Comparable Rates of Gene Loss and Functional Divergence After Genome Duplications Early in Vertebrate Evolution

Joseph H. Nadeau* and David Sankoff†

**Genetics Department, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106-4955 and †Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada*

Manuscript received March 3, 1997
Accepted for publication July 16, 1997

ABSTRACT

Duplicated genes are an important source of new protein functions and novel developmental and physiological pathways. Whereas most models for fate of duplicated genes show that they tend to be rapidly lost, models for pathway evolution suggest that many duplicated genes rapidly acquire novel functions. Little empirical evidence is available, however, for the relative rates of gene loss *vs.* divergence to help resolve these contradictory expectations. Gene families resulting from genome duplications provide an opportunity to address this apparent contradiction. With genome duplication, the number of duplicated genes in a gene family is at most 2^n, where n is the number of duplications. The size of each gene family, *e.g.*, 1, 2, 3, . . . , 2^n, reflects the patterns of gene loss *vs.* functional divergence after duplication. We focused on gene families in humans and mice that arose from genome duplications in early vertebrate evolution and we analyzed the frequency distribution of gene family size, *i.e.*, the number of families with two, three or four members. All the models that we evaluated showed that duplicated genes are almost as likely to acquire a new and essential function as to be lost through acquisition of mutations that compromise protein function. An explanation for the unexpectedly high rate of functional divergence is that duplication allows genes to accumulate more neutral than disadvantageous mutations, thereby providing more opportunities to acquire diversified functions and pathways.

GENE duplication is believed to play an important role during evolution by providing opportunities to evolve new gene functions that can lead to novel morphologies, physiologies and behaviors (HALDANE 1933a,b; OHNO 1970; HOOD *et al.* 1975; TARTOF 1975; FRYXELL 1996; SHARMAN and HOLLAND 1996). Two contrasting theoretical expectations describe the fate of duplicated genes. Many models show that most duplicated genes are rapidly lost because they acquire null mutations (HALDANE 1993b; FISHER 1935; FERRIS and WHITT 1977, 1979; BAILEY *et al.* 1978; ALLENDORF 1979; KIMURA and KING 1979; TAKAHATA and MARUYAMA 1979; LI 1980, 1982; MARUYAMA and TAKAHATA 1982; WATTERSON 1983; BASTEN and OHTA 1992; HUGHES and HUGHES 1993). Duplicated genes persist only if mutations create new and eventually essential protein functions, an event that is predicted to occur rarely (HALDANE 1933; OHNO 1970; KIMURA and OHTA 1974; OHTA 1980, 1988; CLARK 1994; WALSH 1995). By contrast, models of pathway evolution suggest that diversification of developmental and physiological functions depends on many genes acquiring novel protein functions and that this is most likely to occur if many genes are duplicated simultaneously (WAGNER 1994, 1996; FRYXELL 1996). Empirical estimates of the rates of gene loss functional divergence are important in assessing these models for the fate of duplicated genes.

We focused on gene families arising from genome duplications that occurred early in vertebrate evolution. With genome duplication, all genes are duplicated simultaneously, hence the number of genes in the family is known and all proteins are functional. With time, some gene family members will be lost and the gene family will become smaller. The relative rates of gene loss *vs.* functional diversification can be estimated from the number of duplicated genes that persist *vs.* those that have been lost. By contrast, tandem duplication and reverse transcription occur continuously and produce an unknown number of duplicates, at unknown times, and perhaps without protein function (WEINER *et al.* 1986).

Chromosome location distinguishes genes that arise from genome duplication *vs.* tandem duplication. Tandem duplication expands or contracts the number of related genes at a given locus. Reverse transcription moves copies of genes to unrelated sites in the genome. However, with genome duplication and subsequent chromosome rearrangements, duplicated genes are part of duplicated segments that often reflect the order and distances of genes in the ancestral genome (LUNDIN 1979, 1989, 1993; NADEAU 1991). Genetic maps for humans and mice provide the strongest evidence for these duplications with families of related genes being found on several chromosomes and with many exam-

Corresponding author: Joseph H. Nadeau, Genetics Department, Case Western Reserve University School of Medicine, 10900 Euclid Ave., Cleveland, OH 44106-4955.
E-mail: jhn4@po.cwru.edu

ples of duplicated chromosome segments (COMINGS 1972; LUNDIN 1979, 1989, 1993; MORIZOT 1986; NADEAU 1991). Considerable evidence suggests that two genome duplications occurred early in vertebrate evolution, the most recent being ~250 myr ago (ATKIN and OHNO 1967; HINEGARDNER 1968; OHNO *et al.* 1968). As a result, most gene families are expected to consist of one, two, three or four genes. Genetic maps support this expectation (LUNDIN 1979, 1989, 1993; NADEAU 1991).

