Preservation of Duplicate Genes by Complementary, Degenerative Mutations

Allan Force,* Michael Lynch,* F. Bryan Pickett,[†] Angel Amores,*
Yi-lin Yan* and John Postlethwait*

**Department of Biology, University of Oregon, Eugene, Oregon 97403 and [†]Department of Biology,*
Loyola University of Chicago, Chicago, Illinois 60626

Manuscript received March 17, 1998
Accepted for publication December 28, 1998

ABSTRACT

The origin of organismal complexity is generally thought to be tightly coupled to the evolution of new gene functions arising subsequent to gene duplication. Under the classical model for the evolution of duplicate genes, one member of the duplicated pair usually degenerates within a few million years by accumulating deleterious mutations, while the other duplicate retains the original function. This model further predicts that on rare occasions, one duplicate may acquire a new adaptive function, resulting in the preservation of both members of the pair, one with the new function and the other retaining the old. However, empirical data suggest that a much greater proportion of gene duplicates is preserved than predicted by the classical model. Here we present a new conceptual framework for understanding the evolution of duplicate genes that may help explain this conundrum. Focusing on the regulatory complexity of eukaryotic genes, we show how complementary degenerative mutations in different regulatory elements of duplicated genes can facilitate the preservation of both duplicates, thereby increasing long-term opportunities for the evolution of new gene functions. The duplication-degeneration-complementation (DDC) model predicts that (1) degenerative mutations in regulatory elements can increase rather than reduce the probability of duplicate gene preservation and (2) the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions rather than the evolution of new functions. We present several examples (including analysis of a new *engrailed* gene in zebrafish) that appear to be consistent with the DDC model, and we suggest several analytical and experimental approaches for determining whether the complementary loss of gene subfunctions or the acquisition of novel functions are likely to be the primary mechanisms for the preservation of gene duplicates.

> For a newly duplicated paralog, survival depends on the outcome of the race between entropic decay and chance acquisition of an advantageous regulatory mutation.
>
> Sidow (1996, p. 717)

> On one hand, it may fix an advantageous allele giving it a slightly different, and selectable, function from its original copy. This initial fixation provides substantial protection against future fixation of null mutations, allowing additional mutations to accumulate that refine functional differentiation. Alternatively, a duplicate locus can instead first fix a null allele, becoming a pseudogene.
>
> Walsh (1995, p. 426)

> Duplicated genes persist only if mutations create new and essential protein functions, an event that is predicted to occur rarely.
>
> Nadeau and Sankoff (1997, p. 1259)

> Thus overall, with complex metazoans, the major mechanism for retention of ancient gene duplicates would appear to have been the acquisition of novel expression sites for developmental genes, with its accompanying opportunity for new gene roles underlying the progressive extension of development itself.
>
> Cooke *et al.* (1997, p. 362)

THE genomes of most organisms contain multiple copies of genes that are closely related in structure and function. Such gene families can arise from tandem duplications, as in the case of the HOX, hemoglobin, and keratin clusters in animals, or from polyploidization events such as those presumed to have preceded the origin of vertebrates (Ohno 1970; Morizot *et al.* 1991; Lundin 1993; Holland *et al.* 1994; Amores *et al.* 1998; Pébusque *et al.* 1998), brewer's yeast (Wolfe and Shields 1997; Seoighe and Wolfe 1998), and many plant species (Lewis 1979). The mechanism that preserves a large proportion of duplicate genes for long time periods, however, is unclear. The classical model predicts that duplicate genes initially have fully overlap-

Corresponding author: Allan Force, Department of Biology, University of Oregon, Eugene, OR 97403. E-mail: force@oregon.uoregon.edu

ping, redundant functions, such that one copy may shield the second copy from natural selection, if gene dosage is not critical. Because deleterious mutations occur much more frequently than beneficial mutations (Lynch and Walsh 1998), the classical model predicts that the most common fate for the duplicate pair should be the fixation of a null allele that prevents normal transcription, translation, and/or protein function, *i.e.*, the formation of a pseudogene at one of the duplicate loci (Haldane 1933; Nei and Roychoudhury 1973; Bailey *et al.* 1978; Takahata and Maruyama 1979; Li 1980; Watterson 1983). Under this model, first elucidated by Ohno (1970), the only mechanism for the permanent preservation of duplicate genes is the fixation of rare beneficial mutations endowing one of the copies with a novel function, while the second copy maintains the original function. The introductory quotations illustrate the extent to which this model is currently the central paradigm in the theory of duplicate gene evolution.

Here we discuss difficulties in the ability of the classical model to explain the preservation of gene duplicates in evolution and then propose a new model that can explain duplicate gene preservation by the fixation of degenerative mutations rather than by the fixation of new beneficial mutations. Next, we present several examples, including original data from the zebrafish *engrailed* genes, consistent with the new model. Finally, we suggest a series of experimental approaches for testing the new model.

Problems with the classical model for the preservation of gene duplicates: Under the simplest model for the fate of duplicate genes (the double-recessive model), the rate at which nonfunctional genes (genes that do not make a functional protein product) become fixed in populations is largely determined by random genetic drift and the null mutation rate (u), provided the product of the effective population size and u is <0.01. Under these conditions, the frequency of individuals homozygous null at both duplicate loci is negligible, and null mutations behave essentially as neutral alleles. The probability that one copy will become nonfunctional is then $\sim 1 - e^{-2ut}$, where t is the number of generations since the two loci have been functionally diploid with respect to meiosis (Nei and Roychoudhury 1973; Takahata and Maruyama 1979; Li 1980; Watterson 1983). This result suggests that most gene duplicates should become nonfunctional with high probability in a relatively short period of time. For example, if u is 10^{-6} per generation, then the mean time to nonfunctionalization is on the order of a few million generations or less.

Three general observations involving species derived from polyploidization events appear to contradict the rapid demise of gene duplicates predicted by the classical model. First, in numerous cases, the fraction of genes preserved is higher than predicted by the classic model. For example, in tetraploid fish lineages, 30–75% of the duplicate protein-coding genes have avoided nonfunctionalization for time spans on the order of 50 to 100 million yr (Allendorf *et al.* 1975; Ferris and Whitt 1979); in *Xenopus laevis*, about half of all duplicate genes have been preserved for 30 million yr (Bisbee *et al.* 1977; Graf and Kobel 1991; Hughes and Hughes 1993); and for the allopolyploidization event in maize, an annual plant, 72% have avoided nonfunctionalization for 11 million yr (Whitkus *et al.* 1992; Ahn and Tanksley 1993; White and Doebley 1998). The fact that most loci observed in these lineages appear to have a nonfunctional member in some related tetraploid species argues against the idea that both duplicate genes are retained due to constraints imposed by gene dosage requirements, at least for the enzyme loci investigated. Although the highest levels of duplicate gene retention in some fish and plant lineages may be due to the incomplete transition to diploid inheritance, similar estimates of duplicate gene preservation have emerged for more ancient polyploidization events for which disomic inheritance has clearly been reestablished, such as duplications that preceded the origin of tetrapods (33%; Nadeau and Sankoff 1997). Second, in *X. laevis*, which became tetraploid ~30 mya, nucleotide substitution patterns are consistent with the action of purifying selection on both copies of the duplicated genes (Hughes and Hughes 1993). Third, for loci that have avoided nonfunctionalization in both duplicate copies, there seems to be a relative paucity of null alleles segregating in extant populations (Ferris and Whitt 1977). Such observations are unexpected for loci involved in an ongoing degenerative process, and suggest the possibility that the duplicate loci are stabilized in populations.

Several attempts have been made to explain the high rate of duplicate gene preservation found by empirical observation. First, surviving duplicate loci in these taxa may have been preserved because new gene functions evolve at a much higher rate than predicted. We are not aware, however, of any convincing evidence that the majority of duplicate copies have acquired new functions that did not already exist in the ancestral genes (Ferris and Whitt 1979). A second possible explanation is that long-term effective population sizes may have been larger than expected, in fact large enough so that selection against double homozygotes prevents the fixation of null alleles at either locus (Takahata and Maruyama 1979; Li 1980; Walsh 1995). This appears not to account for the case of *X. laevis* (Hughes and Hughes 1993). The population size requirements for the preservation of gene duplicates by selection against double nulls over hundreds of millions of years may be prohibitively extreme. A third possible explanation for the discrepancy between theory and observation is that the rate of gene loss has been slowed by some form of natural selection against heterozygous carriers of null alleles (Bailey *et al.* 1978; Takahata and Maruyama

1979; Li 1980; Hughes and Hughes 1993; Clark 1994; Nowak *et al.* 1997).

Gene structure and duplicate gene preservation: An alternative reason for the failure of the classical model to explain the fates of most duplicate loci may be an overly simplistic view of gene structure. Although a general assumption of the classical model is that the properties of a gene may be adequately subsumed under a single function, genes often have several functions, each of which may be controlled by different DNA regulatory elements (see the following reviews for a number of examples: Piatigorsky and Wistow 1991; Hughes 1994; Kirchhamer *et al.* 1996; Arnone and Davidson 1997). A case in point is the *cut* locus in *Drosophila melanogaster* (Jack 1985; Liu *et al.* 1991; Jack and DeLotto 1995). Genetic and molecular analyses demonstrate that a 120-kb region of DNA upstream of the *cut* promoter drives tissue-specific expression, and that many spontaneous recessive mutant alleles result from insertions of transposable elements into this region. The regulatory mutation alleles fall into five complementation classes, with varying effects on tissue-specific expression (in Malpighian tubules, spiracles, central nervous system, specific portions of wing and leg imaginal discs, and embryonic and adult external sensory organs), as well as on morphology and viability. Similar complementation groups involving regulatory-region mutations are known for other developmental genes in *D. melanogaster*, including *cubitus interruptus* (Slusarski *et al.* 1995) and *Ultrabithorax* (Bender *et al.* 1983).

