The Evolutionary Fate and Consequences of Duplicate Genes

Michael Lynch, *et al.*

Science **290**, 1151 (2000);
DOI: 10.1126/science.290.5494.1151

$\varepsilon \sim N(0, i + jG)$. Although growth exhibits a slight decline with stand age in some cases, using alternative functional forms that allow for a decline do not significantly change the results presented here. We specify mortality as an exponential random variable with mean $M = \mu B$, where the mortality rate μ is a constant. Although the mortality rate exhibits a slight decline with stand age in some cases, using alternative functional forms that allow for a decline do not significantly change the results presented here (the stand mortality rate reflects mortality from various sources, including thinning, windthrow, fire, and selective harvesting, which may exhibit different trends with respect to stand age). We obtained maximum likelihood estimates of the growth and mortality parameters using a simulated annealing algorithm as we do in all subsequent analyses.

16. Given $G(A)$ and μ, we can calculate $B(A)$ for any value of $B(0)$. We estimate $B(0)$ as the value that provides the best fit to $B(A)_{obs}$. Here, we assume that $B(A)_{obs}$ is normally distributed with constant variance.
17. See (*12*).
18. In this paper we report results for linear forms of $h(t)$ and $f(t)$, although alternative forms give similar results.
19. Changes in mortality disproportionately affect old high-biomass stands, because the amount of biomass lost to mortality is small relative to growth in young low-biomass stands but not in old high-biomass stands. Thus, if mortality rates have decreased, current vital rates will closely predict the biomass of younger stands but the predicted biomass of older stands will exceed the observed biomass. In this case, a nonzero α will provide a better fit to $B(A)_{obs}$ than a nonzero β. On the other hand, if growth rates have increased, the predicted biomass will exceed the observed biomass in both young and old stands. In this case, a nonzero β will provide a better fit to $B(A)_{obs}$ than a nonzero α.
20. The estimates in Table 1 indicate that the rate of biomass accumulation has not increased in MN, MI, and FL. In VA and NC, there has been an increase in the rate of biomass accumulation, but we estimate that this increase is due to decreases in mortality rather than increases in growth. The disproportionate effect of changes in growth and mortality on old versus young stands allows us to partition increased accumulation between increases in growth and decreases in mortality. Yet, one can always conceive of complicated scenarios in which changes in mortality exactly balance changes in growth in both old stands and young stands, making partitioning impossible. For this reason, we also present independent evidence that confirms our conclusion that there has been a reduction in mortality in Virginia and North Carolina rather than an increase in growth. See (*12*).
21. These estimates are based on allometric equations used to estimate the aboveground biomass of trees. Thus, if N deposition, CO_2 fertilization, or climate change had a pronounced effect on tree allometry (by significantly increasing the ratio of root biomass to total tree biomass), then the fraction of total net ecosystem production (above and below ground) due to growth enhancement could be greater than the fraction of ANEP due to growth enhancement. Furthermore, growth enhancement may represent a small fraction of ANEP, because the effects of N deposition and CO_2 fertilization are balanced by the effects of other factors such as ozone and calcium depletion.
22. We calculate regional-level ANEP as the sum of the ANEP of natural stands that have not been clear-cut or otherwise heavily disturbed, and the ANEP of all the remaining plots, including clear-cut plots, plantations, and plots that changed from forest to nonforest and vice versa. For the remaining plots, we assumed zero biomass for plots classified as nonforest.
23. We obtain 95% confidence limits by calculating the fraction of ANEP due to growth enhancement using the range of parameter values for which the log-likelihood of the data is within 1.92 of the maximum log-likelihood (*17*). The highest and lowest fraction are the reported limits for the fraction of ANEP due to growth enhancement.
24. This research was conducted under the auspices of the Carbon Modeling Consortium (CMC), which is supported by the Office of Global Programs and the Geophysical Fluid Dynamics Laboratory of the NOAA. Support was also provided by the Andrew W. Mellon foundation. The advice and insights provided by our colleagues at the CMC and elsewhere are gratefully acknowledged, in particular E. Shevliakova, J. Sarmiento, and M. Gloor.

5 July 2000; accepted 29 September 2000

The Evolutionary Fate and Consequences of Duplicate Genes

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

Gene duplication has generally been viewed as a necessary source of material for the origin of evolutionary novelties, but it is unclear how often gene duplicates arise and how frequently they evolve new functions. Observations from the genomic databases for several eukaryotic species suggest that duplicate genes arise at a very high rate, on average 0.01 per gene per million years. Most duplicated genes experience a brief period of relaxed selection early in their history, with a moderate fraction of them evolving in an effectively neutral manner during this period. However, the vast majority of gene duplicates are silenced within a few million years, with the few survivors subsequently experiencing strong purifying selection. Although duplicate genes may only rarely evolve new functions, the stochastic silencing of such genes may play a significant role in the passive origin of new species.

