). Such scaling need not reflect the pattern for total population size, given that it does not account for total species ranges. Moreover, the total abundance of a species need not reflect the more evolutionarily relevant genetic effective population size (N_e), which determines the degree to which gene frequencies are faithfully transmitted across generations. For example, a large population can behave genetically like a small one if a minor fraction of individuals contribute to the reproductive pool or if beneficial chromosomal segments periodically sweep through the population. Insight into long-term effective population sizes can be acquired from the nucleotide variation at silent sites in protein-coding genes (i.e., sites at which a nucleotide substitution leaves the encoded amino acid unchanged). The rate of introduction of new variation per site in two randomly compared alleles is $2u$ (twice the mutation rate per nucleotide), whereas the expected rate of loss of variation from neutral sites is $1/(2N_e)$ in a randomly mating diploid population. At equilibrium, the average number of nucleotide substitutions at neutral sites is $4N_eu$, with slight modifications required for other modes of inheritance (*1*). Thus, levels of silent-site variation among random alleles within a species provide an estimate of the composite parameter N_eu.

In a broad phylogenetic sense, there is an inverse relationship between organism size and N_eu. Proceeding from top to bottom of Fig. 1A, with two exceptions (*Streptococcus pyogenes* and *Pseudomonas aeruginosa*), all surveyed prokaryotes have $N_eu > 0.025$, whereas, with the exception of the malarial parasite *Plasmodium falciparum* and the ciliate *Tetrahymena thermophila*, the physically larger unicellular eukaryotes have $0.0035 < N_eu < 0.025$. For the still larger vascular plants and invertebrates, $0.00077 < N_eu < 0.0037$, whereas for vertebrates, $0.00027 < N_eu < 0.0010$. N_e can be disentangled from u by noting that the mutation rate per base per cell division ranges from 5×10^{-11} to 5×10^{-10}, with an average value of $\sim 2.3 \times 10^{-10}$ (*6*). This implies that N_e is generally greater than 10^8 for prokaryotes and often in the range of 10^7 to 10^8 for unicellular eukaryotes. The number of germline cell divisions per generation is ~10 in nematodes and ~25 in flies (*6*), implying that N_e is in the range of $\sim 10^5$ to 10^6 for invertebrates; the number of germline cell divi-

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA. [2]Department of Computer and Information Science, University of Oregon, Eugene, OR 97403, USA.

*To whom correspondence should be addressed. E-mail: mlynch@bio.indiana.edu

sions in vertebrates is ~100 (*6*), implying that N_e is on the order of 10^4 to 10^5.

These results probably underestimate the disparity in N_e among unicellular and multicellular species for two reasons. First, selectively driven codon bias can reduce silent-site variation below the neutral expectation, and any such bias would be greater for large populations, where selection is more efficient (*7*). Second, the majority of unicellular species in Fig. 1 are pathogens, and their genetic effective sizes may be highly influenced by that of their larger host, as in the case of the human malarial parasite *Plasmodium falciparum*. Thus, although the preceding calculations are approximations and the scaling of estimated N_e with actual population size is less than linear (*8*), the power of random genetic drift appears to vary by several orders of magnitude between the smallest unicellular and largest multicellular species.

The above estimates apply only to the past ~$4N_e$ generations for each taxon, whereas many of the gross features of genomes must have emerged over much longer time scales. However, the strong negative relationship between genome size and estimates of recent N_eu (Fig. 1B) is consistent with the idea that these estimates also reflect longer term conditions, with individual taxa experiencing temporal fluctuations around the predicted values. Moreover, the continuity of this relationship between prokaryotes and eukaryotes suggests that the cellular changes associated with the prokaryote-eukaryote transition are not major determinants of genome size and complexity.

The number of functioning genes within a genome reflects the long-term stochastic interplay between gene origin by various duplication mechanisms and gene loss by mutational silencing, which must be reflected in the smaller genomes of unicellular species relative to multicellular species. To estimate these rates, we have introduced evolutionary demographic techniques that use the divergence of silent sites as a relative measure of the age of a duplicate pair (*9*, *10*). The age distribution of all duplicates within a completely sequenced genome is typically L-shaped, suggestive of a steady-state stochastic birth-death process, from which the rate of birth and loss of duplicate genes can be estimated (*1*, *9*, *10*).