The widespread existence of complementation classes within eukaryotic gene loci indicates that gene expression patterns are typically controlled by multiple (and often modular and independent) regulatory regions associated with distinct protein-coding domains (Arnone and Davidson 1997). With the explicit assumption that these principles involving complementation between alleles at the same locus can be extended to complementation between two duplicate loci, we suggest that the regulatory complexity inherent in many gene classes is an essential, but previously missing, component of models for the evolutionary fate of duplicate genes. Further justification for this argument derives from substantial evidence showing spatial and temporal partitioning of expression patterns for gene duplicates in a wide variety of species (Ferris and Whitt 1979; Hughes and Hughes 1993; Ekker *et al.* 1995; Lee *et al.* 1996; Raff 1996; Gerhart and Kirschner 1997). To formally incorporate the issue of expression pattern complexity into models of gene duplication, we focus here on subfunctions that affect different gene expression domains during development. Here we adopt an operational definition of a subfunction as a specific subset of a gene's function that, when mutated, establishes a distinct complementation group, as in the *cut* example above (Liu *et al.* 1991; Jack and DeLotto 1995). A subfunction might involve the expression of a gene in a specific tissue, cell lineage, or developmental stage, or individual functional domains within the polypeptide coding portion of the gene.

The model presented below outlines how degenerative mutations in regulatory subfunctions can facilitate the preservation of duplicate genes, in the absence of any positive selection for beneficial mutations, by partitioning the repertoire of gene expression patterns of ancestral alleles. This model is quite distinct from the classical model, under which degenerative mutations can only lead to gene loss and beneficial mutations are the only route to gene preservation.

GENE PRESERVATION BY COMPLEMENTARY DEGENERATIVE MUTATIONS (SUBFUNCTIONALIZATION)

Following a polyploidization event, genomic redundancies exist at several levels: duplicate chromosomes, duplicate genes, and duplicate regulatory regions driving gene expression. Each level of redundancy is subject to processes of mutation and random genetic drift, which can lead to loss of function by chromosome loss, gene inactivation, or loss of individual regulatory elements. If duplicate chromosomes lose different genes, then for the organism to remain viable, the two chromosomes must complement each other by jointly retaining functional copies of all genes present on the original ancestral chromosome. Likewise, if duplicate genes lose different regulatory subfunctions, then they must complement each other by jointly retaining the full set of subfunctions present in the original ancestral gene. We refer to the complementary loss of duplicate genetic elements by degenerative mutation as the duplication-degeneration-complementation (DDC) process. The unique feature that distinguishes the DDC process from the classical model is that degenerative mutations facilitate rather than hinder the preservation of duplicate functional genes. In the following discussion, we focus on duplications of entire chromosomes or genomes rather than tandem gene duplications because we wish to exclude for now complications caused by uncertainty about the extent of the original duplication and local homogenization events caused by unequal crossing over or gene conversions (Zhou and Li 1996).

Under the general DDC model, the process of duplicate gene evolution occurs in two phases (Figure 1). During phase I, genes may experience one of three alternative fates, the first two of which correspond to outcomes under the classical model. First, one copy may incur a null mutation in the coding region, which subsequently drifts to fixation, leading to gene loss (nonfunctionalization). Nonfunctionalization can also occur if all of the regulatory regions of one duplicate are destroyed. Second, one copy may acquire a mutation conferring a new function, which becomes fixed through positive Darwinian selection (neofunctionaliza-

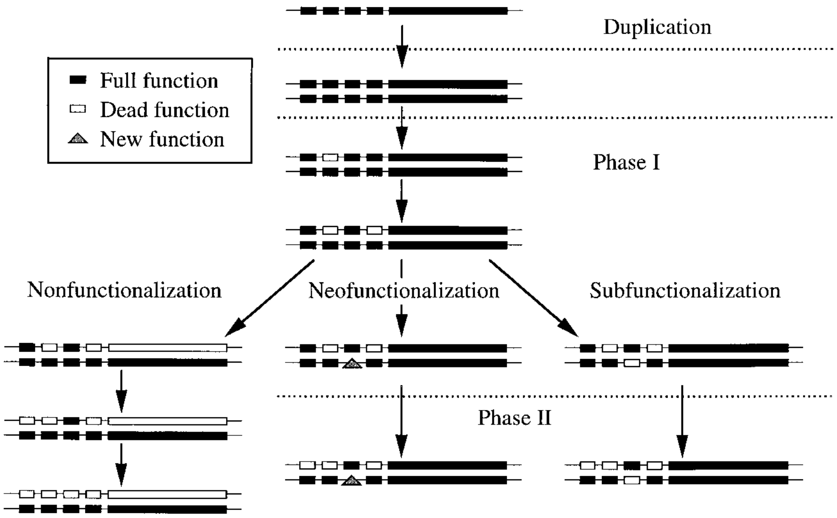

Figure 1.—Three potential fates of duplicate gene pairs with multiple regulatory regions. The small boxes denote regulatory elements with unique functions, and the large boxes denote transcribed regions. Solid boxes denote intact regions of a gene, while open boxes denote null mutations, and triangles denote the evolution of a new function. Because the model focuses on mutations fixed in populations, the diagram shows the state of a single gamete. In the first two steps, one of the copies acquires null mutations in each of two regulatory regions. On the left, the next fixed mutation results in the absence of a functional protein product from the upper copy. Because this gene is now a nonfunctional pseudogene, the remaining regulatory regions associated with this copy eventually accumulate degenerative mutations. On the right, the lower copy acquires a null mutation in a regulatory region that is intact in the upper copy. Because both copies are now essential for complete gene expression, this third mutational event permanently preserves both members of the gene pair from future nonfunctionalization. The fourth regulatory region, however, may still eventually acquire a null mutation in one copy or the other. In the center, a regulatory region acquires a new function that preserves that copy. If the beneficial mutation occurs at the expense of an otherwise essential function, then the duplicate copy is preserved because it retains the original function.

tion). It is now thought that such mutations may often involve changes in regulatory regions (Grenier *et al.* 1997; Shubin *et al.* 1997; Palopoli and Patel 1998). Assuming this new function results in the loss of an essential ancestral function, neofunctionalization insures the preservation of the nonmutated copy. [In principle, neofunctionalization can also occur if one or both copies acquire a new regulatory region without altering existing subfunctions, as pointed out by Sidow (1996)]. Third, each duplicate may experience loss or reduction of expression for different subfunctions by degenerative mutations. In such a case, the combined action of both gene copies is necessary to fulfill the requirements of the ancestral locus (subfunctionalization). If this happens, then complementation of subfunctions between duplicate genes will preserve both partially degenerated copies. In phase II of the DDC model, duplicate genes preserved either by neofunctionalization or subfunctionalization undergo random resolution of persisting redundant subfunctions, as the accumulation of degenerative mutations eliminates each subfunction in one or the other copy.

Subfunctionalization can occur by two different routes: qualitative or quantitative. Under qualitative subfunctionalization, which we model below and illustrate in Figure 1, one duplicate copy goes to fixation for a complete loss-of-subfunction mutation, and the second locus subsequently acquires a null mutation for a different subfunction. In contrast, quantitative subfunctionalization results from the fixation of reduction-of-expression mutations in both duplicates. In this case, once the total regulatory efficiency of a subfunction in both copies has been reduced to a threshold level determined by organismal requirements, any further degradation of the subfunction from either copy may be opposed by purifying selection.

Mutations that cause subfunctions to degrade may occur by several mechanisms, including nucleotide substitutions, deletions, inversions, insertions of transposable elements, slippage / replication errors, and unequal crossing over between repeated transcription-factor binding sites. Transposable elements may generate many subfunctional alleles. For example, *P*, *copia*, and *gypsy* elements are known to be mutagenic when they insert into 5′ regions of Drosophila genes (Kidwell and Lisch 1997). Species with a recent history of polyploidization, for example, maize, appear to have such insertions commonly in the 5′ and 3′ regions of genes, whereas in species lacking a recent polyploidization event, such insertions are infrequent (White *et al.* 1994; Wessler *et al.* 1995). Such transposable element insertions, presumably in regulatory DNA, may be tolerated in recently evolved polyploid species because of the redundancy of their regulatory elements.

The probability of subfunctionalization: The arguments presented above suggest that the DDC process could make both gene duplicates essential, but can it account for the high levels of duplicate gene preservation observed in polyploid lineages? Here we consider a simple model that suggests that, with reasonable parameter values, the DDC process can account for a significant fraction of preserved duplicate genes.

Consider the situation in which both members of a recently duplicated gene have z independently mutable subfunctions, all of which are essential, at least in single copy, and all of which mutate at identical rates (u_r) to alleles lacking the relevant subfunction. Letting u_c be the rate at which null mutations arise in the coding region, the null mutation rate for the locus is then $u_c + zu_r$ per gene copy. We assume that conditions are such that one functional allele (of four possible allele copies) of a given duplicated gene pair is sufficient for wild-type function (the double recessive model), and that beneficial mutations are rare relative to degenerative mutations. Provided the product of population size and genic mutation rate is <0.01 (Li 1980), the frequency of double null homozygotes will be sufficiently low such that all allele frequencies will evolve in an effectively neutral manner. The rate of fixation of a mutation in a population will then be approximately equal to the rate of mutation at the level of the gene (Kimura 1983).