Duplications of individual genes, chromosomal segments, or entire genomes have long been thought to be a primary source of material for the origin of evolutionary novelties, including new gene functions and expression patterns (*1–3*). However, it is unclear how duplicate genes successfully navigate an evolutionary trajectory from an initial state of complete redundancy, wherein one copy is likely to be expendable, to a stable situation in which both copies are maintained by natural selection. Nor is it clear how often these events occur.

Theory suggests three alternative outcomes in the evolution of duplicate genes: (i) one copy may simply become silenced by degenerative mutations (nonfunctionalization); (ii) one copy may acquire a novel, beneficial function and become preserved by natural selection, with the other copy retaining the original function (neofunctionalization); or (iii) both copies may become partially compromised by mutation accumulation to the point at which their total capacity is reduced to the level of the single-copy ancestral gene (subfunctionalization) (*1–12*). Because the vast majority of mutations affecting fitness are deleterious (*13*), and because gene duplicates are generally assumed to be functionally redundant at the time of origin, virtually all models predict that the usual fate of a duplicate-gene pair is the nonfunctionalization of one copy. The expected time that elapses before a gene is silenced is thought to be relatively short, on the order of the reciprocal of the null mutation rate per locus (a few million years or less), except in populations with enormous effective sizes (*11, 12*).

These theoretical expectations are only partially consistent with the limited data that we have on gene duplication. First, comparative studies of nucleotide sequences suggest that although both copies of a gene may often accumulate degenerative mutations at an accelerated rate following a duplication event, selection may not be relaxed completely (*14–16*). Second, the frequency of duplicate-gene preservation following ancient polyploidization events, often suggested to be in the neighborhood of 30 to 50% over periods of tens to hundreds of millions of years (*17–20*), is unexpectedly high.

Further insight into the rates of origin of duplicate genes and their evolutionary fates can now be acquired by using the genomic databases that have emerged for several species. We focused on nine taxa for which large numbers of protein-coding sequences are available through electronic databases: human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), nematode (*Caenorhabditis elegans*), fly (*Drosophila melanogaster*), the plants *Arabidopsis thaliana* and *Oryza sativa* (rice), and the yeast *Saccharomyces cerevisiae.* For each of these species, the complete set of available open reading frames was screened to eliminate se-

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

*To whom correspondence should be addressed. E-mail: mlynch@oregon.uoregon.edu

quences that were unlikely to be functional proteins (*21*). Each sequence retained after this initial filtering was then compared against all other members of the intraspecific set to identify pairs of gene duplicates, which were then analyzed for the degree of nucleotide divergence (*21*). The analyses for *C. elegans*, *D. melanogaster*, and *S. cerevisiae* were based on the complete genomic sequences available for these species.

The traditional approach to inferring the magnitude of selective constraint on protein evolution focuses on codons, comparing the rates of nucleotide substitution at replacement and silent sites (*7, 15, 16*). With this sort of analysis, only the cumulative pattern of nucleotide substitution is identified, making it difficult to determine whether duplicate genes typically undergo different phases of evolutionary divergence, e.g., an early phase of near neutrality followed by a later phase of selective constraint. Some clarification of this issue can be achieved by considering the features of sets of gene duplicates separated by an array of divergence times.

Under the assumption that silent substitutions are largely immune from selection and accumulate at a stochastic rate that is proportional to time, we take the number of substitutions per silent site, S, separating two members of a pair of duplicates to be a measure of the relative age of the pair. Letting R denote the number of substitutions per replacement site, a net (cumulative) selective constraint since the time of origin of a pair of duplicates will be reflected in an R/S ratio < 1, whereas a net acceleration of protein evolution will be revealed by an R/S ratio > 1. Complete relaxation of selection will result in $R/S \approx 1$. For the duplicate genes that we have identified, there is often considerable scatter around the neutral expectation when $S < 0.05$ (Fig. 1), suggesting that early in their history, many gene duplicates experience a phase of relaxed selection or even accelerated evolution at replacement sites. The progressive decline of R/S beyond this point reflects a gradual increase in the magnitude of selective constraint. The vast majority of gene duplicates with $S > 0.1$ exhibits an R/S ratio $\ll 1$.

From the qualitative behavior of the cumulative R/S ratio, some insight into the temporal development of increasing selective constraint on duplicate-gene evolution can be obtained by considering a simple model in which R declines relative to S, according to the function

$$\frac{dR}{dS} = \frac{1}{a - be^{-mS}} \quad (1)$$

Under this model, assuming positive m, the ratio of rates of replacement to silent substitutions initiates with an expected value of $1/(a - b)$ at $S = 0$ (reflecting the evolutionary properties of newly arisen duplicates) and declines to $1/a$ as $S \to \infty$ (reflecting ancient duplicates). Integrating this equation, the expected cumulative number of substitutions per replacement site (R) can be described as a function of the cumulative number of substitutions per silent site (S),

$$R = \frac{1}{am}\left[mS - \ln\left(\frac{a - b}{a - be^{-mS}}\right)\right] \quad (2)$$

The parameters a, b, and m can then be estimated by performing least-squares analysis on the pairwise gene-specific estimates of R and S (*22*).