Our strategy in this article is to develop models in which values can be estimated for the critical parameters that determine the distribution of gene family sizes in humans and mice. All models assume that at least one member of each gene family will remain active. If a gene provides an essential function, it is assumed that the need for this function will remain following genome duplication. The models we propose enabled us to explore several major questions arising from the history of genome duplications: what are the relative rates of gene loss *vs.* functional divergence, did the second duplication occur before resolution of gene fate (loss or functional divergence) following the first duplication, and to what extent does functional compensation operate among duplicate genes even after they have begun to diverge functionally.

The model of genome duplications: After the first genome duplication, each gene within the haploid complement will be represented twice (Figure 1). In some cases, one of these duplicates is eventually lost through mutations and their families are again composed of a single member. In other cases, both copies are retained after duplication. More specifically, one of the two family members loses function rapidly, gene family members retain overlapping functions for some time pending resolution of loss *vs.* divergence, or each family member acquires novel and essential functions assuring their preservation. Equilibrium is reached when each gene family has lost all but one of its exemplars, or else the two or more remaining genes have all become functionally distinct. Because most mutations are null or inactivating with respect to function of the gene product, loss *vs.* divergence is usually resolved quickly (HALDANE 1933b; FISHER 1935; FERRIS and WHITT 1977, 1979; BAILEY *et al.* 1978; ALLENDORF 1979; KIMURA and KING 1979; TAKAHATA and MARUYAMA 1979; LI 1980, 1982; MARUYAMA and TAKAHATA 1982; WATTERSON 1983; BASTEN and OHTA 1992; HUGHES and HUGHES 1993). The distribution of gene family sizes documents the outcome of these alternative fates of gene loss and functional divergence.

The outcome after the second genome duplication depends on the fate of the first genome duplication (Figure 1). If both of the original duplicates remain from the first genome duplication, then gene families consist of four genes immediately after the second duplication. There are four possible outcomes after the second duplication: all four members of a gene family could be kept, three could be kept and one lost, two could be kept and two lost, or one could be kept and three lost; the fifth possibility, namely, lose all four copies, is assumed to be rare because at least one copy of an essential gene is expected to be retained. By contrast, if one of the two original gene family members has been lost and only one remains after the first duplication, families will consist of only two genes immediately after the second duplication. With time, two outcomes are possible, either both members of the gene family are kept, or else one of them is lost and one gene with the original function is retained. Again, we assume that at least one family member will be retained because it provides an essential function. Thus, after two genome duplications, families may consist of one, two, three or four genes.

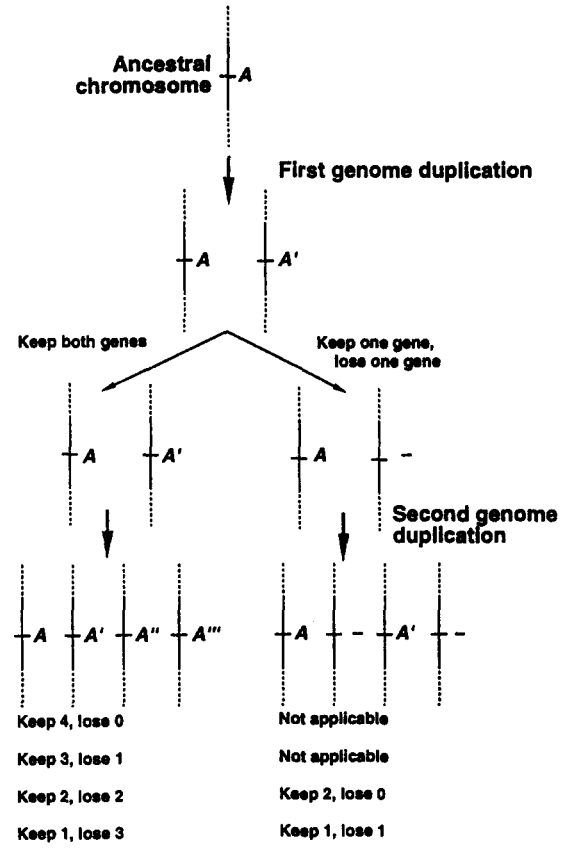

FIGURE 1.—Alternative fate of genes following successive genome duplications. We assume that at least one member of each gene family must be retained, *e.g.*, the "keep 0, lose 4" and the "keep 0, lose 2" options are selectively disadvantageous and hence are not represented. *A*, *A'*, *A''* and *A'''* designate members of the *A* gene family. These may have identical, overlapping, or unique functions.—indicates that gene function has been lost.