Now imagine that one of the duplicate gene copies experiences a fixation event. Assuming there is more than one subfunction, the probability that the gene survives this event (and does not become a pseudogene) is the total regulatory-region mutation rate divided by the total mutation rate for the two copies

$$\text{Prob (survival of first fixation event)} = \frac{zu_r}{u_c + zu_r}. \quad (1)$$

Following the elimination of one of the z subfunctions from the first gene copy, the second copy must maintain this subfunction, because complete loss of an essential subfunction from both duplicates would be lethal. Thus, the permissible mutation rate for the second copy becomes $(z - 1)u_r$. Additional null mutations can occur in the remaining $(z - 1)$ regulatory subfunctions or in the coding region in the partially degraded first copy. Therefore, the total rate (summed over both copies) for the second mutational event is $[u_c + 2(z - 1)u_r]$. The probability of subfunctionalization upon this second event, $P_{S,2}$, is equal to the probability that the coding regions have survived the first hit multiplied by the probability that the second mutation occurs in a complementary subfunction in the second copy,

$$P_{S,2} = \left(\frac{zu_r}{u_c + zu_r}\right)\left(\frac{(z-1)u_r}{u_c + 2(z-1)u_r}\right). \quad (2)$$

Following this logic, it can be seen that $(z - 1)$ distinct series of mutational events can lead to duplicate-gene preservation by subfunctionalization—the first two null mutations in regulatory regions may occur on different gene copies, two may initially occur on the same copy followed by a third on the second copy, three may initially occur on the same copy followed by a fourth on the second copy, and so on. The probability of each of these additional pathways to subfunctionalization, *i.e.*, $(i - 1)$ consecutive regulatory-region null mutations on one copy followed by one on the other, is given by the generalization of Equation 2,

$$P_{S,i} = \left(\frac{zu_r}{u_c + zu_r}\right)\prod_{j=0}^{i-2}\left(\frac{(z-j-1)u_r}{u_c + 2(z-j-1)u_r}\right). \quad (3)$$

The total probability of gene preservation by subfunctionalization, P_S, is obtained by summing this quantity over $i = 2$ to z,

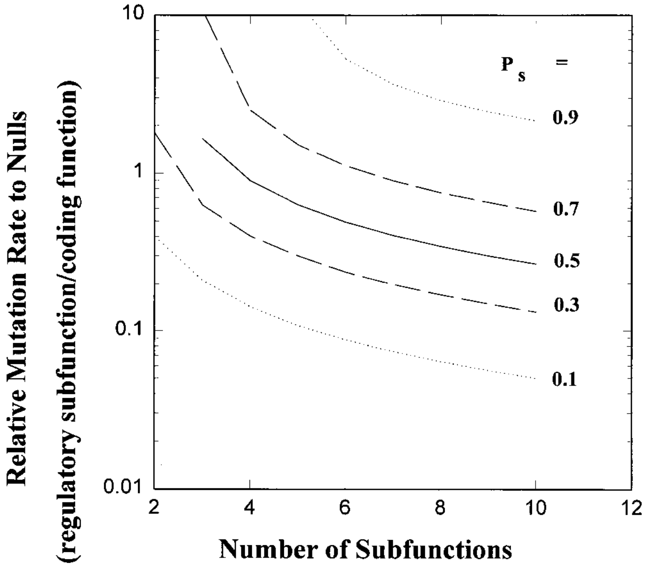

Figure 2.—Combinations of relative null mutation rates to regulatory and coding regions (u_r/u_c) and number of subfunctions (z) that yield various probabilities (P_S) of duplicate gene preservation by the DDC process. The probablility of duplicate gene preservation increases with the number of regulatory elements.

$$P_S = \sum_{i=2}^{z} P_{S,i}, \quad (4)$$

and the probability of nonfunctionalization is equal to $1 - P_S$. From this expression, we see that the probability of duplicate-gene preservation increases with the number of regulatory regions and with the mutation rate per regulatory region (Figure 2). More regulatory regions provide more targets for subfunctionalization that can be hit without penalty, and an increasing mutation rate per subfunction reduces the relative probability of fixation of a null mutation in the coding region before complementation.

The DDC process leads to subfunctionalization with high probability given reasonable parameter values. For example, if there are five subfunctions and the mutation rate per subfunction is 10% of the coding region null rate, then the probability of subfunctionalization is 0.1, and if the mutation rate per subfunction is 30% that of the null rate, then the probablitity of subfunctionalization is 30% (Figure 2). Generally, if the total rate of subfunctional mutations (zu_r) exceeds the null rate in the coding region by more than approximately fourfold, then the probability of gene preservation by subfunctionalization exceeds 50%. The complexity and size of regulatory regions of eukaryotic genes (Kirchhamer *et al.* 1996; Arnone and Davidson 1997) suggests that these conditions may be met frequently.

Time scales for subfunctionalization and resolution:

Using the model presented above, the mean time to gene preservation conditional on its actual occurrence can be obtained by treating the times to mutational events as geometrically distributed variables. The rate of occurrence of an initial regulatory-region null mutation is $2zu_r$, because each of the two copies contains z mutational targets. As noted above, subsequent to this initial event, zero to $(z - 2)$ additional degenerative mutations may be incurred by the first-hit copy before the first mutation on the opposite copy. The mean time to subfunctionalization conditional on the occurrence of $(i - 1)$ consecutive regulatory-region null mutations on one copy followed by one on the other is then

$$\bar{t}_{S,i} = \frac{1}{u_r}\left(\frac{1}{2z} + \sum_{j=1}^{i-1}\frac{1}{z-j}\right). \tag{5a}$$

The mean time to subfunctionalization is then

$$\bar{t}_S = \sum_{i=2}^{z}\frac{\bar{t}_{S,i}\,P_{S,i}}{P_S}. \tag{5b}$$

As in the classical model, these expressions indicate that the fates of duplicate genes are generally determined in a relatively short period (on an evolutionary time scale; Figure 3A). For example, if $u_r = 10^{-7}$/yr, $\bar{t}_S$ is on the order of 4 million yr or less provided the number of regulatory regions is greater than five, and even with $z < 5$ it does not exceed 12.5 million yr. Thus, under the DDC model, most duplicate genes that are destined to be preserved by subfunctionalization are expected to become so within a few million years. With a regulatory-region mutation rate x times that in the figure, the mean time to subfunctionalization would be divided by x.

Unless there are only two initial regulatory regions, some regulatory regions (as many as $z - 2$) will likely remain to be resolved over evolutionary time after the initial subfunctionalization event. The fraction of regulatory regions that is expected to be resolved at the time of gene preservation by subfunctionalization is

$$P_r(0) = \sum_{i=2}^{z}\frac{iP_{S,i}}{zP_S}. \tag{6}$$

This fraction depends only weakly on the ratio of coding-region to regulatory-region mutation rates, and is <0.5 if the number of regulatory regions exceeds five (Figure 3B). Thus, we anticipate that after the preservation of duplicate genes by the DDC process, a substantial fraction of regulatory subfunctions will typically remain to be resolved in phase II. Assuming that the occurrence of mutations that destroy regulatory regions is a Poisson process, for any site that is unresolved at the time of gene preservation, the probability that it is still unresolved after t further time units is simply $P_0(t) = e^{-2tu_r}$. The number of unresolved sites at time t then follows a binomial distribution with parameter $P_0(t)$.

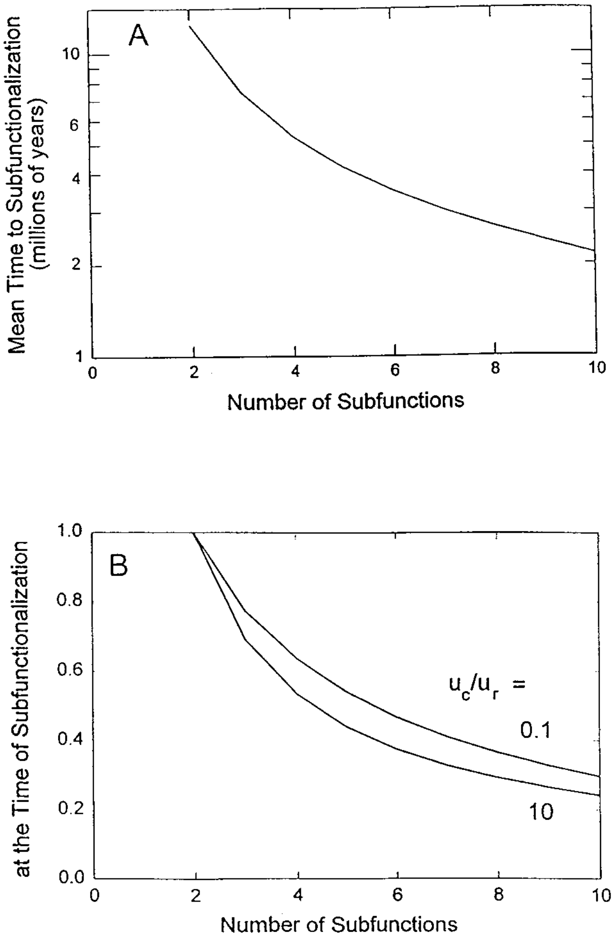

Figure 3.—(A) The mean time to subfunctionalization as a function of the number of subfunctions (z), for the situation in which $u_r = 10^{-7}$/yr. (B) The fates of gene pairs are determined in a short time, on an evolutionary scale, and the ratio of the mutation rate in regulatory and coding regions is a weak determinant of the expected degree of resolution at the time of subfunctionalization.

The molecular nature of subfunctions and the preservation of genetic redundancy: The preceding theory assumes that individual regulatory subfunctions are independently mutable, with single mutations being sufficient to eliminate a subfunction. Under this simple scenario, the various subfunctions within duplicate genes preserved by the DDC process are expected to be resolved randomly, with each copy retaining about half of its subfunctions within the limits of binomial sampling. However, while we define subfunctions by their mutational properties such that they are members of distinct complementation classes, this definition does not describe how such subfunctions are arranged on the DNA molecule. Regulatory regions for different subfunctions are often partially overlapping or embedded, leading to the situation where the number of expression domains exceeds the number of complementation groups (Jack and DeLotto 1995; Kirchhamer *et al.* 1996). Some of the central issues are illustrated in Figure

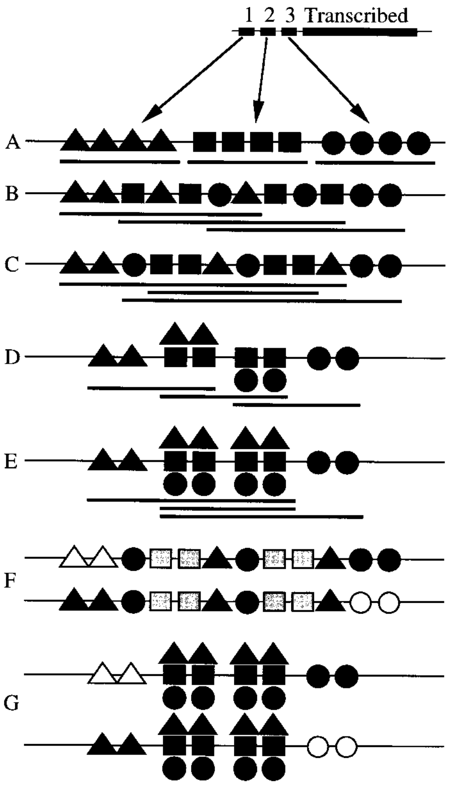

Figure 4.—Overlapping and embedded regulatory elements. All transcription-factor binding sites shown are assumed to be essential for each expression domain. (A) Independent regulatory regions with independent transcription-factor binding sites. (B) Overlapping regulatory regions with independent sites. (C) Overlapping and embedded regulatory regions with independent sites. (D) Overlapping regulatory regions with shared sites. (E) Embedded regulatory regions with shared sites. (F) Resolution of overlapping regulatory regions with independent sites (derived from C) leading to quantitative resolution of regulatory region 2 after 1 and 3 are destroyed on paralogous copies. (G) Resolution of embedded regulatory regions with shared sites leading to true redundancy for regulatory region 2 after 1 and 3 are destroyed on paralogous copies.