Given the inherently stochastic nature of molecular evolutionary processes, Eq. 2 describes the average rate of accumulation of amino acid–replacing substitutions fairly well, explaining more than 50% of the variance in the data in all cases (Fig. 1). Moreover, the pattern is quite similar across species. The estimates of dR/dS at low S are all < 1, with a narrow range of 0.37 to 0.46 and a mean value of 0.43 (SE = 0.01), and dR/dS gradually declines to asymptotic values in the range of 0.022 to 0.106 (mean = 0.053, SE = 0.009) (Table 1). These results imply that, early in their evolutionary history, duplicate genes tend to be under moderate selective constraints with the rate of amino acid substitution averaging about 43% of the neutral expectation. The efficiency of purifying selection subsequently increases approximately 10-fold, to the point at which only about 5% of amino acid–changing mutations are able to rise to fixation.

Some caveats in the interpretation of these results are in order. First, the nucleotide divergence statistics describe the average pattern of molecular evolution. Individual codons may, in many cases, deviate substan-

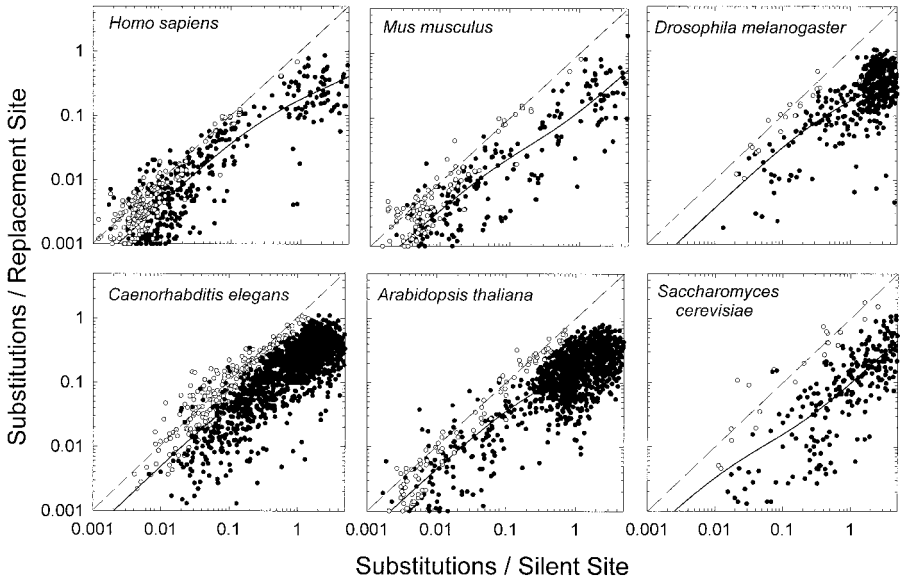

Fig. 1. Cumulative numbers of observed replacement substitutions per replacement site as a function of the number of silent substitutions per silent site. Each point represents a single pair of gene duplicates. The dashed line denotes the expectation under the neutral model, whereas the solid line is the least-squares fit of Eq. 2 to the data (*22*). Open points denote gene pairs for which the ratio R/S is not significantly different from the neutral expectation of 1.

Table 1. Fitted coefficients for the function describing cumulative replacement substitutions per replacement site versus silent substitutions per silent site, Eq. 2, and for the function describing the rate of loss of young duplicates, Eq. 3. The value r^2 gives the proportion of variance in the observed values described by the model; standard errors are in parentheses.

Species	Equation 2				Equation 3	
	m	$(dR/dS)_{S=0}$	$(dR/dS)_{S=\infty}$	r^2	d	r^2
H. sapiens	0.412	0.442	0.038	0.759	23.9 (2.0)	0.954
M. musculus	6.574	0.388	0.106	0.730	13.9 (3.2)	0.698
G. gallus	0.829	0.382	0.032	0.720	–	–
Danio rerio	0.857	0.450	0.022	0.677	–	–
D. melanogaster	0.564	0.372	0.050	0.533	8.2 (1.6)	0.766
C. elegans	0.547	0.500	0.062	0.647	7.0 (1.5)	0.735
A. thaliana	0.695	0.458	0.043	0.750	17.6 (5.0)	0.605
O. sativa	0.500	0.412	0.034	0.540	–	–
S. cerevisiae	20.357	0.433	0.090	0.531	7.5 (2.4)	0.538

tially from the norm. Second, for gene pairs with $S > 1$, potentially large inaccuracies in the estimates of nucleotide divergence are expected to result from multiple substitutions per site. Nevertheless, as can be seen in Fig. 1, the patterns that we describe are fully apparent within the subset of gene duplicates with $S < 1$. Third, although we have taken special precautions to avoid the inclusion of nonfunctional gene duplicates in our analyses (*21*), in the absence of actual expression pattern data, we cannot be certain that all of the genes we have included are functional. However, the fact that most of the pairs that we have identified have $R/S < 1$ and that many pairs with small S have $R/S \gg 1.0$ suggests that we have not inadvertently included many pseudogenes in our analyses.

