Journal of Heredity 2009:100(5):605–617
doi:10.1093/jhered/esp047
Advance Access publication July 13, 2009

Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates

MATTHEW W. HAHN

From the Department of Biology and School of Informatics, 1001 E. 3rd Street, Indiana University, Bloomington, IN 47405.

Address correspondence to M. Hahn at the address above, or e-mail: mwh@indiana.edu.

Abstract

Determining the evolutionary forces responsible for the maintenance of gene duplicates is key to understanding the processes leading to evolutionary adaptation and novelty. In his highly prescient book, Susumu Ohno recognized that duplicate genes are fixed and maintained within a population with 3 distinct outcomes: neofunctionalization, subfunctionalization, and conservation of function. Subsequent researchers have proposed a multitude of population genetic models that lead to these outcomes, each differing largely in the role played by adaptive natural selection. In this paper, I present a nonmathematical review of these models, their predictions, and the evidence collected in support of each of them. Though the various outcomes of gene duplication are often strictly associated with the presence or absence of adaptive natural selection, I argue that determining the outcome of duplication is orthogonal to determining whether natural selection has acted. Despite an ever-growing field of research into the fate of gene duplicates, there is not yet clear evidence for the preponderance of one outcome over the others, much less evidence for the importance of adaptive or nonadaptive forces in maintaining these duplicates.

Key words: *Adaptation, DDC, duplication, paralogs, positive selection*

In his widely cited but rarely read classic, Susumu Ohno (1970) made the first serious case for the importance of gene duplication in evolution. Although a number of earlier geneticists recognized the power gene duplicates held in allowing organisms to functionally diversify (reviewed in Taylor and Raes 2004), Ohno was the first to both gather the evidence on duplication and outline the various evolutionary fates of duplicated genes. Since his time, evidence for the ubiquity of gene duplication and its role in biological adaptation has only increased: Every bacterial and eukaryotic genome sequenced has revealed high numbers of paralogous genes (Zhang 2003) and careful studies of individual cases have uncovered a variety of apparently adaptive functions carried out by paralogous duplicates (e.g., Piatigorsky et al. 1988; Zhang et al. 1998; Hittinger and Carroll 2007).

Despite the progress that has been made in identifying gene duplicates, the evolutionary mechanisms responsible for the initial maintenance of a duplicated gene remain undetermined in the vast majority of cases. In fact, the ambiguity of many of the evolutionary mechanisms proposed for gene duplicate maintenance has led to confusion over what exact hypothesis is being tested. Often the long-term fate of duplicates (over dozens of millions of years) is described without an analysis of the proximate mechanisms that originally maintained the duplicates, which is the main question to be answered (Force et al. 1999). For a duplicate to be maintained means that loss of the duplicate results in a fitness cost. Adding to this confusion is what I view as a conceptual deficiency in the field: We have not been clear in distinguishing evolutionary outcomes from the population genetic models of mechanisms that result in these outcomes. Though Ohno was not a population geneticist, he very clearly outlined the 3 major outcomes of gene duplication (Figure 1): 1) the evolution of a new function in one of the duplicates ("neofunctionalization," a term actually coined by Force et al. [1999]), 2) the division of ancestral functions among duplicates ("subfunctionalization"), and 3) the conservation of all functions in both duplicates ("gene conservation"). These outcomes are distinct from the models that have subsequently been proposed as the underlying processes responsible for them, and only by distinguishing between these 2 levels can we begin to distinguish among the different possible histories of duplicated genes.