4. The situation modeled above is equivalent to a setting in which the spatial arrangement of transcription-factor binding sites allows the independent resolution of the subfunctions ($z = 3$ in Figure 4A). Within overlapping (Figure 4, B and D) or embedded (Figure 4, C and E) regulatory regions, the transcription-factor binding sites may be either interdigitated and acting independently (Figure 4, B and C) or shared, with the same DNA binding site(s) being required for more than one expression domain (Figure 4, D and E). In these cases, some mutational events can knock out more than one expression domain at a time, effectively reducing the number of subfunctions.

Complexities involving the physical arrangement of regulatory regions on the DNA may help explain, without invoking positive selection, how the same expression domains may be preserved by both gene duplicates (Pickett and Meeks-Wagner 1995). In Figure 4C, for example, regulatory regions 1 and 3 are partially overlapping and regulatory region 2 is embedded in the region of overlap. If in this situation, one gene duplicate suffers a degenerative mutation in the nonoverlapping portion of regulatory region 1, and the other gene duplicate experiences a degenerative mutation destroying the nonoverlapping portion of regulatory region 3, then regulatory region 2 will be sheltered from mutation events that destroy two or more adjacent binding sites because such a mutation in that region would completely eliminate subfunction 1 or 3 (Figure 4F). Resolution of subfunction 2 may then proceed only by small-scale mutational events that partially degrade the quantitative functionality of individual transcription-factor binding sites. If on the other hand, the binding sites for regulatory region 2 are completely shared with both regulatory regions 1 and 3 (Figure 4, E and G), then mutational events that inactivate regions 1 and 3 on opposite copies will indefinitely preserve the embedded regulatory region 2 in both copies. Thus, the resolution of overlapping and/or embedded subfunctions provides a potential explanation for the maintenance of some of the redundancy observed between duplicate genes without invoking positive selection for duplicate gene expression (Nowak *et al.* 1997).

The topology of regulatory regions may also help explain unidirectional and bidirectional divergence of gene duplicates observed by Ferris and Whitt (1979). For some enzyme-encoding loci, gene duplicates show unidirectional divergence—the enzyme products of one locus predominate in all tissues in which the two loci are expressed. This situation would be most common at gene loci with overlapping and/or embedded arrangements of regulatory regions, because the elimination of the embedded element from one gene copy (element 2 in Figure 4E, for example) would prevent the fixation of alleles in which the overlapping regulatory elements (elements 1 and 3 in Figure 4E) are destroyed on the other gene duplicate. Under bidirectional divergence, either locus may be more highly expressed in any given tissue. This situation should be most frequent for gene loci with independently arranged regulatory regions.

DDC and dosage effects: In some situations, gene dosage requirements might increase the probability that both gene duplicates are preserved. The theoretical model developed above assumes that for each subfunction, activity of only one of the four alleles of the two gene duplicates is sufficient for survival. It is possible, however, that after gene duplication some subfunctions must remain intact in more than one of the four alleles to ensure optimal fitness. For instance, consider a gene with three separate subfunctions. After duplication, the first subfunction may be sufficient for survival if intact in a single allele, the second subfunction may be sufficient in two alleles, but the third subfunction might be required in three of the four alleles. In such a case, the first and second subfunctions could be resolved to either duplicate gene by the principles of DDC. The third subfunction, however, would have to be maintained by both gene duplicates to have at least three active alleles. In such cases, dosage requirements would provide the initial gene preservation mechanism, but complementary loss of other subfunctions or acquisition of a new function could reinforce the initial preservation event. Note that this type of dosage effect provides an alternative mechanism to shared embedded elements (Figure 4) for retaining a specific expression domain by both gene duplicates.

In some cases, gene dosage requirements might cause the partitioning of subfunctions to be favored by positive selection. For example, consider a situation in which activity of all four alleles of a duplicated gene pair in a certain tissue or time is deleterious. In such a case, the fixation of a nonfunctional or subfunctional allele might be accelerated by positive selection. Note that a case like this differs from the formal model proposed above, which assumes that drift and purifying selection is usually sufficient for the fixation of subfunctional alleles. In these cases, mutations of subfunctions that would be deleterious in the single-copy genes before duplication would become beneficial after duplication. Because this might increase the rate of fixation of subfunctional alleles while simultaneously increasing the rate of fixation of nonfunctional alleles, the overall effect on the probability of duplicate gene preservation is not clear. Future experimental and modeling work may help to define these more complex interactions between gene dosage, population size, the mutation rates to subfunctional, coding null, and neofunctional alleles, and the roles of purifying and positive selection in duplicate gene preservation. It is hoped that the near-neutral DDC model provided here can act as a null hypothesis for testing these and other more complex possibilities.

Possible examples of the DDC process: Here we present several possible examples of the general DDC process and the way in which it can account for observed patterns of duplicate gene expression. Additionally, we suggest experiments that could falsify the DDC model as an explanation for these specific cases. We consider here a pair of duplicate *engrailed* genes in zebrafish and the *ZAG1* and *ZMM2* gene pair in maize. Analysis of such cases must identify gene duplicates, determine whether they arose by tandem duplication or by duplication of large chromosome regions, infer ancestral functions of the unduplicated parent gene, and finally determine whether the distribution of gene functions between duplicated genes can be explained by the complementary sharing of ancestral functions or only by the acquisition of novel functions.

Engrailed genes in zebrafish: Tetrapods have two members of the *engrailed* gene family, called *En1* and *En2* (Joyner and Martin 1987; Gardner and Barald 1992). Zebrafish, however, has been reported to have three *engrailed* genes (*eng1*, *eng2*, and *eng3*; Ekker *et al.* 1992), and here we report additional sequence and expression data of a fourth *eng* gene that we recently cloned, *eng1b* (Amores *et al.* 1998). Phylogenetic analysis rooted with *engrailed* genes of amphioxus (Holland *et al.* 1997) and a new lamprey engrailed gene, *enga*, shows that *eng1*/*eng1b* and *eng2*/*eng3* are ancient duplicate gene pairs (Figure 5). The tree shows that *eng1* and *eng1b* are both orthologous to tetrapod *En1*, and that *eng2* and *eng3* are both orthologous to tetrapod *En2*. The phylogenetic tree shows that the duplication event(s) that produced the *eng1*/*eng1b* and *eng2*/*eng3* gene pairs occurred in the lineage of zebrafish after it diverged from the lineage of tetrapods. The duplication that produced the two main clades of vertebrate *engrailed* genes occurred before that divergence.

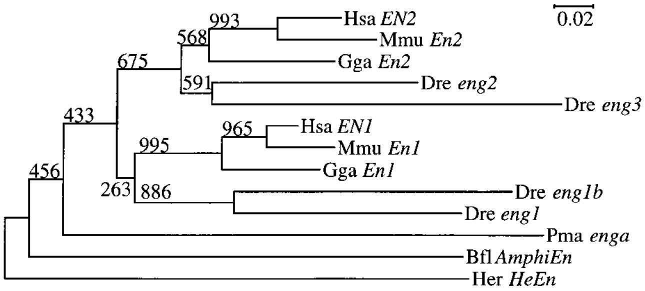

Figure 5.—Phylogeny of *engrailed* family genes. Sequences were aligned by eye with the assistance of CLUSTAL X, and only the unambiguous portions of the alignment were employed in the final analysis. Phylogenetic tree based on the neighbor-joining method (Saitou and Nei 1987) with multiple substitution correction. Numbers are bootstrap values based on 1000 runs. Lamprey *enga* was cloned by the polymerase chain reaction using two primers [forward, 5′ATI AA(A G) ATI TGG TT(T C) CA(A G) AA(T C) AA3′; and reverse, 5′TG(G A) TT(G A) TAI A(G A)I CC(T C) TGI GCC AT3′] designed to amplify the conserved Engrailed amino acid sequences IKIWFQNK and MAQGLYNH, respectively. The primers were used to screen a lamprey cosmid library constructed from a single ammocete larva according to the Stratagene SuperCos cloning kit. The same primers were used to obtain additional sequence from zebrafish eng1b (Amores *et al.* 1998). Species abbreviations: Hsa, *Homo sapiens*; Mmu, *Mus musculus*; Gga, *Gallus gallus*; Dre, *Danio rerio*; Pma, *Petromyzon marinus*; Bfl, *Branchiostoma floridae*; Her, Heliocidaris erythrogramma. GenBank accession nos.: Hsa *EN2*, J03066; Mmu *En2*, L12705; Gga *En2*, L12697; Dre *eng2*, X68446; Dre *eng3*, X68447; Hsa *EN1*, L12699; Mmu *En1*, Y00201, M11987; Gga *En1*, L12695; Dre *eng1b*, AF071237; Dre *eng1*, X68445; Pma *enga* AF129401; Bfl *AmphiEn*, U82487; Her *HeEn*, U58775.