Assuming that the number of silent substitutions increases approximately linearly with time, the relative age-distribution of gene duplicates within a genome can be inferred indirectly from the distribution of S (*23*). For all species, the highest density of duplicates is contained within the youngest age classes, with the density dropping off very rapidly with increasing S (Fig. 2). For *Arabidopsis*, there is a conspicuous secondary peak in the age distribution centered around $S = 0.8$, which is consistent with conclusions from comparative mapping data that the lineage containing this species experienced an ancient polyploidization event (*24*). Using an estimated rate of silent-site substitution of 6.1 per silent site per billion years (*25*), this event dates to approximately 65 million years ago. Unfortunately, this type of analysis cannot shed much light on the debate over whether complete genome duplications preceded the divergence of ray-finned fishes and tetrapods (*1–3, 26–28*). With a divergence time between these two lineages at approximately 430 million years ago (*29*), the average S for a pair of older duplicates would be expected to be in excess of 1.0. Levels of substitution of this magnitude are estimated with a great degree of inaccuracy, which would weaken the signature of ancient genome-duplication events.

For levels of divergence less than $S = 0.25$, problems with saturation effects in the estimation of substitutions per site should be minimal, and the time scale is short enough that it is reasonable to expect the rate of evolution at silent sites to be approximately constant. If the origin and loss of duplicates is then viewed as having been an essentially steady-state process over the time period $S = 0$ to 0.25, the rate of loss of gene duplicates can be estimated by using the survivorship function

$$N_S = N_0 e^{-dS} \quad (3)$$

where N_S is the number of duplicates observed at divergence level S, and N_0 and d are fitted constants obtained by linear regression of the log-transformed data (Fig. 3) (*30*). For the species for which adequate data are available for analysis, the loss coefficients fall in the range of $d = 7$ to 24, with a mean value of 13.0 (SE = 2.8) (Table 1). For $d = 7$, 13, and 24, the half-life of a gene duplicate on the scale of S is 0.099, 0.053, and 0.029, respectively, and 95% loss is expected at 4.3 times these S values. Thus, assuming they are not nonfunctional at the time of origin, most gene duplicates are apparently nonfunctionalized by the time silent sites have diverged by only a few percent.

Some insight into the absolute time to duplicate-gene loss can be acquired for the groups in which estimated rates of nucleotide evolution at silent sites are available. The average estimate of d for mouse and human is 18.9, which, using an average rate of silent substitution in mammalian genes of 2.5 per silent site per billion years (*31*), translates to 7.3 million years. The estimates of d for the two invertebrates *Drosophila* and *Caenorhabditis* are very similar, averaging to 7.6. Although a direct estimate of the rate of silent substitution is not available for nematodes, indirect evidence suggests that the rate of molecular evolution in *C. elegans* is elevated relative to that in other invertebrates (*32*). Using the estimated rate of silent-site substitution in *Drosophila* of 15.6 per silent site per BY (*7*), we obtain a possibly upwardly biased estimate of 2.9 million years as the average half-life of duplicate genes in invertebrates. For *Arabidopsis*, $d = 17.6$, which translates into a half-life of 3.2 million years using the silent substitution rate cited above.

Finally, we note that for the three species for which the complete genomic sequence is available, the rate of origin of gene duplicates can be estimated from the abundance of the very youngest pairs. For *D. melanogaster*, there are 10 pairs of duplicates with $S < 0.01$, which translates to a rate of origin of approximately 31 new duplicates per genome per million years, or by using the estimated 13,601 genes per genome (*33*), to 0.0023 per gene per million years. There are 32 identifiable duplicates in yeast with $S < 0.01$. Although no direct estimates of the rate of nucleotide substitution exist for fungi, there is no evidence that the fungal rate is very different from that of animals or plants either. Using the average silent substitution rate for mammals, *Drosophila*, and vascular plants (8.1 per nucleotide site per BY), the crudely estimated number of new duplicates arising in the yeast genome per million years is 52; with a total genome of approximately 6241 open reading frames, this translates to 0.0083 per million years. The rate of origin of gene duplicates in *C. elegans* over the past few hundred thousand years appears to be substantially greater than that for *D. melanogaster* and *S. cerevisiae*. There are 164 pairs of gene dupli-