Numerous researchers have proposed methods for distinguishing among the various population genetic models, using both sequence and functional data. But due to ambiguity of the proposed models, incomplete alternative hypotheses, and weak power of most of the tests (or all 3

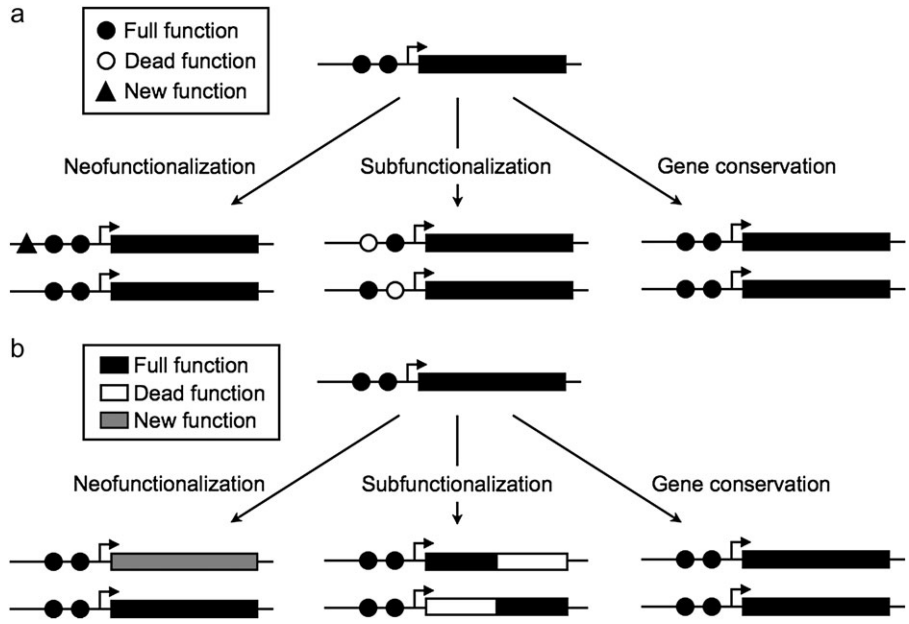

Figure 1. Outcomes of duplication that maintain the new copy. In each panel, the 3 outcomes of gene duplication are shown, either by (**a**) regulatory sequence changes or by (**b**) coding sequence changes.

together), the interpretations of many of these studies are questionable. Many of the tests proposed to distinguish among mechanisms are instead tests for the action of adaptive natural selection, which can be consistent with a number of different outcomes and models. Contrary to multiple claims in the literature, it is my opinion that there is no convincing evidence for the preponderance of one evolutionary outcome over another, much less the prevalence of a single population genetic model for reaching that outcome. In this review, I attempt to describe the different models for the maintenance of duplications and the predictions each makes about the evolutionary histories of gene duplicates. I also stress throughout the prescience displayed by Ohno in anticipating many of the ideas that have found new life in the modern genomic era.

Molecular Mechanisms of Gene Duplication

The molecular mechanisms responsible for the duplication of genetic material have been reviewed many times (e.g., Li 1997; Lynch 2007a), but it is worthwhile briefly discussing them in light of models for the maintenance of gene duplicates in order to clarify terminology. There appear to be 4 major mechanisms by which DNA is duplicated: 1) unequal crossing-over, 2) duplicative (DNA) transposition, 3) retrotransposition, and 4) polyploidization. Though the precise contribution of each type of duplication to any single genome is generally not known, estimates of the frequency of each can be made based largely on the locations of paralogs across the genome.

Unequal Crossing-Over

Unequal crossing-over ("tandem duplication") appears to be a common contributor of new genetic material. Estimates from *Arabidopsis thaliana*, *Mus musculus*, *Rattus norvegicus*, *Homo sapiens*, and *Saccharomyces cerevisiae* put the number of tandemly arrayed duplicates between 10% and 20% of all genes, though the exact meaning of "tandem" can differ in each paper (Arabidopsis Genome Initiative 2000; Drouin 2002; Shoja and Zhang 2006). The number in *Caenorhabditis elegans* appears to be much higher: Almost 70% of new gene duplicates are located directly next to each other (Katju and Lynch 2003), but more than half of these are not in the same orientation (duplicates are expected to be in the same orientation with unequal crossing-over). How these numbers correspond to the frequency of unequal crossing-over is unclear for 2 reasons. First, they may be underestimates of the contribution of crossing-over as initially tandem genes are moved farther apart by the insertion of DNA between them. Second, they may be overestimates of crossing-over because although it is clear how unequal crossing-over results in an increase (and decrease) of copy number when there are already multiple paralogs in tandem, it is less clear how this mechanism acts to duplicate only a single gene—with only one copy there is no opportunity for mispairing during meiosis. It may be that other repeated sequences surrounding a single gene can facilitate this process, but it is not known how commonly this occurs. Third, there appears to be yet another mechanism for producing closely linked duplicates via strand switching of the replication machinery (Lee et al. 2007). This mechanism will result in tandem duplicates that do not necessarily have head-to-tail orientations and may therefore explain newly duplicated genes in