Gene family data: Gene lists and chromosome localizations were obtained from the Mouse Genome Database (as posted February 1, 1996) and the Human Genome Database (as posted May 29, 1996). To identify gene family members, emphasis was usually placed on

interpretation of experts working on particular gene families. To be included in the analysis, family members must have extensive DNA sequence similarity implying homology through duplication rather than motif similarity alone or convergent evolution.

Another important criteria was that gene families must have arisen through duplication of a chromosome segment, a whole chromosome, or the entire genome with the important result that family members are located at distant genomic locations and often as part of known duplicated segments that include members of other linked gene families (LUNDIN 1979, 1989, 1993; NADEAU 1991). With tandem duplication or unequal crossingover, by contrast, gene families expand and contract locally and their size does not reflect genome duplication. Moreover, expansion and contraction of family size occurs continuously, rather than synchronously as occurs with genome duplication. For these reasons, closely linked family members such as each Hox complex were counted as a single entity.

Genes that occurred only once in the genome were excluded because of uncertainty about their origins: are they the only remaining family member, did they originate after the last genome duplication, does their sequence evolve so rapidly that other family members are difficult to identify with standard methods, or has insufficient effort been made to find duplicates in as yet uncharacterized areas of the genome? The latter problem is less pertinent with known multigene families, because the discovery of one duplicate gene tends to motivate a search for additional family members. These sources of spurious singe-gene "families" would lead to serious underestimates of the rate of functional divergence were they to be included in the dataset for analysis.

For humans, 276 families were identified, of which 188 had two members, 67 had three members, and 21 had four members. For the mouse, 176 families were identified. Of these, 120 had two members, 45 had three members, and 11 had four members. An example from humans of a family with two members is aconitase-1 and -2 on chromosomes *9p22* and *22q11.2*, respectively. An example of a family with three members is the retinoic acid receptors alpha, beta and gamma on chromosome *17q12*, *3p24.3* and *12q13*, respectively, and an example of a family with four members is the homeobox complexes A, B, C and D on chromosomes *7p15*, *17q21*, *12q12* and *2q31*, respectively. Presence of families with three or four members is evidence for at least two genome duplications (COMINGS 1972; LUNDIN 1979, 1989, 1993; MORIZOT 1986; NADEAU 1991; SHARMAN and HOLLAND 1996). Families with more than four members are found occasionally, but their origins are probably complex and will be the subject of a separate study.

Competing models: The first model we elaborate (*Section 3.1*) assumes that the two genome duplications occurred within a short time span compared to the time necessary to resolve gene loss *vs.* functional divergence. The latter two models assume that loss *vs.* divergence of the duplicate genes resulting from the first genome duplication is completely or partially decided before the second duplication. The first of these two models (*Section 3.2*) assumes that functional divergence is resolved before the second duplication. The final model (*Section 3.3*) assumes that at the time of the second duplication, functional divergence may not be complete and that the two persistent duplicate genes may still compensate each other. This model plays an important role in evaluating competing hypotheses about genome duplication and in the construction of a combined model that permits variability in the divergence patterns among different gene families.

In terms of proportions of gene families with two, three or four members, the data sets described in the previous section have only two degrees of freedom. Thus, to assess and compare the statistical fit of the three models, and hence the hypotheses they embody, they each should be based on one parameter only. (Any reasonable two-parameter model with two degrees of freedom can be solved to fit data exactly.) To assure consistency among our three models, all are based on the same "hidden" variable ψ, which represents the probability of gene loss rather than divergence, without taking into account the constraint that not all genes in a family can be lost. Because one of the outcomes of the unconstrained process (all members of a gene family are lost) is unobservable, due to natural selection, there is no way of estimating ψ directly from the observed proportions. Based on ψ, however, we can predict the proportion of two-, three- and four-member gene families by conditioning the process in the identical manner for each of the three models on the constraint that an entire gene family cannot be eliminated.