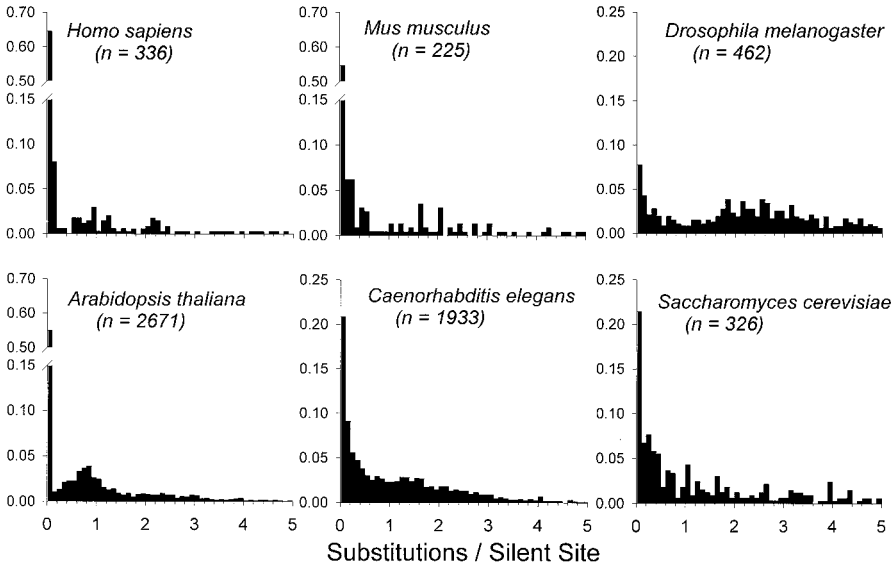

Fig. 2. Frequency distributions of pairs of duplicates as a function of the number of silent substitutions per silent site.

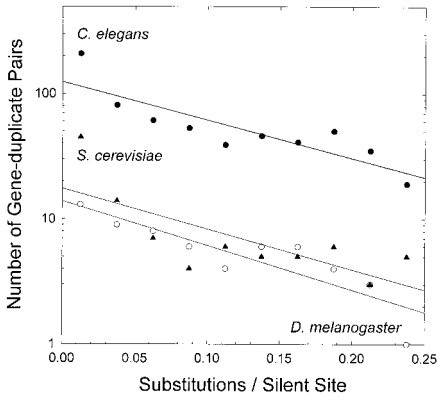

Fig. 3. Survivorship curves for gene duplicates, based on the complete genomic sequences of *C. elegans* (●), *D. melanogaster* (○), and *S. cerevisiae* (▲). The fitted parameters for these and other species are contained in Table 1.

cates with $S < 0.01$ in *C. elegans*. Again using the rate of silent-site substitution from *Drosophila*, the rate of origin of new duplicates in this species is at least 383 per genome per million years; with a genome size of approximately 18,424 open reading frames (*33*), this translates to a per-gene rate of duplication of 0.0208 per million years.

These estimated rates of origin of new gene duplicates could be inflated if gene conversion keeps substantial numbers of older duplicates appearing as if they were younger. Of the young duplicates identified in the previous paragraph, 100% of those in *Drosophila*, 56% of those in *Saccharomyces*, and 71% of those in *Caenorhabditis* are located on the same chromosome. However, although significant, the correlation between *S* and the physical distance between duplicates residing on the same chromosome tends to be quite weak, and many spatially contiguous gene duplicates are highly divergent (see figure at www.csi.uoregon.edu/projects/genetics/duplications). In addition, a genome-wide analysis of *C. elegans* suggests that gene-conversion events arise only rarely in duplicate genes and are largely concentrated in multigene families (*34*). Such multigene families have been excluded from our analyses (*21*).

These results suggest a conservative estimate of the average rate of origin of new gene duplicates on the order of 0.01 per gene per million years, with rates in different species ranging from about 0.02 down to 0.002. Given this range, 50% of all of the genes in a genome are expected to duplicate and increase to high frequency at least once on time scales of 35 to 350 million years. Thus, even in the absence of direct amplification of entire genomes (polyploidization), gene duplication has the potential to generate substantial molecular substrate for the origin of evolutionary novelties. The rate of duplication of a gene is of the same order of magnitude as the rate of mutation per nucleotide site (*7*).

However, the fate awaiting most gene duplicates appears to be silencing rather than preservation. For the species that we have examined, the average half-life of a gene duplicate is approximately 4 million years, consistent with the theoretical predictions mentioned above (*11, 12*). The contrast between the high rate of silencing observed in this study and the high level of duplicate-gene preservation that occurs in polyploid species (*17–20*) may be reconciled if dosage requirements play an important role in the selective environment of gene duplicates. Polyploidization preserves the necessary stoichiometric relationships between gene products, which may be subsequently maintained by stabilizing selection, whereas duplicates of single genes that are out of balance with their interacting partners may be actively opposed by purifying selection.