Although fairly large standard errors are associated with species-specific estimates, the average rates of gene duplication for unicellular and multicellular (metazoan) eukaryotes are not significantly different on the time scale of silent-site divergence (Table 1). Only downwardly biased estimates of the birth rates of prokaryotic genes can be obtained (*1*), but the averages based on 73 taxa are still ~50% of the values for eukaryotes. Thus, over a wide phylogenetic range, chromosomal events that result in gene duplications appear to occur at rates that are roughly proportional to those of mutations causing nucleotide substitutions, perhaps because both types of events reflect activities during replication.

In contrast, on the scale of silent-site divergence, duplicate genes are lost much more slowly in multicellular than in unicellular eukaryotes (Table 1), and there is a clear tendency for the half-life of duplicate genes to increase with genome size, again with a continuous transition between prokaryotes and eukaryotes (Fig. 2). Thus, by correlation, the ability of a newly arisen gene to survive the accumulation of mutations increases with decreasing effective population size. Because deleterious mutations are expected to accrue more easily in small populations, this counterintuitive result sheds some light on the processes that may be influencing the longevity of duplicate genes.

Preservation of both members of a duplicate pair can be promoted when one member of the pair acquires a beneficial mutation at the expense of an original essential function retained by the other (neofunctionalization)

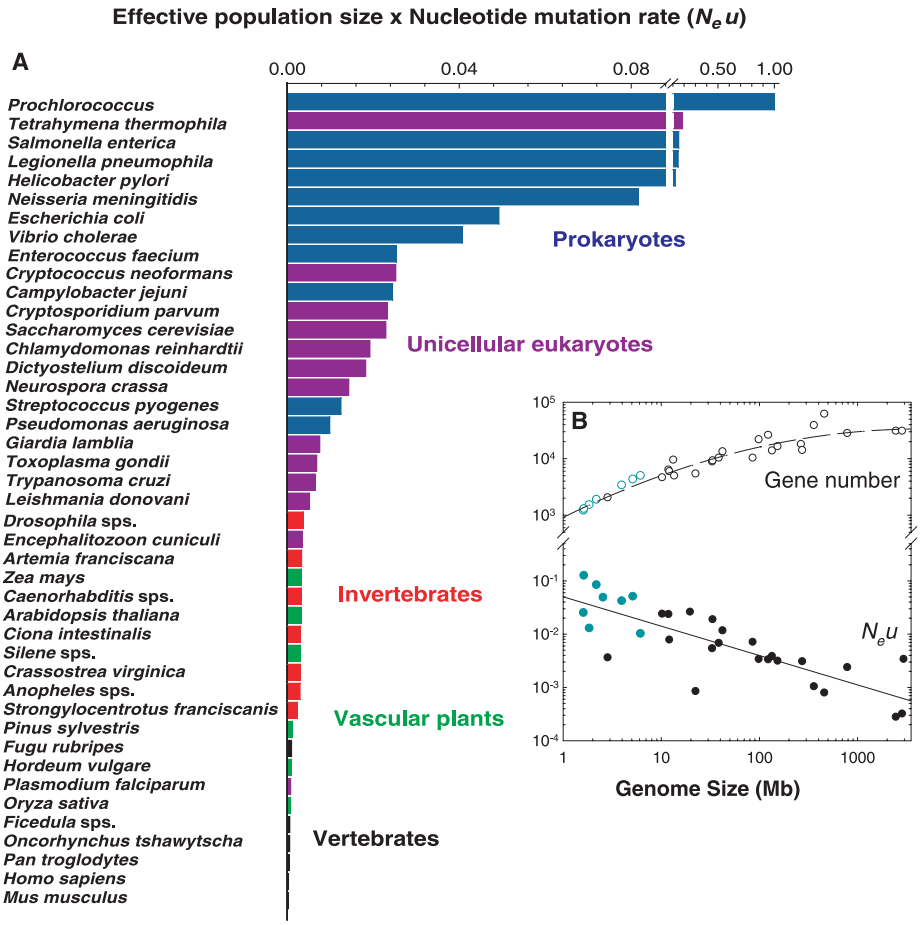

Fig. 1. (**A**) Estimates of the composite parameter N_eu for a phylogenetically diverse assemblage of species. (**B**) The relationship between estimated N_eu, total gene number, and genome size. Data for prokaryotes are plotted in blue. The log-log regression of N_eu versus genome size is highly significant, with an intercept of −1.30 ± 0.40, a slope of −0.55 ± 0.07, and $r^2 = 0.659$, df = 28 (*1*). The number of species plotted differs between graphs because genome structure information is not available for all species with N_eu estimates.