opposite orientations without having to invoke unequal crossing-over and subsequent inversions.

Duplicative (DNA) Transposition

Duplicative transposition of DNA sequences can be accomplished by 1 of 2 main pathways: nonallelic homologous recombination (NAHR) or nonhomologous end joining (NHEJ; reviewed in Paques and Haber 1999). The difference in the 2 pathways is largely based on whether homologous sequences are used as a template during double-strand break repair, and this difference can also be used to infer the mechanism by which individual genes are duplicated (unequal crossing-over is a form of NAHR, albeit involving closely linked sequences). Bailey et al. (2003) found an enrichment of transposable elements at the junctions of interchromosomally duplicated sequences in humans, a pattern also recently found in *Drosophila melanogaster* (Fiston-Lavier et al. 2007). Recombination between these nonallelic homologous sequences can result in the duplication of the intervening sequences, which can then lead in turn to more duplications because of pairing between the new paralogs (Bailey et al. 2003). But other studies in humans have also found multiple cases with no repetitive DNA or long stretches of homologous sequence at duplication breakpoints, suggesting the action of NHEJ (Linardopoulou et al. 2005). Due to the relatively low proportion of duplicated sequences arranged in tandem in the human genome, it has been proposed that duplicative transposition (of one mechanism or another) is the major mode of duplication in humans (Samonte and Eichler 2002). Consistent with this, Friedman and Hughes (2004) found that almost two-thirds of young gene duplicates ($K_S << 1$) in both human and mouse are on different chromosomes; this is in contrast to the 89% of new duplicates found on the same chromosome in *C. elegans* (Katju and Lynch 2003) and the 96% found on the same chromosome among 12 *Drosophila* genomes (Heger and Ponting 2007), though different methodologies were used in these papers. The apparent large differences in mutational mechanism of duplication among these species are especially surprising given the similarity in rate and maintenance of duplicates (Lynch and Conery 2000; Hahn, Han, and Han 2007; Hahn et al. 2007a). As mentioned above, the fraction of paralogs arranged in tandem can be an underestimate of the contribution of unequal crossing-over. The fact that many paralogs lie on different chromosomes, however, argues that duplicative transposition (or retrotransposition, see below) is a major force in gene duplication among mammals.

Retrotransposition

Retrotransposed duplicate genes result from the reverse transcription of mRNA into cDNA that is then inserted into a new genomic position (Brosius 1991). Because reverse transcription uses fully processed mRNAs, the newly duplicated paralogs lack introns and have a poly-A tail. Retrotransposed duplicates (or "retrogenes") also do not bring any of the flanking noncoding DNA with them and consequently are much less likely to be expressed after duplication. Recent studies have found that retrogenes that land near other coding regions or even in the introns of expressed coding sequences are much more likely to be expressed than those that land far from coding sequences (Vinckenbosch et al. 2006). The number of retrogenes maintained in both mammals (Pan and Zhang 2007) and *Drosophila* (Hahn, Han, and Han 2007) is lower than the number maintained by DNA-based intermediates (i.e., unequal crossing-over and duplicative transposition), despite the fact that the mutation rate forming new retrocopies is higher (Pan and Zhang 2007; Jun et al. 2008). The lack of functional regulatory DNA is likely to be the reason that very few of these paralogs are maintained for long—only ~120 functional retrotransposed gene copies have been maintained in the human genome over the past 63 million years (Vinckenbosch et al. 2006). However, the asymmetry in duplication of flanking sequences between RNA-based and DNA-based mechanisms is not complete: Only about 50% of DNA-based duplications result in a new gene copy that contains all the original exons (Katju and Lynch 2003). It is therefore likely that many paralogs lack the flanking noncoding DNA necessary for transcription, regardless of the molecular mechanisms responsible for their duplication.