Despite the rather narrow window of opportunity for evolutionary exploration by gene duplicates, such genes may play a prominent role in the generation of biodiversity by promoting the origin of postmating reproductive barriers (*35, 36*). Consider a young pair of functionally redundant duplicate genes in an ancestral species. If a geographic isolating event occurs, a random copy will be silenced in the two sister taxa with very high probability within the next one to 2 million years. The probability that alternative copies will be silenced in the two sister taxa is 0.5, so if the copies are unlinked and the two taxa are then brought back together, there will be a 0.0625 probability that an F_2 derivative will be a double-null homozygote for the two loci. With tens to hundreds of young, unresolved gene duplicates present in most eukaryotic genomes, such genes may provide a common substrate for the passive origin of isolating barriers. Moreover, this process does not simply rely on gene duplicates in ancestral species. With rates of establishment of 0.002 to 0.02 duplicates per gene per million years and a moderate genome size of 15,000 genes, we can expect on the order of 60 to 600 duplicate genes to arise in a pair of sister taxa per million years, many of which will subsequently experience divergent resolution.

The passive build-up of reproductive isolation induced by gene duplicates, with no loss (and in most cases, no gain) of fitness in sister taxa, provides a simple mechanism for speciation that is consistent with the Bateson-Dobzhansky-Muller model (*37*), without requiring the presence of negative epistatic interactions between gene products derived from isolated genomes. The microchromosomal repatterning induced by recurrent gene duplication is also consistent with the chromosomal model for speciation (*38*), without requiring the large-scale rearrangements that are typically thought to be necessary (*39*). Finally, the time scale of the process is consistent with what we know about the average time to postreproductive isolation (*40, 41*).

References and Notes

1. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Heidelberg, Germany, 1970).
2. P. W. H. Holland, J. Garcia-Fernandez, N. A. Williams, A. Sidow, *Development (suppl.)*, p. 125 (1994).
3. A. Sidow, *Curr. Opin. Genet. Dev.* **6**, 715 (1996).
4. J. B. S. Haldane, *Am. Nat.* **67**, 5 (1933).
5. H. J. Muller, *Genetics* **17**, 237 (1935).
6. J. B. Walsh, *Genetics* **110**, 345 (1985).
7. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1999).
8. A. Force *et al.*, *Genetics* **151**, 1531 (1999).
9. A. L. Hughes, *Proc. R. Soc. London Ser. B* **256**, 119 (1994).
10. A. Stoltzfus, *J. Mol. Evol.* **49**, 169 (1999).
11. M. Lynch, A. Force, *Genetics* **154**, 459 (2000).
12. G. A. Watterson, *Genetics* **105**, 745 (1983).
13. M. Lynch, B. Walsh, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA, 1998).
14. W.-H. Li, in *Population Genetics and Molecular Evolution*, T. Ohta, K. Aoki, Eds. (Springer-Verlag, Berlin, 1985), pp. 333–352.
15. M. K. Hughes, A. L. Hughes, *Mol. Biol. Evol.* **10**, 1360 (1993).
16. B. S. Gaut, J. F. Doebley, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6809 (1997).
17. S. D. Ferris, G. S. Whitt, *J. Mol. Evol.* **12**, 267 (1979).
18. J. H. Nadeau, D. Sankoff, *Genetics* **147**, 1259 (1997).
19. A. Amores *et al.*, *Science* **282**, 1711 (1998).
20. J. F. Wendel, *Plant Mol. Biol.* **42**, 225 (2000).
21. For each organism, the complete set of available putative amino acid sequences was downloaded from GenBank and stored in a local file. We first filtered out possible nonfunctional protein sequences by removing all those that did not start with methionine, and all sequences that were annotated as known or suspected pseudogenes or transposable elements were also discarded. We then used BLAST [S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997)] to compare all pairs of protein sequences in the set, retaining those pairs for which the alignment score was below 10^{-10}. To avoid the inclusion of large multigene families (and as a secondary guard against the inclusion of transposable elements), we further excluded all genes that identified more than five matching sequences. All remaining sequences that were similar to known transposable elements were also removed at this point. As a final step in preparing the data set for analysis, we attempted to minimize the noise inherent in poorly aligned sequences by adopting a conservative strategy for retaining only unambiguously aligned sequences. Using the protein alignment generated by BLAST as a guide, at each alignment gap we scanned to the right and discarded all sequence until an anchor pair of identical amino acids was located, then retained all subsequent sequence (exclusive of the anchor site) provided another match was found within the next six amino acid sites. This procedure was iterated to the right and to the left of each alignment gap until acceptable anchor sites were encountered. We used a two-out-of-seven rule because the probability of a lower level of sequence identity arising by chance is >0.05, and we excluded anchor sites to minimize upward bias in estimated sequence similarities. Using the final set of amino acid alignments as guides, we then retrieved and aligned all of the necessary nucleotide sequences. The numbers of nucleotide substitutions per silent and replacement sites were estimated by using the maximum-likelihood procedure in the PAML package, version 2.0k [Z. Yang, *Comput. Appl. Biosci.* **13**, 555 (1997)]. Estimated rates of nucleotide substitution can be highly sensitive to the relative rates of occurrence of transitions and transversions when the amount of sequence divergence is high. Therefore, before our final analyses, we obtained species-specific estimates of the transition/transversion bias by considering the observed substitutions at all fourfold redundant sites in all pairs of sequences that were similar enough that multiple substitutions per site were unlikely to have occurred (we used only proteins for which the divergence at such sites is ≤15%, after verifying that the transition/transversion ratio is essentially constant below this point). The estimated transition/transversion ratios were then treated as constants in the maximum-likelihood analyses.
22. The parameter estimates for Eq. 2 are those that minimize the sum of squared deviations between predicted and observed values. In these analyses, the natural logarithms of observed and expected values were used, so as not to give undue weight to sequence pairs with large *R*. Pairs of sequences for which S or $R = 0$ were not included in this analysis. In addition, for human and mouse, we excluded gene pairs for which $S < 0.01$ in the curve-fitting procedure to avoid the accidental inclusion of alleles at the same locus; this problem should be negligible for the other large genome projects, where the species are either largely selfing (plants and *C. elegans*) or the final data exclude the possibility of allelic inclusion (*D. melanogaster*). The final data, including gene numbers, can be located at www.csi.uoregon.edu/projects/genetics/duplications.
23. Such analysis implicitly assumes an unbiased sample of gene duplicates, which should be valid for most