To illustrate, consider the duplication of a single gene whose functionality is unique. After enough time has elapsed for the fate of the two copies to be resolved, both copies can persist or either one can persist. Selective pressures rule out the loss of function of both. This can be modeled by the conditional probabilities

$$\Pr[\text{both survive}] = \frac{(1-\psi)^2}{1-\psi^2} = \frac{1-\psi}{1+\psi}$$

$$\Pr[\text{only one survives}] = \frac{2\psi(1-\psi)}{1-\psi^2} = \frac{2\psi}{1+\psi}.$$

For the two models (*Sections 3.2* and *3.3*) with a long time interval between the two genome duplications, the value of ψ, and hence the dynamics of loss and divergence, are assumed to be the same after the second duplication as after the first.

Section 3.1: The short-interval model with neither gene loss nor functional divergence between successive duplications: If the time interval between successive duplications is

short, the fate of a pair of genes issuing from the first duplication will not be decided before the second duplication: not enough mutations will have accumulated to silence either of the genes and not enough time will have elapsed for one to acquire a new function. What happens when the second duplication event gives rise to four exemplars? Because functional divergence has not occurred, as many as three genes in each family may lose function without causing a selective disadvantage. Thus

$$\Pr[\text{four genes survive}] = \frac{(1-\psi)^4}{1-\psi^4},$$

$$\Pr[\text{three genes survive}] = \frac{4\psi(1-\psi)^3}{1-\psi^4},$$

$$\Pr[\text{two genes survive}] = \frac{6\psi^2(1-\psi)^2}{1-\psi^4},$$

$$\Pr[\text{one gene survives}] = \frac{4\psi^3(1-\psi)}{1-\psi^4}.$$

Section 3.2: The long-interval model with complete functional divergence before the subsequent duplication: If the time interval between successive duplications is long, the fate of a pair of genes issuing from the first duplication will be settled before the second duplication: two genes with divergent and essential functions are retained or one gene is lost and one is retained. For the case of two genes with fully divergent functions before the second duplication, what happens when the second duplication event gives rise to two new duplicate pairs? In this model, we postulate that because of functional divergence, the genes in one new pair cannot functionally compensate for the genes in the other. One but not both genes in each pair may lose function, since loss of both would be selectively disadvantageous.

In the case where, after the first duplication only one copy is retained, the second duplication gives rise to just one pair of genes. These may acquire divergent functionality so that both are retained; alternatively no new functionality emerges and one copy is lost. Thus

$$\Pr[\text{four genes survive}] = \frac{(1-\psi)^3}{(1+\psi)^3},$$

$$\Pr[\text{three genes survive}] = \frac{4\psi(1-\psi)^2}{(1+\psi)^3},$$

$$\Pr[\text{two genes survive}] = \frac{2\psi(1-\psi)(3\psi+1)}{(1+\psi)^3},$$

$$\Pr[\text{one gene survives}] = \frac{4\psi^2}{(1+\psi)^2}.$$

These probabilities are derived by taking account of the fact that after the second duplication, if there are two pairs of duplicates, the two pairs evolve independently as if each were the result of a single gene duplication described by the illustrative equations before *Section 3.1.*

Section 3.3: The long-interval model with partial functional divergence: In this model, as in *Section 3.2,* the loss or divergence of duplicate genes is resolved before the second duplication. As in *Section 3.1,* however, we postulate that functional compensation can occur among any of the two, three or four genes in a family, despite ongoing functional divergence, perhaps because the time scale for complete divergence and loss of compensatory function tends to be longer than that for gene loss and perhaps longer than the interval between genome duplications. Then

$$\Pr[\text{four genes survive}] = \frac{(1-\psi)^5}{(1+\psi)(1-\psi^4)},$$

$$\Pr[\text{three genes survive}] = \frac{4\psi(1-\psi)^4}{(1+\psi)(1-\psi^4)},$$

$$\Pr[\text{two genes survive}] = \frac{6\psi^2(1-\psi)^3}{(1+\psi)(1-\psi^4)} + \frac{2\psi(1-\psi)}{(1+\psi)^2},$$

$$\Pr[\text{one gene survives}] = \frac{4\psi^3(1-\psi)^2}{(1+\psi)(1-\psi^4)} + \frac{4\psi^2}{(1+\psi)^2}.$$

Choosing among the models: We assessed the three models only on the basis of the numbers of two-, three- and four-gene families in the humans and mouse databases, for reasons given above. We estimated ψ in the models in *Section 3.1, 3.2* and *3.3* by maximum likelihood applied to the data summarized in Table 1, which also contains the results of this analysis.