To determine whether the zebrafish *eng* gene pairs originated in chromosome-scale duplications or local tandem duplications, we mapped the *eng1b* locus and compared it to the genome locations of other *engrailed* genes in mammals and zebrafish (Logan *et al.* 1989; Amores *et al.* 1998; Postlethwait *et al.* 1998). The results showed that the zebrafish genes *eng1* and *eng1b* are linked to *dlx2* and *dlx5* on two different chromo-

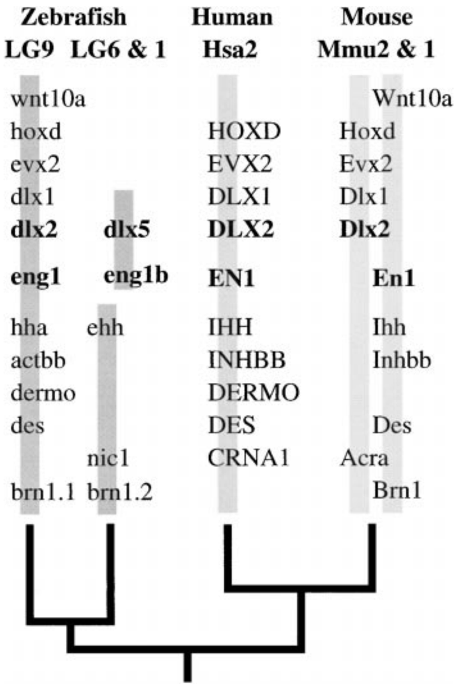

Figure 6.—Syntenic relationships of *engrailed-1* genes. Zebrafish appears to have two copies of a single chromosome segment present in the last common ancestor of teleosts and mammals, one copy with *eng1* and *dlx2* and the other copy with *eng1b* and *dlx5*. One copy of this segment has been broken by a chromosome rearrangement, and appears in LG6 and LG1. In humans, as in zebrafish LG9, this segment is in a single chromosome arm, Hsa2q, but in mouse, this segment is on two chromosomes. Orthologous genes are on the same horizontal line. Syntenies are displayed, but gene orders are ignored to ease the comparisons of orthologues.

somes (Figure 5). Prior analysis had shown that the gene duplicates *dlx2* and *dlx5* are both zebrafish orthologues of tetrapod *DLX2* (Stock *et al.* 1996; Ellies *et al.* 1997), which is linked to *EN1* on human chromosome 2q (Ozcelik *et al.* 1992). Likewise, *eng2* and *eng3* genes are linked to the gene duplicates *shh* and *twhh* on two different zebrafish chromosomes (Postlethwait *et al.* 1998), and both of these zebrafish *hedgehog* genes are orthologous to tetrapod *SHH* (Zardoya *et al.* 1996), which is linked to *EN2* on human chromosome 2q (Logan *et al.* 1989). These mapping results show that zebrafish has two copies of large chromosome segments surrounding the zebrafish *engrailed* genes, and that these syntenic regions are conserved with mammalian genomes (Figure 6). We conclude that the two gene pairs *eng1*/*eng1b* and *eng2*/*eng3* arose by duplication events that involved large chromosome sections, consistent with their origin in a genome duplication event in ray-finned fish (Amores *et al.* 1998; Postlethwait *et al.* 1998).

Note that two independent data sets, gene phylogenies based on sequence information and chromosomal locations based on genetic mapping data, concur that the tetrapod *En1* gene is an outgroup to the two zebrafish duplicates *eng1*/*eng1b*. Therefore, *En1* can be used as an outgroup to infer the ancestral shared expression domains of *eng1* and *eng1b*.

Although the expression patterns of *engrailed* genes are complex, here we focus on expression patterns of the *engrailed-1* gene family in two groups of cells. Zebrafish *eng1* is expressed in the pectoral appendage bud, while *eng1b* is expressed in a specific set of neurons in the hindbrain/spinal cord (Figure 7). Is either of these expression domains due to neofunctionalization? Or were both present in the progenitor gene before duplication and one domain lost by each duplicate? Examining the most recent unduplicated outgroup would allow one to infer the state of the ancestral gene. In the absence of information from the most recent outgroup, tetrapods can provide appropriate data. In mouse and chicken, *En1* is expressed in both expression domains, the developing pectoral appendage bud, and in specific neurons of the hindbrain and spinal cord (Joyner and Martin 1987; Davis *et al.* 1991; Gardner and Barald 1992). Therefore it appears that following the duplication event that produced *eng1* and *eng1b* in the ray-finned lineage, the *eng1* ancestor retained the pectoral appendage bud expression and lost the hindbrain/spinal cord neuron expression, while the *eng1b* ancestor lost the pectoral appendage bud expression and retained the hindbrain/spinal cord neuron expression as hypothesized in Figure 7.

Is this a case of gene preservation by subfunctionalization? These data suggest complementary loss of expression, which is consistent with the DDC model. A definitive test of this hypothesis will require identification of the regulatory elements responsible for these expression domains in zebrafish, fish that share the *eng1*/*eng1b* duplication, fish that diverged from the lineage giving rise to zebrafish before the duplication event, and tetrapods, including mouse and chicken. In zebrafish, there appear to be many examples similar to *engrailed*, including duplicates of *msx* genes (Ekker *et al.* 1997), *nkx* genes (Lee *et al.* 1996), and *hedgehog* genes (Ekker *et al.* 1995).

ZAG1 and ZMM2 in maize: As a second possible example of the preservation of gene duplicates by subfunctionalization, consider the duplicate genes known as *ZAG1* and *ZMM2* in the maize genome, which originated via an allotetraploidization event between two closely related grasses about 11 mya (Goodman *et al.* 1980; Wendel *et al.* 1986; Helentjaris *et al.* 1988; Gaut and Doebley 1997; White and Doebley 1998). *ZAG1* and *ZMM2* are apparent orthologues of the single-copy floral homeotic genes known as *AGAMOUS* in Arabidopsis and as *PLENA* in Antirrhinum, both of which are expressed in carpels and stamens (Yanofsky *et al.* 1990; Coen and Meyerowitz 1991; Bradley *et al.* 1993). *ZAG1* is expressed at high levels throughout maize carpel development, but it is expressed at only low levels in stamen primordia (Mena *et al.* 1996). In contrast,

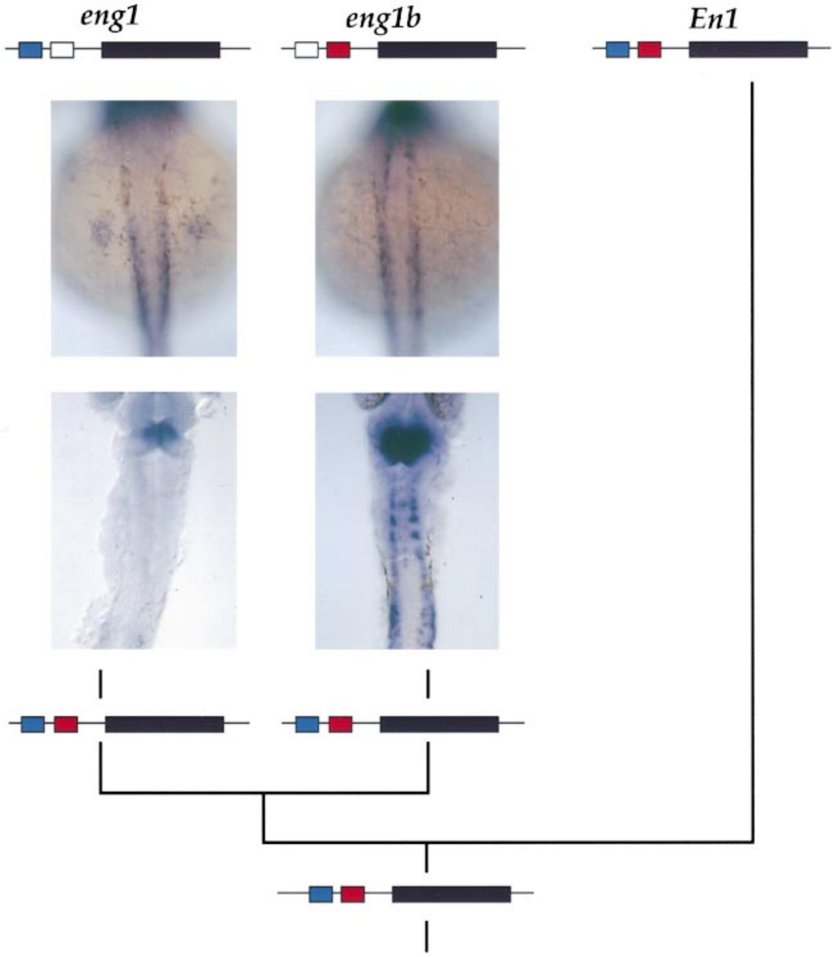

Figure 7.—Expression pattern of zebrafish *eng1* and *eng1b*. Expression of *eng1* (left) and *eng1b* (right) transcript in 28.5-hr zebrafish embryos. Dorsal view of trunk showing the pectoral appendage bud (top), which expresses *eng1*, and dorsal view of head (bottom) showing the expression of *eng1b* in a specific set of hindbrain and spinal neurons. *In situ* hybridization was performed according to Jowett *et al.* (1995) and Thisse *et al.* (1993). The *eng1b* probe was derived from the 3′ untranslated region by the polymerase chain reaction using using two primers (forward, GCTCATGGCT CAAGGACTCTA; and reverse, AACATTG GACTAAACGTAAAACTT) and cloned into the TA cloning vector (Invitrogen, San Diego), while the *eng1* probe was a gift from Monte Westerfield. The figure shows a hypothesized evolutionary scenario for the tissue-specific patterns of expression of the *engrailed* family members in zebrafish on the left and in tetrapods on the right. Solid and open boxes indicate full expression and lack of expression, respectively. Blue boxes represent the pectoral appendage regulatory element and red boxes indicate the neuron regulatory element.

ZMM2 is highly expressed in maize stamens but not at all in the immature carpel, although it increases in late female flower development (Figure 8). The phenotype of a *ZAG1* null mutant is limited to early carpel development, suggesting that late *ZMM2* expression rescues later carpel development in *ZAG1* mutants.