analysis herein, because they are based on either completely sequenced genomes or random sequencing projects.
24. D. Grant, P. Cregan, R. C. Showemaker, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4168 (2000).
25. This estimate is the average of two surveys based on analyses of multiple genes in vascular plants (*7, 42*).
26. L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
27. A. L. Hughes, *J. Mol. Evol.* **48**, 565 (1999).
28. A. P. Martin, *Am. Nat.* **154**, 111 (1999).
29. P. E. Ahlberg, A. R. Milner, *Nature* **368**, 507 (1994).
30. In large genomic databases like those analyzed here, some sequencing errors may inflate the apparent level of divergence, but this error should be independent of *S*, and in any event, is unlikely to add more than 0.01 to individual estimates of *S*. Thus, the impact of such error on our statistical analyses should be negligible.
31. This estimate is the average of the results obtained in three broadly compatible studies, all of which surveyed a large number of genes [S. Easteal, C. Collet, *Mol. Biol. Evol.* **11**, 643 (1994); T. Ohta, *J. Mol. Evol.* **40**, 56 (1995); (*7*)].
32. A. M. A. Aguinaldo *et al.*, *Nature* **387**, 489 (1997).
33. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
34. C. Semple, K. H. Wolfe, *J. Mol. Evol.* **48**, 555 (1999).
35. C. R. Werth, M. D. Windham, *Am. Nat.* **137**, 515 (1991).
36. M. Lynch, A. Force, *Am. Nat.*, in press.
37. H. A. Orr, *Genetics* **144**, 1331 (1996).
38. M. J. D. White, *Modes of Speciation* (Freeman, San Francisco, CA, 1978).
39. G. Fischer, S. A. James, I. N. Roberts, S. G. Oliver, E. J. Louis, *Nature* **405**, 451 (2000).
40. H. R. Parker, D. P. Philipp, G. S. Whitt, *J. Exp. Zool.* **233**, 451 (1985).
41. J. A. Coyne, H. A. Orr, *Evolution* **51**, 295 (1997).
42. M. Lynch, *Mol. Biol. Evol.* **14**, 914 (1997).
43. This research was supported by NIH grant RO1-GM36827. We thank A. Force and A. Wagner for helpful comments.

22 June 2000; accepted 4 October 2000

The Genetic Legacy of Paleolithic *Homo sapiens sapiens* in Extant Europeans: A Y Chromosome Perspective

Ornella Semino,[1,2]*† Giuseppe Passarino,[2,3]† Peter J. Oefner,[4] Alice A. Lin,[2] Svetlana Arbuzova,[5] Lars E. Beckman,[6] Giovanna De Benedictis,[3] Paolo Francalacci,[7] Anastasia Kouvatsi,[8] Svetlana Limborska,[9] Mladen Marcikiæ,[10] Anna Mika,[11] Barbara Mika,[12] Dragan Primorac,[13] A. Silvana Santachiara-Benerecetti,[1] L. Luca Cavalli-Sforza,[2] Peter A. Underhill[2]

A genetic perspective of human history in Europe was derived from 22 binary markers of the nonrecombining Y chromosome (NRY). Ten lineages account for >95% of the 1007 European Y chromosomes studied. Geographic distribution and age estimates of alleles are compatible with two Paleolithic and one Neolithic migratory episode that have contributed to the modern European gene pool. A significant correlation between the NRY haplotype data and principal components based on 95 protein markers was observed, indicating the effectiveness of NRY binary polymorphisms in the characterization of human population composition and history.