Because all three models are based on a single parameter ψ, they are directly comparable. As formalized in the next section, the short-interval model is clearly inferior to both of the two long-interval models, indicating that the time elapsed between the two genome duplications was long enough so that the process of duplicate gene loss could at least partially run its course. It is unclear whether or not the remaining pairs of duplicate genes had diverged enough to lose the ability to compensate for each other functionally. The evidence for the two species examined point in opposite directions, despite the fact that this evidence pertains to the same evolutionary events in early vertebrate evolution, long before the split in the lineages leading to humans and mice.

Combined model: There is no reason to believe that the fate of all gene families will be resolved simultaneously. Suppose a fraction θ of genes evolve as in the model with complete divergence (*Section 3.2*) and $1-\theta$ show functional overlap (*Section 3.3*). Since our data for each species contain two degrees of freedom, whereas our models are both based on the same single

TABLE 1

Maximum likelihood analysis of human and mouse data under short interval, long interval with complete divergence, and long interval without complete divergence models

	Human gene families				Mouse gene families			
Model:		3.1	3.2	3.3		3.1	3.2	3.3
Inter-duplication time:		Short	Long	Long		Short	Long	Long
Complete divergence:	n	No	Yes	No	n	No	Yes	No
Family size								
4	21	13.5	18.4	21.3	11	8.0	10.4	13.6
3	67	80.8	73.6	64.3	45	50.2	45.2	41.0
2	188	181.7	184.0	190.4	120	17.8	120.4	121.4
Total	276	276.0	276.0	276.0	176	176.0	176.0	176.0
ψ		0.60 ± 0.05	0.50 ± 0.06	0.43 ± 0.05		0.61 ± 0.07	0.52 ± 0.08	0.43 ± 0.07
ΔL			2.63	3.09			0.78	0.34

Confidence intervals for ψ were calculated according to the 1-LOD rule. ΔL represents increase log likelihood of the long-interval models over the short-interval model.

parameter ψ, we can solve for ψ and θ by substituting the observed relative frequencies in any two of the three equations:

$$\Pr[\text{four genes survive}] = \theta\frac{(1-\psi)^3}{(1+\psi)^3} + (1-\theta)\frac{(1-\psi)^5}{(1+\psi)(1-\psi^4)},$$

$$\Pr[\text{three genes survive}] = \theta\frac{4\psi(1-\psi)^2}{(1+\psi)^3} + (1-\theta)\frac{4\psi(1-\psi)^4}{(1+\psi)(1-\psi^4)},$$

$$\Pr[\text{two genes survive}] = \theta\frac{2\psi(1-\psi)(3\psi+1)}{(1+\psi)^3} + (1-\theta)\left[\frac{6\psi^2(1-\psi)^3}{(1+\psi)(1-\psi^4)} + \frac{2\psi(1-\psi)}{(1+\psi)^2}\right].$$

In the human data, the solution of these simultaneous equations is $\psi = 0.44$ and $\theta = 0.22$, and in the mouse data $\psi = 0.50$ and $\theta = 0.84$.

In both cases, the likelihood of the solution is higher, but not significantly higher, than either of the models in *Sections 3.2* and *3.3*, and the 1-LOD confidence interval for ψ includes the entire interval [0,1]. Thus, more gene families will have to be documented and added to our database before we can be precise about the proportions of gene families corresponding to the alternative models.

The analogous exercise, combining the model in *Section 3.1* with that of *Sections 3.2* or *3.3*, while not biologically meaningful, because with whole genome duplications all gene families must have the same interduplication time, nonetheless allows us to compare the models statistically. Thus, with the human data, adding any proportion of θ of the short interduplication time model to either of the long-interduplication times models does not improve the likelihood significantly. By contrast, the two-parameter model is significantly better by the log-likelihood test than the short interduplication time model alone. The maximum likelihood estimate of θ, the proportion of the short interduplication time model is zero, or very close to it. For these data then, the short-interduplication time model is rejected. The less numerous mouse data, however, do not by themselves confirm these test results, although contrary indications are not apparent, *i.e.*, the long interduplication time models both have higher likelihood than the short interduplication time model. Moreover, combining the mouse and human samples strengthens the significance levels of the tests to reject the short-interduplication time model.

Comments on the significance of ψ and θ: The values of ψ suggest that the rate of gene loss is comparable to the rate of functional divergence, regardless of the model proposed to account for patterns of loss *vs.* divergence. The similarity between ψ values in the "combined model" for humans and mice is also found among solutions to the three one-parameter models (range = 0.43–0.60 for humans and = 0.43–0.61 for mice; Table 1). That $\psi = \sim 0.5$ means that $(1-\psi)/(1+\psi) = \sim 0.33$. In other words, duplicate genes are almost as likely to acquire a novel and essential function as they are to be lost through deleterious mutation. These results should apply to all duplicated genes, regardless of their mode of origin, *e.g.*, tandem duplication.