The DDC model can explain these data by suggesting that the ancestral genes to *ZAG1* and *ZMM2* were both expressed strongly in the developing stamens and carpels in the allotetraploid ancestor of maize shortly after the polyploidization event, as *AGAMOUS* and *PLENA* are today in Arabidopsis and Antirrhinum. We hypothesize that reciprocal regulatory mutations in the *ZAG1*/*ZMM2* duplicates complemented each other, thereby preserving both genes that exist in today's maize. After the allotetraploidization event, degenerative regulatory mutations decreased the expression of *ZAG1* in stamens but not in carpel, while other regulatory mutations eliminated the expression of *ZMM2* in the early carpel but not in the stamens. If this hypothesis is correct, then, maize plants doubly homozygous for *ZMM2* and *ZAG1* null mutations should produce plants that phenocopy *AGAMOUS* and *PLENA* mutants in Arabidopsis and Antirrhinum. In addition, molecular analysis of the promoters of this gene family in maize, its close relative sorghum, Arabidopsis, and Antirrhinum should identify conserved regulatory elements that became partitioned after gene duplication.

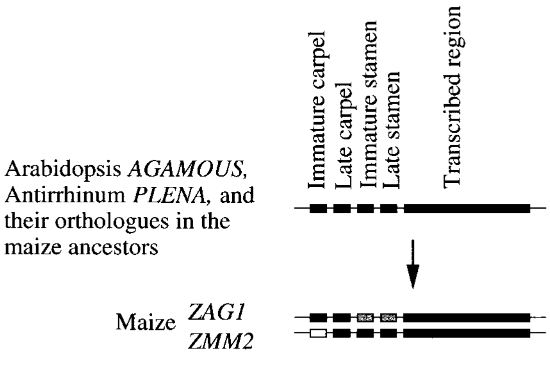
Immature carpel
Late carpel
Immature stamen
Late stamen
Transcribed region
Arabidopsis AGAMOUS, Antirrhinum PLENA, and their orthologues in the maize ancestors
Maize
ZAG1
ZMM2

Figure 8.—The subfunctionalization hypothesis to explain the expression of duplicated genes in maize. Small solid boxes indicate active regulatory elements, gray boxes indicate regulatory elements with reduced activity, and open boxes indicate null activity alleles for the regulatory element.

Hoxa1 and Hoxb1 in mouse: A third possible example of DDC in duplicate genes involves the *Hoxa1* and *Hoxb1* genes in mouse (Figure 9). These genes reside in duplicate *Hox* clusters, groups of closely linked genes that

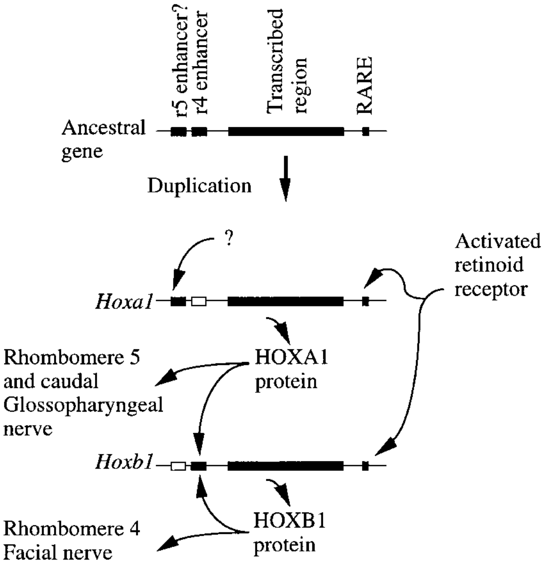

Figure 9.—Can the DDC hypothesis explain the expression of duplicated *Hox* genes in mouse? Small solid boxes indicate a conserved RARE enhancer, medium boxes indicate active rhombomere regulatory elements, and large boxes indicate the transcribed coding regions. The "r5 enhancer?" is an uncharacterized element that is expected to drive expression from the r3/r4 border, through r5 to more caudal segments, and also to be necessary for glossopharyngeal nerve development. Open boxes indicate hypothetical dead regulatory elements. The r5 enhancer element in *Hoxa1* has not yet been identified but is predicted to be present.

encode a family of DNA-binding proteins that specifies fate along the anterior-posterior axis of bilaterian animals (Lewis 1978; Krumlauf 1994). The four tetrapod *Hox* clusters arose from two sequential duplication events, probably whole genome duplications, before the divergence of ray-finned and lobe-finned fishes 420 mya (Duboule and Dollé 1989; Holland and Garcia-Fernandez 1996; Amores *et al.* 1998; Postlethwait *et al.* 1998). The *Hoxa* and *Hoxb* clusters are probably sisters derived from the second duplication event (Kappen and Ruddle 1993; Zhang and Nei 1996; Amores *et al.* 1998). Thus, for this case, entire genes and all their regulatory elements were duplicated, which might not be true of a tandem duplication.

Hoxb1 and *Hoxa1* cooperate to pattern anterior ectodermal and mesodermal derivatives of vertebrate embryos (Figure 9). *Hoxa1* is important for segment identity in rhombomere 5 (r5) of the hindbrain and for the development of the glossopharyngeal nerve as well as more caudal rhombomeres (Dupé *et al.* 1997; Studer *et al.* 1998). Analogously, *Hoxb1* is important for specifying the identity of rhombomere 4 (r4) and development of the facial nerve (Pöpperl *et al.* 1995; Gavalas *et al.* 1998). A 70-bp-long element, called the r4 autoregulatory enhancer, is conserved between mouse, chick, and pufferfish (Pöpperl *et al.* 1995). This element functions to elevate the expression of *Hoxb1* and restrict it to rhombomere 4. Judging from expression and functional analysis, an analogous element probably drives the early expression of *Hoxa1* in rhombomere 5 and more caudal segments, but this element has yet to be defined molecularly. The DDC model predicts that the unduplicated ancestor of these two genes possessed regulatory elements that drove expression of this *Hox* gene in both late r4 and early r5 expression domains, and that after duplication, these two subfunctions resolved divergently between the two paralogous copies (Figure 9).

In addition to the independent roles of *Hoxa1* and *Hoxb1* just discussed, these two genes have early redundant roles, including expression in broadly overlapping territories and activation of some of the same downstream targets (Studer *et al.* 1994, 1996; Dupé *et al.* 1997; Maconochie *et al.* 1997; Chen and Ruley 1998). Some of the shared roles are important for their independent roles. Both genes contain a retinoic acid response element (RARE) in their 3′ region that initiates expression of the adjacent gene in neuroectoderm and mesoderm (Langston *et al.* 1997; Figure 9). The *Hoxa1* RARE is required not only for *Hoxa1* expression in rhombomere 5, but also for the proper level of *Hoxa1* expression in the somitic and presomitic mesoderm (Thompson *et al.* 1998). The requirement of a single conserved regulatory element (the RARE) for expression in different domains (rhombomere 5 and mesoderm) is a property expected of embedded regulatory elements as diagrammed in Figure 4, D, E, and G. The DDC model predicts that when subfunctions with shared embedded regulatory elements become divergently resolved, then redundant, overlapping expression domains will be retained, as observed for the *Hoxa1*/*Hoxb1* duplicated gene pair in portions of the neuroectoderm and the mesoderm.

What experiments can distinguish whether the current subfunctions of murine *Hoxa1* and *Hoxb1* duplicates arose by subfunctionalization, neofunctionalization, or some other model? A critical issue is whether the r4 and r5 enhancers were present in the ancestral gene before duplication. One can infer the state of the ancestral gene by examining an outgroup that diverged from the lineage of tetrapods just before the duplication event that produced the *Hoxa1* and *Hoxb1* genes. The lamprey might be such an outgroup (Pendleton *et al.* 1993; Amores *et al.* 1998; Carr *et al.* 1998) and experiments should be conducted to see if it is, and if so, to determine whether the corresponding functions were likely to have already been present in the ancestral gene before duplication.

In summary, the examples discussed above provide data that are consistent with the DDC model, and in some cases are more readily explained by the DDC model than the classical model. Further experiments need to be done to firmly establish which route of duplicate gene preservation was employed in each case.

Testing the DDC and classical models: As we noted earlier, even the most basic premise of the classical model of duplicate gene evolution—that gene duplicates are preserved only by the evolution of new functions—has never been tested. Because deleterious mutations are much more common than beneficial mutations, we believe that the DDC process provides a reasonable (and parsimonious) alternative explanation for at least some cases of long-term preservation of gene duplicates. Unlike the classical model, the mutational mechanisms that lead to gene preservation by DDC are distinct from those responsible for the origin of new gene functions. On the other hand, by expanding the time period for which genes are exposed to selection, the preservation of duplicates by the DDC process facilitates subsequent opportunity for the evolution of new functions. If the evolution of new gene functions is the *only* mechanism of duplicate gene preservation, then it should be possible to empirically reject our alternative subfunctionalization hypothesis. We now consider some potentially fruitful avenues for future research.