Polyploidy

The fourth major mechanism of duplicate gene formation is polyploidization. Whole-genome duplications result in new gene copies of every gene in a genome and, obviously, all the flanking regulatory sequences. Though every gene is duplicated, only 10–30% of all genes are maintained in the genome for very long (Byrne and Wolfe 2005; Maere et al. 2005; Paterson et al. 2006). The type or function of genes maintained after polyploidization appears to differ from those duplicated by smaller scale mechanisms: Many of the genes kept after whole-genome duplications exhibit dosage effects (reviewed in Birchler and Veitia 2007; Conant and Wolfe 2008). The possibility of this difference was noted by Ohno (1970, p. 98): "Concordant duplication of all gene loci creates no problem with regard to the dosage relationship of functionally related genes," as opposed to single gene duplication events that do not maintain dosage balance among interacting genes. Though an excess of duplicates in these categories have not necessarily held up in studies of additional taxa (e.g., Barker et al. 2008), polyploidy events are likely to have had a large impact on genome evolution and gene duplication overall.

Segmental Duplication

One last comment on the molecular mechanisms of gene duplicates is required. The term "segmental duplication" has recently become very popular, largely because of the excellent work of Evan Eichler and colleagues (Bailey et al. 2001, 2002; She et al. 2006). Though this term initially denoted duplications of large stretches of the genome (e.g., Birchler and Levin 1991), it has come to take on many

meanings. One can find the term referring to long duplications (i.e., containing more than one gene), any duplication event that is not due to polyploidization, any duplication longer than an arbitrary threshold, any duplication with paralogous sequences more than 90% similar, any DNA-based duplication event, or only duplicative transposition events (this last usage is likely due to the fact that many of the best studied cases in humans are due to duplicative transposition). It is therefore being used as both a description of the minimum length and age of a duplicate as well as a molecular mechanism of duplication. But the length criterion for identifying segmentally duplicated loci has slowly been shrinking due to improved experimental and computational methods, and this has begun to change the meaning of the term. For instance, a recent paper by Jiang et al. (2007) looks at segmental duplicates with a minimum length of 1 kb in humans and explicitly identifies retrotransposition as one of the contributors to segmental duplicates (because even single exons can be longer than 1 kb). As it is therefore no longer clear precisely how the term segmental duplication relates to either the mechanism or the length of a duplication, we should be mindful of how it is applied.

Outcomes and Models of Gene Duplication

As outlined in the Introduction, Ohno identified 3 main outcomes in the evolution of gene duplicates. Though his book is often characterized as stressing only neofunctionalization (again, not his original term), Ohno actually gives equal billing to all the outcomes in the book section entitled "Why Duplication?". Below I describe these outcomes and the population genetic models that detail the evolutionary paths to these endpoints (Figure 2). I attempt to stress the different predictions each model makes, without going into the mathematical details of each; a more focused review of the theoretical results can be found in Walsh (2003).

Gene Conservation ("Duplication for the Sake of Producing More of the Same" [Ohno 1970])

The first mechanism for maintaining a gene duplicate addressed by Ohno was to simply increase the number of genes coding for a protein. In this scenario, both loci maintain the original functions, and it has therefore come to be known as "gene conservation" (Zhang 2003). Multiple authors have also recently proposed that this is a major force in duplicate gene retention (Kondrashov et al. 2002; Kondrashov and Kondrashov 2006; Sugino and Innan 2006). Ohno proposed 2 possible models for why these duplicates would maintain the original functions, though they are not necessarily mutually exclusive.

Redundancy

The first model posits that a second gene could provide functional redundancy if the original locus was disabled by mutation (Figure 2). Multiple researchers have addressed this model and have shown theoretically that it is not likely to have played a large role in evolution because the strength of selection maintaining the duplicate would be on the order of the mutation rate to null alleles (Clark 1994; Lynch et al. 2001; O'Hely 2006). However, as pointed out by Lynch (2007a), this result does not imply that duplicates maintained by other forces cannot act as mutational buffers over time.