The values of θ shed light on progress toward resolving loss *vs.* divergence and the extent of functional complementation among gene family members that were present at the second genome duplication. Although more variable than ψ, neither of the θ values, 0.22 for humans *vs.* 0.84 for mice, approaches 0 or 1, indicating that each of the two component models is

appropriate for a significant number of families. The greater variability of θ as compared to ψ probably reflects the fact that both the partial and complete functional divergence models individually fit the data rather well, so that the estimate of θ is highly dependent on small data fluctuations. These fluctuations could result, for example, from biases in identifying gene family members either in the laboratory or in databases, or could simply reflect sampling error.

Comparison of theoretical predictions with the empirical data lead to rejection of one, but not all, of the proposed models. The "short interval model with neither gene loss nor functional divergence" (*Section 3.1*) is clearly inferior to the other models (Table 1). The data strongly support the argument that loss *vs.* divergence processes were being resolved when the second genome duplication occurred. By contrast, the likelihood for the "combined model" was not more than 1.0 more than the likelihood over the better of the individual models for either the human or mouse data, indicating that the data are ambiguous about the extent to which compensation *vs.* divergence was resolved among those pairs of duplicated genes where neither was lost. Gene families follow different patterns of loss, compensation and divergence; when the second genome duplication occurred, some divergent genes had already lost their original function, while many others continued to show functional compensation.

Finally, these results suggest that the two genome duplications occurred within a relatively short evolutionary interval, an interval that was long enough for genes resulting from the first duplication to begin deciding their fate, but short enough so that fate was not fully resolved when the second genome duplication occurred. The fate of duplicated genes is thought to be resolved rapidly during evolution (HALDANE 1933a,b; FISHER 1935; FERRIS and WHITT 1977, 1979; BAILEY *et al.* 1978; ALLENDORF 1979; KIMURA and KING 1979; TAKAHATA and MARUYAMA 1979; LI 1980, 1982; MARUYAMA and TAKAHATA 1982; WATTERSON 1983; BASTEN and OHTA 1992; HUGHES and HUGHES 1993; CLARK 1994; WALSH 1995). Partial functional divergence therefore implies that the second duplication followed soon, but not immediately, after the first duplication. Moreover, successful genome duplication is a rare event in vertebrate evolution (OHNO 1970). As a result, these two genome duplications may have occurred within the life span of a species that was predisposed to resolving tetraploidy that results from genome duplication.

DISCUSSION

The counter-intuitive result that the rates of gene loss and functional divergence have similar magnitude can be explained if presence of duplicated genes buffers the consequences of null mutations, thereby providing greater opportunities for new functions to evolve (KIMURA 1983; LI 1985; CLARK 1994). Previous analyses assumed that the probability of gene loss was at least an order of magnitude greater than the rate of functional divergence (KIMURA and OHTA 1974). This argument applies to genes with novel and essential functions where natural selection directly tests the consequences of each mutation. However, even when null mutations create neutral alleles, the rate of gene loss is thought to exceed the rate of functional divergence. With duplicated genes, gene families have greater flexibility to accommodate mutations that modify functions because of complementation among partially divergent gene family members. Mutations that are deleterious when only a single gene is present may be neutral if a gene duplicate compensates for loss of function. The high mutation rate that occurs immediately after gene duplication is consistent with this interpretation (GOODMAN 1976; LI and GOJOBORI 1983; LI 1985).

To what extent do current gene inventories retain enough information to reconstruct early patterns of gene family evolution? We extracted some 270 duplicate human gene families from the literature. This might seem to constitute but a small percentage of the 3000–6000 known genes. It corresponds rather well, however, to the proportion that known genes represent of the total estimated number, 50,000–100,000, of human genes. In other words, given that only ~5% of human genes have been identified in these inventories, when one member of a pair of duplicate genes is discovered, the chances are only ~5% that the other member will also be known, assuming that gene family members are independently discovered. Thus 270 families, *i.e.*, 5% of 3000–6000 genes, are what we would expect given the current state of knowledge of the human genome.