1. Phylogenetic analysis: The subfunctionalization model predicts that the sum of subfunctions in preserved gene duplicates will be equal to the total subfunctions in the ancestral gene. This prediction is clearly distinct from the position of the classical model, which suggests that gene preservation is dependent upon the acquisition of new *cis*-regulatory regions driving novel expression patterns during development (Sidow 1996; Cooke *et al.* 1997). To test these alternative hypotheses, the evolutionary time frame must be short enough to preclude the possibility that genes initially preserved by subfunctionalization will have also subsequently acquired new functions. This then requires the analysis of recently preserved duplicates on a cladogram that also allows the inference of ancestral expression patterns from appropriate outgroups. For example, to explain the derivation of the triplicate Drosophila genes *paired*, *gooseberry*, and *gooseberry-neuro*, which have conserved protein function but distinct developmental functions, Li and Noll (1994) suggested that following duplication "genes may acquire new functions by changes in their regulatory regions generating an altered expression" without considering the possibility that these three genes simply result from the differential loss of subsets of the expression domains of the ancestral gene. A phylogenetic analysis of closely related outgroup species with single gene copies would distinguish between the classical and DDC models.
2. Mutation rate to subfunctional alleles: The simple subfunctionalization model discussed here requires that the total subfunction mutation rate relative to the total null mutation rate be on the order of 0.3 or larger to achieve at least a 10% probability of duplicate gene preservation, and on the order of 0.7 or larger to achieve at least a 50% probability of gene preservation (Table 1). If the relative rate of formation of subfunctional alleles is not within this range, then subfunctionalization as modeled is unlikely to be a major mechanism of duplicate gene preservation. Experiments must be designed to measure the mutation rate to subfunctional, neofunctional, and nonfunctional alleles to test critically the various models. If empirical studies demonstrate that the rate of mutation to subfunctional alleles is too low relative to the rate of coding null mutations, then this particular subfunctionalization model is falsified.
3. Regulatory region complexity: The subfunctionalization model predicts that the probability of gene preservation should be higher for more complex genes (with larger numbers of subfunctions), particularly for genes in which the regulatory regions for subfunctions are spatially independent on the DNA because more complex genes provide more targets for subfunction mutations. Testing this prediction requires the molecular characterization of regulatory regions for various genes in species with duplicated genes and comparison with closely related species without the duplications.
4. Multiple polyploidization events: The subfunctionalization model makes specific predictions about the probability of duplicate gene preservation after closely spaced polyploidization events, such as those thought to have occurred early in vertebrate phylogeny (Holland *et al.* 1994; Holland and Garcia-Fernandez 1996; Nadeau and Sankoff 1997; Amores *et al.* 1998). The DDC model suggests that duplicate loci preserved by subfunctionalization after the first polyploidization event (Figure 1, right) will have fewer subfunctions than the parent locus before duplication (Figure 1, top). Therefore, because theory predicts that the likelihood of preservation depends on the number of subfunctions (Figure 2), the probability that both duplicate loci will be preserved after the second round of duplication is diminished relative to the first polyploidization event. If, on the

TABLE 1

Critical values of $zu_r/(u_c + zu_r)$ required for specific probabilities of subfunctionalization (P_S), given for different numbers of subfunctions (z) obtained from Equations 3 and 4

	P_S		
z	0.1	0.3	0.5
2	0.448	0.774	1.000
3	0.384	0.654	0.832
5	0.351	0.598	0.759
10	0.332	0.57	0.728
∞	0.316	0.548	0.708

other hand, a single duplicate survives the first round with all of the original subfunctions intact (Figure 1, left), then after the second round of duplication, the probability of duplicate preservation will be approximately the same as in the first event. If the level of gene preservation does not change between polyploidization events, then subfunctionalization is an unlikely explanation for the preservation of duplicate genes. Data from the *HOX* complexes of vertebrates suggest that the level of duplicate preservation does indeed decline with subsequent duplication events. Assuming the (AB)(CD) model of *HOX* cluster duplication (Kappen and Ruddle 1993; Zhang and Nei 1996; Amores *et al.* 1998), then 11 of 14 gene pairs (13 *HOX* cluster genes plus *EVX*; 79%) were both preserved after the first duplication (Amores *et al.* 1998). After the second duplication, in the lineage leading to mouse, 16 of 25 gene pairs (64%) were preserved, while after the third duplication, which occurred in the ray-finned fish lineage leading to zebrafish (Amores *et al.* 1998), only 13 of 45 gene pairs (29%) have been preserved to the present in duplicate copies. With the caveat that a few genes may remain to be discovered in the zebrafish, and that all genes still present in duplicate copy at the time of the second and third duplication events had been permanently preserved by some mechanism, the results conform at least to the qualitative expectations of the DDC model.

5. Population size: Finally, because the subfunctionalization model assumes that selection against mutant alleles is negligible, the predictions for this model should be fulfilled most closely in populations of relatively small size, because under these conditions, the incidence of double null homozygotes and the appearance of beneficial mutations is minimal. On the other hand, because neofunctionalization relies entirely on rare beneficial mutations, it should be a more common mechanism of gene preservation in species with large effective sizes.

CONCLUSION

Because it focuses on the regulatory complexity of genes and the preservational nature of degenerative mutations, the DDC model provides a different perspective on the evolutionary consequences of gene duplications than that of the classical model, which focuses primarily on the nonfunctionalizing properties of degenerative mutations in coding regions and on the neofunctionalizing properties of beneficial mutations. The DDC model organizes a rather disparate collection of observations and principles on gene structure, and we believe that most of these principles are established firmly enough that subfunctionalization must be taken seriously as a testable model for explaining the evolutionary fate of duplicate genes.

We thank Bruce Appel, Susan Bassham, Ian Boussy, Bruce Draper, Greg Gibson, Bill Jackman, Chuck Kimmel, Craig Miller, Cecilia Moens, Katherine Swan, Han Wang, and Monte Westerfield for critical comments on an early draft of this manuscript and James Langeland for a larval *Petromyzon marinus*. This research evolved out of a collaborative effort fostered in part by a National Science Foundation Research Training Grant in Genetic Mechanisms of Evolution and a National Institutes of Health Training Grant in Developmental Biology, which supported A.F., by National Institutes of Health grants to M.L. (R01-GM36827) and J.P. (P01-HD22486 and R01-RR10715), and a U.S. Department of Agriculture grant to F.B.P. (97-35304-4658).

LITERATURE CITED

Ahn, S., and S. D. Tanksley, 1993 Comparative linkage maps of the rice and maize genomes. Proc. Natl. Acad. Sci. USA **90:** 7980–7984.

Allendorf, F. W., F. M. Utter and B. P. May, 1975 Gene duplication within the family Salmonidae: II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations, pp. 415–432 in *Isozymes*, edited by C. L. Markert. Academic Press, New York.

Amores, A., A. Force, Y.-L. Yan, L. Joly, C. Amemiya *et al.*, 1998 Zebrafish *hox* clusters and vertebrate genome evolution. Science **282:** 1711–1714.

Arnone, M. I., and E. H. Davidson, 1997 The hardwiring of development: organization and function of genomic regulatory systems. Development **124:** 1851–1864.

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.

Bender, W., M. Akam, F. Karch, P. A. Beachy, M. Peifer *et al.*, 1983 Molecular genetics of the Bithorax complex in *Drosophila melanogaster*. Science **221:** 23–29.

Bisbee, C. A., M. A. Baker, A. C. Wilson, H. A. Irandokht and M. Fischberg, 1977 Albumin phylogeny for clawed frogs (*Xenopus*). Science **195:** 785–787.

Bradley, D., R. Carpenter, H. Sommer, N. Hartley and E. Coen, 1993 Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of Antirrhinum. Cell **72:** 85–95.

Carr, J. L., C. S. Shashikant, W. J. Bailey and F. H. Ruddle, 1998 Molecular evolution of *Hox* gene regulation: cloning and transgenic analysis of the lamprey *HoxQ8* gene. J. Exp. Zool. **280:** 73–85.

Chen, J., and H. E. Ruley, 1998 An enhancer element in the *EphA2* (*Eck*) gene sufficient for rhombomere-specific expression is activated by HOXA1 and HOXB1 homeobox proteins. J. Biol. Chem. **273:** 24670–24675.

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

Coen, E. S., and E. M. Meyerowitz, 1991 War of the whorls: genetic interactions controlling flower development. Nature **353:** 31–37.

Cooke, J. M. A., M. A. Nowak, M. Boerlijst and J. Maynard Smith, 1997 Evolutionary origins and maintenance of redundant gene expression during metazoan development. Trends Genet. **13:** 360–364.

Davis, C. A., D. P. Homyard, K. J. Millen and A. L. Joyner, 1991 Examining pattern formation in mouse, chicken and frog embryos with an *En*-specific antiserum. Development **2:** 287–298.

Duboule, D., and P. Dollé, 1989 The structural and functional organization of the murine HOX gene family resembles that of *Drosophila* homeotic genes. EMBO J. **8:** 1497–1505.

Dupé, V., M. Davenne, J. Brocard, P. Dollé, M. Mark *et al.*, 1997 In vivo functional analysis of the *Hoxa-1* 3′ retinoic acid response element (3′RARE). Development **124:** 399–410.

Ekker, M., J. Wegner, M.-A. Akimenko and M. Westerfield, 1992 Coordinate expression of three zebrafish *engrailed* genes. Development **116:** 1001–1010.

Ekker, S. C., A. R. Ungar, P. Greenstein, D. P. vonKessler, J. A.

Porter *et al.*, 1995 Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. Curr. Biol. **5:** 944–955.

Ekker, M., M. Akimenko, M. Allende, R. Smith, G. Drouin *et al.*, 1997 Relationships among *msx* gene structure and function in zebrafish and other vertebrates. Mol. Biol. Evol. **14:** 1008–1022.

Ellies, D., D. Stock, G. Hatch, G. Giroux, K. Weiss *et al.*, 1997 Relationship between the genomic organization and the overlapping embryonic expression patterns of the zebrafish *dlx* genes. Genomics **45:** 580–590.

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 the differential regulation of duplicate genes after polyploidization. J. Mol. Evol. **12:** 267–317.

Gardner, C. A., and K. F. Barald, 1992 Expression patterns of engrailed-like proteins in the chick embryo. Dev. Dyn. **193:** 370–388.

Gaut, B. S., and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809–6814.

Gavalas, A., M. Studer, A. Lumsden, F. Rijli, R. Krumlauf *et al.*, 1998 *Hoxa1* and *Hoxb1* synergize in patterning the hindbrain, cranial nerves and second pharyngeal arch. Development **125:** 1123–1136.

Gerhart, J., and M. Kirschner, 1997 *Cells, Embryos, and Evolution*. Blackwell Science, Malden, MA.

Goodman, M. M., C. W. Stuber, K. Newton and H. H. Weissinger, 1980 Linkage relationships of 19 enzyme loci in maize. Genetics **96:** 697–710.

Graf, J. D., and H. R. Kobel, 1991 *Xenopus laevis:* practical uses in cell and molecular biology, pp. 19–34 in *Methods in Cell Biology*, edited by B. K. Kay and H. B. Peng. Academic Press, New York.

Grenier, J. K., T. L. Garber, R. Warren, P. M. Whitington and S. Carroll, 1997 Evolution of the entire arthropod *Hox* gene set predated the origin and radiation of the onychophoran/arthropod clade. Curr. Biol. **7:** 547–553.

Haldane, J. B. S., 1933 The part played by recurrent mutation in evolution. Am. Nat. **67:** 5–9.