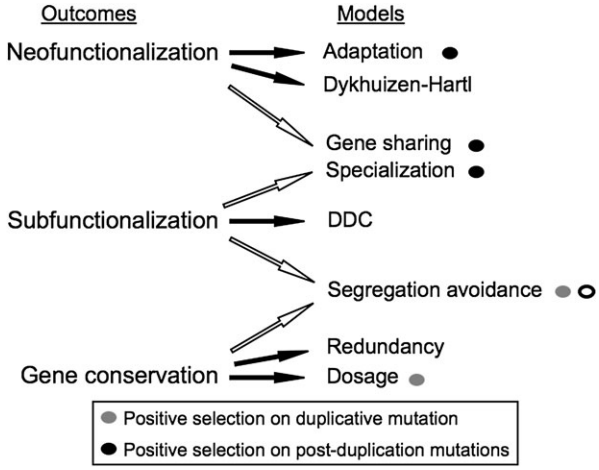

Figure 2. Outcomes and population genetic models of gene duplication. The 3 outcomes of gene duplication that maintain the new copy are shown, along with the population genetic models that have been proposed for each. Many of the models have been categorized into multiple outcomes and are linked to outcomes via unfilled arrows. Also shown are the proposed effects of positive (adaptive) natural selection on each model. The unfilled circle represents a misclassification of data from this model as positive selection on postduplication mutations.

Dosage

The second possibility for why exact copies of duplicated genes are maintained is that there is an advantage to producing more of a gene. Although it is certainly true that increased levels of protein production can be accomplished by increasing expression levels at a single locus, duplicating a gene may have an equivalent effect. According to Ohno (p. 59): "When the metabolic requirement of an organism dictates the presence of an enormous amount of a particular gene product, the incorporation of multiple copies of the gene locus by the genome often fulfills that requirement." The most commonly cited example of this phenomenon is the array of highly duplicated ribosomal RNAs needed for development and other translationally intensive stages. A pair of recent studies may also highlight the selective advantage of having more of the same gene. Perry et al. (2007) studied variation in the number of duplicates of the salivary amylase gene (*AMY1*) among humans. They found that human populations that consume starch-rich diets had on average more copies of *AMY1* per individual and that

this translated into higher protein levels and enhanced ability to break down starches. As these *AMY1* duplicates are very young and presumably extremely similar in sequence, it is likely that selection simply favored more copies with identical functions, consistent with the dosage model. In a study of immunity genes among 12 *Drosophila* genomes, Sackton et al. (2007) found that duplicated antimicrobial genes—the effector proteins that have microbial killing abilities—do not evolve rapidly at the nucleotide level but do evolve rapidly in copy number. This pattern is in contrast to the recognition genes of the immune system, which appear locked in a coevolutionary arms race with infective microbes. The authors propose that the large numbers of conserved effector proteins are needed because there is selection for high rates of translation on infection. Such a conclusion would be consistent with the gene conservation outcome and the dosage model.

The most difficult issue in demonstrating that gene conservation is the outcome maintaining gene duplicates is the need to demonstrate that the 2 paralogs have exactly the same function. Whether the function is defined as breadth of expression, enzyme efficiency, or some other aspect of protein performance, support for the gene conservation outcome will often require a large set of negative results. As the above examples show, it may be easier to find definitive proof among newly duplicated genes that have not had a chance to diverge, though this means that the long-term prospects for maintenance are unknown. Some authors have predicted that gene conversion among paralogs is expected with gene conservation in order to maintain high sequence similarity (Sugino and Innan 2006), regardless of which of the 2 models is correct. Though gene conversion or high rates of unequal crossing-over are certainly occurring among the rRNA duplicates, as long as there is strong negative selection on the paralogs there does not appear to be any required association between conversion and gene conservation. In addition, the conclusion that the observed correlation between gene expression and rates of gene conversion in yeast is due to selection for the maintenance of ancestral gene function (Sugino and Innan 2006) can more readily be explained by the fact that yeast undergo mRNA-mediated gene conversion (Derr and Strathern 1993; Pyne et al. 2005; Storici et al. 2007). More highly expressed genes are therefore mechanistically more likely to undergo gene conversion (Pyne et al. 2005).