Theoretical studies make contradictory predictions about the fate of duplicated genes and pathway evolution. Traditional models show that most duplicated genes are rapidly lost (HALDANE 1933a,b; FISHER 1935; FERRIS and WHITT 1977, 1979; BAILEY *et al.* 1978; ALLENDORF 1979; KIMURA and KING 1979; TAKAHATA and MARUYAMA 1979; LI 1980, 1982; MARUYAMA and TAKAHATA 1982; WATTERSON 1983; BASTEN and OHTA 1992; HUGHES and HUGHES 1993; CLARK 1994; WALSH 1995), a conclusion that is not supported by the large number of multigene families in both humans and mice (Table 1). The critical parameters in these models are effective population size, mutation rate, fitness, time, and the ratio of advantageous to deleterious mutations. For some values of these parameters, it has been proposed that ~1% of gene duplicates acquire a new function and that ~99% lose function (WALSH 1995). Obviously these numbers do not fit the observed distribution of family sizes for humans or mice. Only when the effective populations size is ~500,000 are the rates of gene loss and functional divergence similar (WALSH 1995). Considerable evidence suggests that effective population

sizes for mammals have been much smaller by many orders of magnitude (BERRY 1981; SAGE 1981). The easiest way to accommodate a substantial rate of divergence is to assume that more mutations than previously thought contribute to functional divergence.

Models for pathway evolution have different expectations about gene fate in that they assume a substantial rate of functional divergence. In particular, mathematical models for evolution of gene networks show that the probability of network diversification is maximized when ~40% of the genes are duplicated simultaneously and if these duplicates are either dispersed through out the genome or are tightly clustered (WAGNER 1994, 1996). Genome duplication is obviously easier than other mechanisms such as tandem duplication to duplicate many genes simultaneously. More importantly, the ~40% estimate can be interpreted as the number of duplicated genes that acquire novel and essential functions, a result that is remarkably consistent with our estimated rate of functional divergence, *i.e.*, $(1 - \psi)/(1 + \psi) = \sim 0.33$, inferred from the distribution of gene family sizes.

The fate of duplicated genes has an important bearing not only on the acquisition of novel gene functions but also on the diversification of physiological pathways. Although the importance of duplication to the evolution of gene function is widely appreciated, their contribution to the evolution of pathways is only beginning to be recognized. Genes obviously do not act in isolation, but cooperate instead in complex networks of protein interactions that together create complex morphologies, physiologies and behaviors. A recent study of the evolution of functionally related gene families revealed striking examples of coevolution with functionally related families showing similar evolutionary histories (FRYXELL 1996). It is evident that early in vertebrate evolution considerable diversification of functions and pathways occurred. Perhaps coevolution of genes in pathways accounts for some of the high rate of functional divergence.

We thank JANAN EPPIG (M.G.D.) and KEN FASMAN (G.D.B.) for providing list of mouse and human genes for this work. Research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. D.S. is a Fellow of the Canadian Institute of Advanced Research.

LITERATURE CITED

ALLENDORF, F. W., 1979 Rapid loss of duplicate gene expression by natural selection. Heredity **43:** 247–258.

ATKIN, N. B., and S. OHNO, 1967 DNA values of four primitive chordates. Chromosoma **23:** 10–13.

BAILEY, G. S., R. T. M. POULTER and P. A. STOCKWELL, 1978 Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. Proc. Natl. Acad. Sci. USA **75:** 5575–5579.

BASTEN, C. J., and T. OHTA, 1992 Simulation study of a multigene family, with special reference to the evolution of compensatory advantageous mutations. Genetics **132:** 247–252.

BERRY, R. J., 1981 Population dynamics in the house mouse, pp. 395–426 in *Biology of the House Mouse,* edited by R. J. BERRY. Academic Press, New York.

CLARK, A. G., 1994 Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91:** 2950–2954.

COMINGS, D. E., 1972 Evidence for ancient tetraploidy and conservation of linkage groups in mammalian chromosomes. Nature **238:** 455–457.

FERRIS, S. D., and G. S. WHITT, 1977 Loss of duplicate gene expression after polyploidization. Nature **265:** 258–260.

FERRIS, S. D., and G. S. WHITT, 1979 Evolution of differential regulation of duplicate genes after polyploidization. J. Mol. Evol. **12:** 267–317.

FISHER, R. A., 1935 The sheltering of lethals. Am. Nat. **69:** 446–455.

FRYXELL, K. J., 1996 The coevolution of gene family trees. Trends Genet. **12:** 364–369.

GOJOBORI, T., W.-H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

GOODMAN, M., 1976 Protein sequences in phylogeny, pp. 141–159 in *Molecular Evolution,* edited by F. J. AYALA. Sinauer Associates, Sunderland, MA.