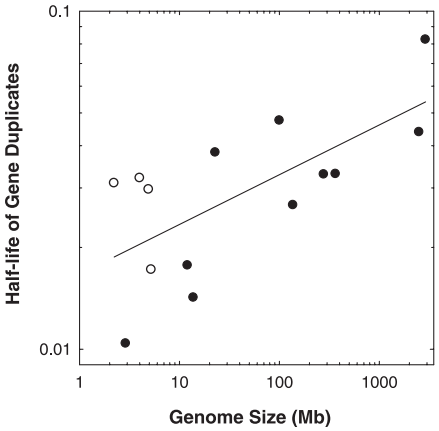

Fig. 2. The average half-lives of duplicate genes, defined as –ln(0.5)/*d*, in eukaryotic (solid circles) and prokaryotic (open circles) species on the time scale of divergence of silent substitutions. The log-log regression is significant at the 5% level, with an intercept of −1.76 ± 0.20, a slope of 0.20 ± 0.05, and $r^2 = 0.548$, df = 12 (*1*).

(*11*). Because degenerative mutations greatly outnumber beneficial mutations, the probability of preservation by rare neofunctionalizing mutations is diminished in small populations. In contrast, preservation by subfunctionalization occurs when both members of a pair are partially degraded by mutations to the extent that their joint expression is necessary to fulfill the essential functions of the ancestral locus (*12*, *13*). The probability of subfunctionalization approaches zero in large populations because the long time to fixation magnifies the chances that secondary mutations will completely incapacitate one copy before joint preservation is complete and because of the weak mutational disadvantage of harboring two coding regions (*14*). The longer retention time of duplicate genes in small populations is inconsistent with the predictions for the neofunctionalization model and opposite to the expected pattern if degenerative mutations only lead to complete nonfunctionalization of duplicate genes (*15*), but it is entirely compatible with expectations under the subfunctionalization model. Thus, although the evolution of multicellularity undoubtedly posed some new selective challenges that were met through neofunctionalization, much of the increase in gene number in multicellular species may not have been driven by adaptive processes, but rather as a passive response to a genetic environment (reduced population size) more conducive to duplicate-gene preservation by subfunctionalization.

Spliceosomal introns are noncoding stretches of RNA that are excised from the transcripts of their host protein-coding genes. The mechanisms by which introns originate remain a mystery, but their broad phylogenetic distribution implies that they and the spliceosome that processes them were present in the stem eukaryote (*16*). The average number of introns per gene in most multicellular species is between four and seven, whereas the average number for most unicellular eukaryotes is less than two. Only two spliceosomal introns have been found in the kinetoplastid *Trypanosoma* (*17*), and only a single one has been found in the diplomonad *Giardia* (*18*). Understanding this uneven phylogenetic distribution of introns is a major challenge for evolutionary genomics.

Although natural selection may eventually exploit introns for adaptive purposes (*16*), newly established introns are expected to impose a selective disadvantage (s) on their host genes by increasing the mutation rate to defective alleles (*19*). Theory suggests that there is a threshold value of $N_e s \approx 1.0$, below which newly arisen introns can freely drift to fixation and above which intron colonization and maintenance are exceedingly improbable. Qualitatively consistent with this hypothesis is a threshold genome size of ~10 Mb, below which introns are very rare and above which they approach an asymptote of about seven per gene (Fig. 3). By transforming scales from Fig. 1B, we found that the maximum value of $N_e u$ that is permissive to intron proliferation is ~0.015. How does this compare with the theoretical expectation of $N_e s \approx 1.0$?