Helentjaris, T., D. Weber and S. Wright, 1988 Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. Genetics **118:** 353–363.

Holland, L. Z., M. Kene, N. A. Williams and N. D. Holland, 1997 Sequence and embryonic expression of the amphioxus engrailed gene (*AmphiEn*): the metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila.* Development **124:** 1723–1732.

Holland, P. W., and J. Garcia-Fernandez, 1996 *Hox* genes and chordate evolution. Dev. Biol. **173:** 382–395.

Holland, P. W. H., J. Garcia-Fernandez, N. A. Williams and A. Sidow, 1994 Gene duplications and the origins of vertebrate development. Development (Suppl.), 125–133.

Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. Ser. B Biol. Sci. **256:** 119–124.

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

Jack, J., and Y. DeLotto, 1995 Structure and regulation of a complex locus: the cut gene of *Drosophila.* Genetics **139:** 1689–1700.

Jack, J. W., 1985 Molecular organization of the cut locus of *Drosophila melanogaster.* Cell **42:** 869–876.

Jowett, T., M. Mancera, A. Amores and Y.-L. Yan, 1995 In situ hybridization to embryo whole mounts and tissue sections: mRNA detection and application to developmental studies, pp. 91–121 in *In situ Hybridization*, edited by M. Clark. Chapman & Hall, Weinheim, Germany.

Joyner, A. L., and G. R. Martin, 1987 *En-1* and *En-2*, two mouse genes with sequence homology to the *Drosophila engrailed* gene: expression during embryogenesis. Genes Dev. **1:** 29–38.

Kappen, C., and F. Ruddle, 1993 Evolution of a regulatory gene family: *HOM/HOX* genes. Curr. Opin. Genet. Dev. **3:** 931–938.

Kidwell, M. G., and D. Lisch, 1997 Transposable elements as sources of variation in animals. Proc. Natl. Acad. Sci. USA **94:** 7704–7711.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kirchhamer, C. V., C.-H. Yuh and E. H. Davidson, 1996 Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc. Natl. Acad. Sci. USA **93:** 9322–9328.

Krumlauf, R., 1994 Hox genes in vertebrate development. Cell **78:** 191–201.

Langston, A. W., J. R. Thompson and L. J. Gudas, 1997 Retinoic acid-responsive enhancers located 3′ of the *Hox A* and *Hox B* homeobox gene clusters. Functional analysis. J. Biol. Chem. **272:** 2167–2175.

Lee, K. H., Q. H. Xu and R. E. Breitbart, 1996 A new *tinman*-related gene, *nkx2.7*, anticipates the expression of *nkx2.5* and *nkx2.3* in zebrafish heart and pharyngeal endoderm. Dev. Biol. **180:** 722–731.

Lewis, E. B., 1978 A gene complex controlling segmentation in Drosophila. Nature **276:** 565–570.

Lewis, W. H., 1979 *Polyploidy: Biological Relevance.* Plenum, New York.

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, X., and M. Noll, 1994 Evolution of distinct developmental functions of three *Drosophila* genes by acquisition of different cis-regulatory regions. Nature **367:** 83–87.

Liu, S., E. McLeod and J. Jack, 1991 Four distinct regulatory regions of the cut locus and their effect on cell type specification in *Drosophila.* Genetics **127:** 151–159.

Logan, C., H. F. Willard, J. M. Rommens and A. L. Joyner, 1989 Chromosomal localization of the human homeobox-containing genes, *EN1* and *EN2.* Genomics **4:** 206–209.

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.

Lynch, M., and J. B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Maconochie, M. K., S. Nonchev, M. Studer, S. K. Chan, H. Pöpperl *et al.*, 1997 Cross-regulation in the mouse *HoxB* complex: the expression of *Hoxb2* in rhombomere 4 is regulated by *Hoxb1.* Genes Dev. **11:** 1885–1895.

Mena, M., B. A. Ambrose, R. B. Meeley, S. P. Briggs, M. F. Yanofsky *et al.*, 1996 Diversification of C-function activity in maize flower development. Science **274:** 1537–1540.

Morizot, D. C., S. A. Slaugenhaupt, K. D. Kallman and A. Chakravarti, 1991 Genetic linkage map of fishes of the genus *Xiphophorus* (Teleostei: Poeciliidae). Genetics **127:** 399–410.

Nadeau, J. H., and D. Sankoff, 1997 Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147:** 1259–1266.

Nei, M., and A. K. Roychoudhury, 1973 Probability of fixation of nonfunctional genes at duplicate loci. Am. Nat. **107:** 362–372.

Nowak, M. H., M. C. Boerlijst, J. Cooke and J. Maynard Smith, 1997 Evolution of genetic redundancy. Nature **388:** 167–171.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.

Ozcelik, T., M. H. Porteus, J. L. R. Rubenstein and U. Francke, 1992 *DLX2* (*TES1*), a homeobox gene of the Distal-less family, assigned to conserved regions on human and mouse chromosomes 2. Genomics **13:** 1157–1161.

Palopoli, M. F., and N. H. Patel, 1998 Evolution of the interaction between *Hox* genes and a downstream target. Curr. Biol. **8:** 587–590.

Pébusque, M.-J., F. Coulier, D. Birnbaum and P. Pontarotti, 1998 Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol. Biol. Evol. **15:** 1145–1159.

Pendleton, J. W., B. K. Nagai, M. T. Murtha and F. H. Ruddle, 1993 Expansion of the *Hox* gene family and the evolution of chordates. Proc. Natl. Acad. Sci. USA **90:** 6300–6304.

Piatgorsky, J., and G. Wistow, 1991 The recruitment of crystallins: new functions precede gene duplication. Science **252:** 1078–1079.

Pickett, F. B., and D. R. Meeks-Wagner, 1995 Seeing double: appreciating genetic redundancy. Plant Cell **7:** 1347–1356.

Pöpperl, H., M. Bienz, M. Studer, S. Chan, S. Aparicio *et al.*, 1995 Segmental expression of *Hoxb-1* is controlled by a highly conserved autoregulatory loop dependent upon *exd*/*pbx*. Cell **81:** 1031–1042.

Postlethwait, J., Y. Yan, M. Gates, S. Horne, A. Amores *et al.*, 1998 Vertebrate genome evolution and the zebrafish gene map [see comments]. Nat. Genet. **18:** 345–349.

Raff, R. A., 1996 *The Shape of Life*. University of Chicago Press, Chicago, IL.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Seoighe, C., and K. H. Wolfe, 1998 Extent of genomic rearrangement after genome duplication in yeast. Proc. Natl. Acad. Sci. USA **95:** 4447–4452.

Shubin, N., C. Tabin and S. Carroll, 1997 Fossils, genes and the evolution of animal limbs. Nature **388:** 639–648.

Sidow, A., 1996 Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genet. Dev. **6:** 715–722.

Slusarski, D. C., C. K. Motsny and R. Holmgren, 1995 Mutations that alter the timing and pattern of *cubitus interruptus* gene expression in *Drosophila melanogaster*. Genetics **139:** 229–240.

Stock, D. W., D. L. Ellies, Z. Y. Zhao, M. Ekker, F. H. Ruddle *et al.*, 1996 The evolution of the vertebrate *Dlx* gene family. Proc. Natl. Acad. Sci. USA **93:** 10858–10863.

Studer, M., H. Pöpperl, H. Marshall, A. Kuroiwa and R. Krumlauf, 1994 Role of conserved retinoic acid response element in rhombomere restriction of *Hoxb-1*. Science **265:** 1728–1732.

Studer, M., A. Lumsden, L. Ariza-McNaughton, A. Bradley and R. Krumlauf, 1996 Altered segmental identity and abnormal migration of motor neurons in mice lacking *Hoxb-1*. Nature **384:** 630–634.

Studer, M., A. Gavalas, H. Marshall, L. Ariza-McNaughton, F. M. Rijli *et al.*, 1998 Genetic interactions between *Hoxa1* and *Hoxb1* reveal new roles in regulation of early patterning. Development **125:** 1025–1036.

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.

Thisse, C., B. Thisse, T. F. Schilling and J. H. Postlethwait, 1993 Structure of the zebrafish *snail1* gene and its expression in wild-type, *spadetail* and *no tail* mutant embryos. Development **119:** 1203–1215.

Thompson, J. R., S. W. Chen, L. Ho, A. W. Langston and L. J. Gudas, 1998 An evolutionary conserved element is essential for somite and adjacent mesenchymal expression of the *Hoxa1* gene. Dev. Dyn. **211:** 97–108.

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.

Wendel, J. F., C. W. Stuber, M. D. Edwards and M. M. Goodman, 1986 Duplicated chromosome segments in *Zea mays L.*: further evidence from hexokinase isozymes. Theor. Appl. Genet. **72:** 178–185.

Wessler, S. R., T. E. Bureau and S. E. White, 1995 LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. **5:** 814–821.

White, S., and J. Doebley, 1998 Of genes and genomes and the origin of maize. Trends Genet. **14:** 327–332.

White, S. E., L. F. Habera and S. R. Wessler, 1994 Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and function. Proc. Natl. Acad. Sci. USA **91:** 11792–11796.

Whitkus, R., J. Doebley and M. Lee, 1992 Comparative genome mapping of Sorghum and maize. Genetics **132:** 1119–1130.

Wolfe, K. H., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708–713.

Yanofsky, M. F., H. Ma, J. L. Bowman, G. N. Drews, K. A. Felmann *et al.*, 1990 The protein encoded by the *Arabidopsis* homeotic gene *Agamous* resembles transcription factors. Nature **346:** 35–39.

Zardoya, R., E. Abouheif and A. Meyer, 1996 Evolutionary analyses of *hedgehog* and *Hoxd-10* genes in fish species closely related to the zebrafish. Proc. Natl. Acad. Sci. USA **93:** 13036–13041.

Zhang, J., and M. Nei, 1996 Evolution of Antennapedia-class homeobox genes. Genetics **142:** 295–303.

Zhou, Y. H., and W. H. Li, 1996 Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. Mol. Biol. Evol. **13:** 780–783.

Communicating editor: A. G. Clark