One pattern expected under the dosage model is that the duplicative mutation is itself fixed by adaptive natural selection because individuals carrying the extra copy have higher fitness. The newly fixed duplicated locus will therefore show a pattern of nucleotide variation consistent with a selective sweep: reduced variation and an excess of low-frequency nucleotide polymorphisms. This signature of selection is also expected under alternative models (see below), but only the dosage model predicts that the 2 resulting loci will be almost identical in sequence after the new duplicate has fixed.

Subfunctionalization ("The Differential Regulation of Former Alleles and Their Transformation to Isozyme Genes" [Ohno 1970])

The second major outcome in the evolution of gene duplicates addressed by Ohno is now known as subfunctionalization (Force et al. 1999). Subfunctionalization can most broadly be defined as the division of ancestral functions among duplicated loci. These functions may comprise expression domains—for instance, expression in multiple tissues—protein operations—for instance, functions carried out by different active sites of the same peptide—or any other genetic function (Lynch 2007a). Because both Ohno (1970) and Force et al. (1999) focused on the division of expression domains among paralogs as the main type of subfunctionalization, gene expression is sometimes the only function associated with this term (Figure 1a). However, I use it to mean any division of labor among resulting duplicates, including protein domains (Force et al. 1999; Stoltzfus 1999). There are multiple models of subfunctionalization, some involving adaptive natural selection and some purposefully devoid of any; I outline these models and their predictions below.

Segregation Avoidance

Squeezed between the chapters on gene conservation and subfunctionalization, Ohno outlined his most detailed population genetic model for the maintenance of gene duplicates. The significance of its placement in the book is revealing as it is still not clear which outcome this model should be associated with. I will refer to this model as "segregation avoidance" (Figure 2); a very similar model was proposed by Spofford (1969) and has recently been examined by other researchers (Lynch et al. 2001; Otto and Yong 2002; Proulx and Phillips 2006). The premise of the segregation avoidance model is very simple: If balancing selection is occurring at a single-copy locus via heterozygote advantage, then homozygotes will be produced every generation regardless of the strength of selection. This is because it is impossible to maintain a population of pure heterozygotes, and the population will experience what is called segregation load. But if one of the balanced alleles at the original locus is duplicated to a new location, then individuals can attain permanent heterozygosity, avoiding segregation load. As I have outlined it, it is not clear whether this model describes gene conservation—after all, no change to the ancestral sequence is required—or subfunctionalization—because the multiple functions of the single-copy locus are now carried out by 2 loci. This distinction is largely semantic, however, and I have therefore represented it as a model of both outcomes in Figure 2.

One of the best (and only) examples of segregation avoidance occurs in the acetylcholinesterase (AChE1) locus of the common mosquito, *Culex pipiens*. An allele that confers resistance to organophosphate pesticides when in heterozygote form has been found as a separate duplicated locus in multiple populations (Bourguet et al. 1996; Lenormand et al. 1998; Labbe et al. 2007). The

2 alleles/paralogs differ at only one amino acid, though this change appears to be enough to provide reduced susceptibility to this class of insecticides. Another set of duplicates possibly fixed because of segregation avoidance are immunity genes. Demuth et al. (2006) found high rates of gene gain and loss in multiple families of immunity genes in mammals and proposed that the heterozygote advantage found at many major histocompatibility complex (MHC) loci (e.g., Hughes and Nei 1988) could be permanently fixed by duplicating alternative alleles. Just as balanced alleles at a single MHC locus are lost due to coevolution of the invading microbe (because they no longer confer an advantage) so might duplicated loci be lost. A similar situation may also hold for R genes in plants (Michelmore and Meyers 1998).