Various types of evidence suggest that the present European population arose from the merging of local Paleolithic groups and Neolithic farmers arriving from the Near East after the invention of agriculture in the Fertile Crescent (*1*–*5*). However, the origin of Paleolithic European groups and their contribution to the present gene pool have been debated (*6*, *7*). Assuming no selection, local differentiation occurred in isolated and small Paleolithic groups by drift (*8*, *9*). Range expansions and population convergences, which occurred at the end of the Paleolithic, were catalyzed by improved climate and new technologies and spread the present genetic characteristics to surrounding areas (*8*). The smaller effective population size of the NRY enhances the consequences of drift and founder effect relative to the autosomes, making NRY variation a potentially sensitive index of population composition. Previously, the distribution of two NRY restriction fragment length polymorphism (RFLP) markers suggested Paleolithic and Neolithic contributions to the European gene pool (*10*). NRY binary markers (*11*) representing unique mutational events in human history allow a more comprehensive reconstruction of European genetic history.

Twenty-two relevant binary markers [4 gathered from the literature and 18 detected by denaturing high-performance liquid chromatography (DHPLC) (*12*)] were genotyped in 1007 Y chromosomes from 25 different European and Middle Eastern geographic regions. More than 95% of the samples studied could be assigned to haplotypes or clades of haplotypes defined by just 10 key mutations (Fig. 1 and Table 1). The frequency distribution of Y chromosome haplotypes revealed here defines the basic structure of the male component of the extant European populations and provides testimony to population history, including the Paleolithic period. Two lineages (those characterized by M173 and M170) appear to have been present in Europe since Paleolithic times. The remaining lineages entered Europe most likely later during independent migrations from the Middle East and the Urals as they are found at higher frequencies and with more variation of linked microsatellites than in other continents (*10*–*14*).

Of the 22 haplotypes that constitute the phylogeny in Fig. 1 (top), Eu18 and Eu19 characterize about 50% of the European Y chromosomes. Although they share M173, the two haplotypes show contrasting geographic distribution. The frequency of Eu18 decreases from west to east, being most frequent in Basques (Fig. 1, bottom, and Table 1). This lineage includes the previously described proto-European lineage that is characterized by the 49a,f haplotype 15 (*10*). In contrast, haplotype Eu19, which is derived from the M173 lineage and is distinguished by M17, is virtually absent in Western Europe. Its frequency increases eastward and reaches a maximum in Poland, Hungary, and Ukraine, where Eu18 in turn is virtually absent. Both haplotypes Eu18 and Eu19 share the derived M45 allele. The lineage characterized by M3, common in Native Americans

[1]Dipartimento di Genetica e Microbiologia, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy. [2]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305–5120, USA. [3]Dipartimento di Biologia Cellulare, Università della Calabria, 87030 Rende, Italy. [4]Stanford Genome Technology Center, 855 California Avenue, Palo Alto, CA 94304, USA. [5]International Medico-Genetic Centre, Hospital Nol, 57 Artem Str, 340000 Donetsk, Ukraine. [6]Department of Oncology, Pathology and Medical Genetics, University of Umeå, S-901 85 Umeå, Sweden. [7]Dipartimento di Zoologia e Antropologia Biologica, Università di Sassari, Via Regina Margherita, 15, 07100 Sassari, Italy. [8]Department of Genetics, Development and Molecular Biology, Aristotle University, 54006 Thessaloniki, Macedonia, Greece. [9]Institute of Molecular Genetics, Russian Academy of Sciences, Kurchatov Square, 2, Moscow 123182, Russia. [10]Clinical Hospital Center Osijek, Department of Pathology Medical School, J Huttlera 4, 31000 Osijek, Croatia. [11]Regionalne Centrum Krwiodawstwa i Krwiolecznictwa w Lublinie–Oddzial w, Zamosciu, ul Legionow 10, 22400 Zamosc, Poland. [12]Samodzielny Publiczny Szpital Wojwodzki im. Papieza Jona Pawla II w, Zamosciu, ul Legionow 10, 22400 Zamosc, Poland. [13]University Hospital Split, Department of Pediatrics, Laboratory for Clinical and Forensic Genetics, Spinèiæeva 1, 21000 Split, Croatia.

*To whom correspondence should be addressed. E-mail: semino@ipvgen.univp.it
†These authors contributed equally to this work.