HALDANE, J. B. S., 1933a The part played by recurrent mutation in evolution. Am. Nat. **69:** 446–455.

HALDANE, J. B. S., 1933b *The Causes of Evolution.* Longwood Green, London.

HINEGARDNER, R., 1968 Evolution of cellular DNA content in teleost fishes. Am. Nat. **102:** 517–523.

HOOD, L., J. H. CAMPBELL and S. C. R. ELGIN, 1975 The organization, expression, and evolution of antibody genes and other multigene families. Annu. Rev. Genet. **9:** 305–353.

HUGHES, M. K., and A. L. HUGHES, 1993 Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis.* Mol. Biol. Evol. **10:** 1360–1369.

KIMURA, M., 1983 *The Neutral Theory of Evolution.* Cambridge University Press, London.

KIMURA, M., and J. T. KING, 1979 Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA **76:** 2858–2861.

KIMURA, M., and T. OHTA, 1974 On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA **71:** 2848–2852.

LI, W.-H., 1980 Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95:** 237–258.

LI, W.-H., 1982 Evolutionary change of duplicated genes, pp. 55–92 in *Isozymes: Current Topics in Biological and Medical Research,* Vol. 6, edited by M. C. RATTAZZI, J. G. SCANDALIOS and G. S. WHITT. A. R. Liss, New York.

LI, W.-H., 1985 Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy, pp. 333–353 in *Population Genetics and Molecular Evolution,* edited by T. OHTA and K. AOKI. Japan Scientific Society Press, Tokyo and Springer Verlag, Berlin.

LI, W.-H., and T. GOJOBORI, 1983 Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1:** 94–108.

LUNDIN, L.-G., 1979 Evolutionary conservation of large chromosomal segments reflected in mammalian gene maps. Clin. Genet. **16:** 72–81.

LUNDIN, L.-G., 1989 Gene homologies with emphasis on paralogous genes and chromosomal regions. Life Sci. Adv. (Genet.) **8:** 89–104.

LUNDIN, L.-G., 1993 Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16:** 1–19.

MARUYAMA, T., and N. TAKAHATA, 1982 Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicate loci. Heredity **46:** 49–57.

MORIZOT, D. C., 1986 Comparative gene mapping evidence for chromosome duplications in chordate evolution. Isozyme Bull. **19:** 9–10.

NADEAU, J. H., 1991 Genome duplication and comparative gene mapping, pp. 269–296 in *Advanced Techniques in Chromosome Research,* edited by K. T. ADOLPH. Marcel Dekker, New York.

OHNO, S., 1970 *Evolution by Gene Duplication.* Springer Verlag, New York.

OHNO, S., U. WOLF and N. B. ATKIN, 1968 Evolution from fish to mammals by gene duplication. Hereditas **59:** 169–187.

OHTA, T., 1980 Evolution and variation of multigene families. Lect. Notes Biomath. **37:** 1–131.

OHTA, T., 1988 Evolution by gene duplication and compensatory advantageous mutations. Genetics **120:** 841–847.

SAGE, R. D., 1981 Wild mice, pp. 40–90 in *The Mouse in Biomedical Research, Vol. 1, History, Genetics and Wild Mice,* edited by H. L. FOSTER, J. D. SMALL and J. G. FOX. Academic Press, New York.

SHARMAN, A. C., and P. W. H. HOLLAND, 1996 Conservation, duplication, and divergence of developmental genes during chordate evolution. Netherl. J. Zool. **46:** 47–67.

TAKAHATA, N., and T. MARUYAMA, 1979 Polymorphism and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. Proc. Natl. Acad. Sci. USA **76:** 4521–4525.

TARTOF, K. D., 1975 Redundant genes. Annu. Rev. Genet. **9:** 355–385.

WAGNER, A., 1994 Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. Proc. Natl. Acad. Sci. USA **91:** 4387–4391.

WAGNER, A., 1996 Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. Biol. Cybern. **74:** 557–567.

WALSH, J. B., 1995 How often do duplicated genes evolve new functions? Genetics **139:** 421–428.

WATTERSON, G. A., 1983 On the time for gene silencing at duplicate loci. Genetics **105:** 745–766.

WEINER, A. M., P. L. DEININGER and A. EFSTRADIATIS, 1986 Nonviral retrotransposons: genes, pseudogenes, and transposable elemants generated by the reverse flow of genetic information. Annu. Rev. Genet. **55:** 631–661.

Communicating editor: A. G. CLARK