The minimum selective disadvantage of an allele that contains a new intron is about equal to the excess-mutation rate to defective alleles caused by alterations at sites involved in splicing. The number of base positions (in the intron and surrounding exons) with nucleotide identities that are essential for proper splicing is unlikely to be less than 10 and is plausibly as high as 30 (*19*). Thus, the net selective disadvantage of an intron-containing allele is at least 10 to 30 times as large as u, not including insertion and deletion mutations, which minimally occur at ~10 to 60% of the rate of substitutions per base (*20*, *21*). Because they can alter the spatial configuration of key splice-site signatures, the number of insertion and deletion events affecting proper splicing must exceed that for substitutions. Thus, the observed threshold value of $N_e u \approx 0.015$ for intron proliferation is reasonably compatible with the theoretical $N_e s \approx 1.0$ threshold.

The rather abrupt increase in the average intron number per gene with increasing genome size is accompanied by a more continuous increase in the average intron length (Fig. 3), which has been observed previously in more phylogenetically restricted surveys (*22*, *23*). The inverse scaling of the average intron length with $N_e u$ [slope of the logarithmic regression (±SEM) on $N_e u = -0.67 \pm 0.22$] is consistent with the hypothesis that population-size reduction diminishes the efficiency of selection against mildly deleterious insertions into introns. Within genomes, the average intron size increases in regions of low recombination (*24*, *25*), which may also be a consequence of localized reductions in effective population size resulting from selective sweeps and/or background selection (*19*, *25*). An alternative hypothesis that intron size acts as a recombination modifier to reduce selective interference among linked sites (*24*) is not easily reconciled with the reduction of intron size and number in compact genomes.

Mobile genetic elements are self-contained genomic units capable of proliferating within their host genomes (*26*, *27*). Hundreds of families of these elements exist within eukaryotes, and almost all of them fit into three major functional categories: DNA-based (cut-and-paste) transposons and the long-terminal repeat (LTR) and non-LTR classes of RNA-dependent (copy-and-paste) retrotransposons. The vast majority of mobile elements are indiscriminate with respect to insertion sites, and as a consequence, their activities often have deleterious effects on the host genome. A broad range of selection coefficients must be associated with insertions in coding regions, regulatory regions, and intergenic spacers, and because mutations with negative fitness consequences $>>1/(2N_e)$ are efficiently eliminated by selection, the fraction of mobile-element insertions capable of drifting to fixation must decline with increasing N_e. Because mobile elements gradually acquire inactivating mutations, the long-term survival of an element family requires the average au-

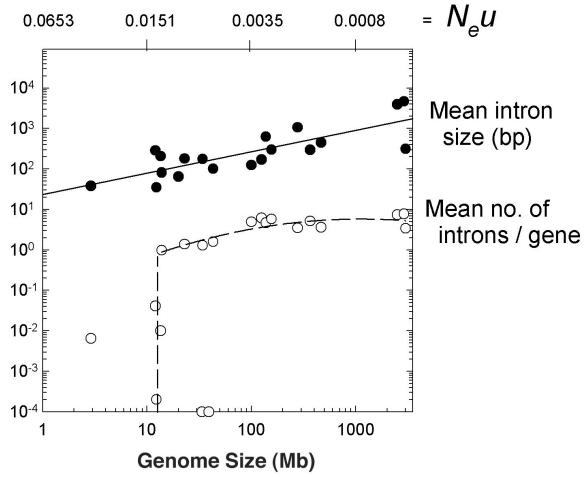

Fig. 3. The relationship between average intron size (solid circles) in base pairs (bp) and intron number (open circles) and genome size. The regression for intron size is highly significant, with an intercept of 1.41 ± 0.36, a slope of 0.51 ± 0.10, and $r^2 = 0.641$, df = 16 (*1*).

Table 1. Average rates of origin (B) and loss (d) of duplicate genes (±SEM). The former is defined as the probability of a gene duplicating over the time span required for a silent-site divergence of 1%. The latter is the exponential rate of loss, such that $D = 1 - e^{-(0.01d)}$, or ~0.01d for small d, is the probability of loss by the time silent sites have diverged by 1%, where e is the base of the natural logarithm (*1*). The analyses are based on gene families containing five or fewer members, and species-specific estimates can be found in the supporting online material (*1*).

Species	B	d
Unicellular eukaryotes	0.00405 ± 0.00130	43.26 ± 10.15
Metazoan species	0.00373 ± 0.00073	17.80 ± 2.52
Prokaryotes	0.00238 ± 0.00038	–

tonomous member to spawn at least one successful insertion in its lifetime. Therefore, there must be a critical value of N_e above which a mobile-element family cannot maintain itself within a host species.

Consistent with theoretical expectations, all three classes of mobile elements appear to have a threshold genome size below which mobile elements are unable to establish, an intermediate range in which only a fraction of species harbor them, and an upper threshold (~100 Mb) above which all species are infected (*1*) (Fig. 4). By extrapolation from the mutation rate cited above, the critical effective population size above which a unicellular eukaryote population appears to be immune to retrotransposon proliferation is $\sim 7 \times 10^7$, whereas for DNA-based transposons it is $\sim 2 \times 10^7$.

The influence of effective population size and mildly deleterious mutation on patterns of gene-sequence evolution has long been appreciated (*28*, *29*), and the preceding results suggest that these forces are central determinants of the types of genomic evolution that are permissible in various phylogenetic lineages. If this hypothesis is correct, then many of the genomic attributes of multicellular organisms did not arise in direct response to selection for new cell types and functions but were indirect consequences of reduced effective population sizes that accompanied an increase in organism size. Although the mechanisms responsible for the initial restructuring of eukaryotic genomes may have been nonadaptive in nature, this would not preclude the secondary deployment of the resultant genomic complexities in adaptive phenotypic evolution. For example, having colonized most protein-coding genes in some species, introns sustained a reliable mechanism for alternative splicing (*30*), and in at least some lineages, they provide an orientation mechanism for the surveillance of defective mRNAs (*16*). In addition, by converting single genes with multiple functions into multiple genes with fewer functions, subfunctionalization provides a mechanism for eliminating pleiotropic constraints on ancestral genes, thereby opening up previously inaccessible evolutionary pathways.

Because genomic infidelities associated with DNA replication are likely to generate observable genomic repatterning over a time scale that is on the order of tens of millions of years, a judicious use of experiments provided by nature will be necessary to test our hypothesis further. Although there is a general tendency for the genome sizes of multicellular species to exceed those of unicellular species, the range in genome size can be up to three orders of magnitude among species with similar levels of cellular and developmental complexity (*31*). In the very near future, we will experience an enormous proliferation of phylogenetically well-distributed genomic data, including those from unculturable organisms. The exceptional species within lineages should provide ideal substrate for testing the ideas outlined here. For example, one general prediction is that carnivores should exhibit the genomic hallmarks of population-size reduction compared with related herbivores. As a general rule, total biomass declines ~10% with increasing trophic level, and because average body size increases at higher levels in the food chain, total population size must decline even more sharply, which is consistent with the substantially lower estimates of N_e for carnivores than for herbivores derived from molecular surveys (*8*). If the theory that we present is correct, and should free-living prokaryotes with sufficiently small long-term N_e be found, we predict that they will harbor many of the same genomic changes that we have described here for eukaryotes.

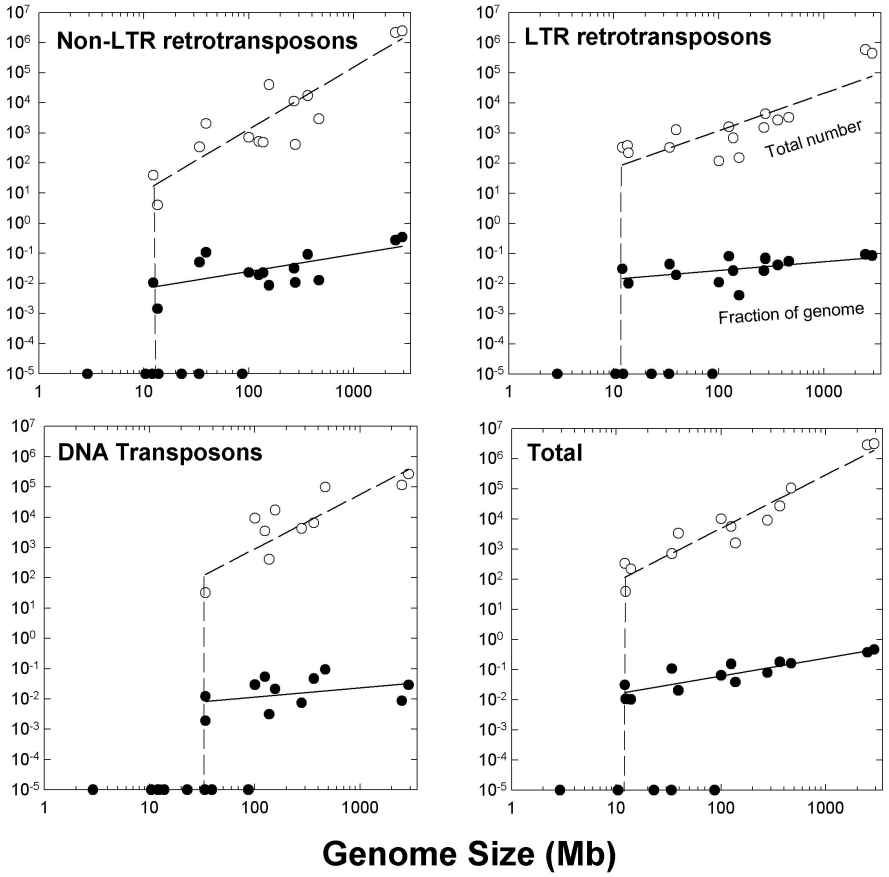

Fig. 4. Expansion of the three major classes of mobile genetic elements with genome size. Species for which the elements are entirely absent are plotted on the *x* axis but not included in the regressions.

References and Notes

1. Materials and methods are available as supporting material on *Science* Online.
2. P. E. Schmid, M. Tokeshi, J. M. Schmid-Araya, *Science* **289**, 1557 (2000).
3. B. J. Enquist, K. J. Niklas, *Nature* **410**, 655 (2001).
4. C. Carbone, J. L. Gittleman, *Science* **295**, 2273 (2002).
5. B. J. Finlay, *Science* **296**, 1061 (2002).
6. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, *Genetics* **148**, 1667 (1998).
7. H. Akashi, *Curr. Opin. Genet. Dev.* **11**, 660 (2001).
8. J. H. Gillespie, *The Causes of Molecular Evolution* (Oxford Univ. Press, New York, 1991).
9. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
10. M. Lynch, J. S. Conery, *J. Struct. Funct. Genomics* **3**, 35 (2003).
11. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Heidelberg, Germany, 1970).
12. A. Force *et al.*, *Genetics* **151**, 1531 (1999).
13. A. Stoltzfus, *J. Mol. Evol.* **49**, 169 (1999).
14. M. Lynch, M. O'Hely, B. Walsh, A. Force, *Genetics* **159**, 1789 (2001).
15. M. Lynch, A. Force, *Genetics* **154**, 459 (2000).
16. M. Lynch, A. O. Richardson, *Curr. Opin. Genet. Dev.* **12**, 701 (2002).
17. G. Mair *et al.*, *RNA* **6**, 163 (2000).
18. J. E. Nixon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3701 (2002).
19. M. Lynch, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6118 (2002).
20. D. A. Petrov, D. L. Hartl, *Mol. Biol. Evol.* **15**, 293 (1998).
21. D. R. Denver, K. Morris, M. Lynch, L. L. Vassilieva, W. K. Thomas, *Science* **289**, 2342 (2000).
22. A. E. Vinogradov, *J. Mol. Evol.* **49**, 376 (1999).
23. M. Deutsch, M. Long, *Nucleic Acids Res.* **27**, 3219 (1999).
24. J. M. Comeron, M. Kreitman, *Genetics* **156**, 1175 (2000).
25. A. B. Carvalho, A. G. Clark, *Nature* **401**, 344 (1999).
26. N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds., *Mobile DNA II* (Am. Soc. Microbiol. Press, Washington, DC, 2002).
27. B. Charlesworth, in *Population Genetics and Molecular Evolution*, T. Ohta, K. Aoki, Eds. (Japan Sci. Soc. Press, Tokyo, 1985), pp. 213–232.
28. T. Ohta, *Nature* **246**, 96 (1973).
29. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, UK, 1983).
30. B. R. Graveley, *Trends Genet.* **17**, 100 (2001).
31. T. R. Gregory, *Biol. Rev.* **76**, 65 (2001).

32. Supported by grants from the NIH and the NSF (to M.L.). We thank J. Gillespie, E. Koonin, and M. Wade for helpful comments.

Supporting Online Material

www.sciencemag.org/cgi/content/full/302/5649/1401/DC1
SOM Text
Table S1

18 July 2003; accepted 8 October 2003