Like the dosage model of gene conservation, the segregation avoidance model predicts that the duplicative mutation itself will be fixed by positive selection because of the advantage it confers. In contrast to the dosage model, however, segregation avoidance also predicts that the new locus will be distinct in sequence from at least one of the haplotypes segregating at the ancestral locus and that the alternative allele at the ancestral locus will subsequently become fixed. So the signature of this type of fixation will be that of a selective sweep of a sequence that differs from the original gene. One complication with interpreting these data is that most researchers assume that any differences observed between paralogs arose after the duplication event. However, in cases like the MHC genes, variation between alternative functional alleles is extremely high and there may even be more nonsynonymous than synonymous polymorphisms per site (i.e., $\pi_A/\pi_S > 1$). These data would then be incorrectly interpreted as adaptive evolution that occurred after the duplication event (i.e., $K_A/K_S > 1$) and would therefore likely be associated with a different model of gene duplication (open circle in Figure 2).

Duplication–Degeneration–Complementation

A subfunctionalization model for the maintenance of duplicates that does not require adaptive mutations was proposed independently by Force et al. (1999) and Stoltzfus (1999). This is often considered the only model of subfunctionalization (but see Conant and Wolfe 2008). In the duplication–degeneration–complementation (DDC) model, 1 of the 2 loci resulting from a duplication event suffers a degenerating mutation that results in the loss of a function. The model is agnostic to whether the mutation was already present in the duplicated allele (from a low-frequency variant at the original locus), appeared during fixation of a duplicated locus, or occurred after the duplicate ultimately fixed. Regardless of when the mutation occurs, if the initially unmutated locus then suffers a degenerative mutation that results in the loss a different function, then the 2 genes complement each other and the organism requires both loci for proper functioning. As no advantageous mutations are required during this process, it can proceed solely via the degeneration of ancestral functions at the daughter loci. Force et al. (1999) actually outlined 2 alternative forms of their DDC model: qualitative and quantitative subfunctionalization. Qualitative subfunctionalization is the more widely considered model and is in fact the model that most of the Force et al. paper addresses. But quantitative subfunctionalization is also an interesting idea, so I will consider it in turn.

In the qualitative subfunctionalization model, the ancestral locus has 2 or more distinct subfunctions that are independently mutable. Again, these subfunctions can be either protein or expression based. The 2 daughter copies then have at least 2 complementary functions, though they may still overlap for other functions. Therefore, to identify cases of qualitative subfunctionalization, one must know the range of functions of both the paralogs and the ancestral locus—which is generally represented by a single-copy version of the same gene in a closely related species. If all we know is the range of overlap of the 2 paralogs, we have no way of determining if nonoverlapping functions are complementary or newly evolved. This has been a problem for recent studies of functional divergence among paralogs that do not have an outgroup data set available (e.g., Wapinski et al. 2007), though it has been possible in a few cases (Huminiecki and Wolfe 2004; Tirosh and Barkai 2007; Semon and Wolfe 2008). Of course simply determining that paralogs have been qualitatively subfunctionalized does not imply that the DDC model is responsible for their maintenance; this would further require evidence for the absence of adaptive evolution postulated by other models of subfunctionalization (see the next section).

In the quantitative subfunctionalization model of DDC, there is only one function—such as expression level or enzyme efficiency—but postduplication it can be carried out by multiple genes. For instance, if an organism requires 100 units of some enzyme and this value can be produced by 2 duplicates each producing 50 units rather than one copy, then we would consider these genes to be quantitatively subfunctionalized. A relationship like this one would arise in much the same manner as in the qualitative DDC model, with the duplication of an ancestral locus able to produce the full complement of enzymes followed by the degeneration of each paralog. As soon as both paralogs lose the ability to individually produce the full 100 units, they would both be necessary to the organism. In this way, 2 half dead duplicates might be just as good as one healthy gene. Data in support of this sort of model will be especially hard to come by because of both limits to the accuracy of experimental methods and the many similar predictions made by this model and the dosage model. If, for example, each duplicate is reduced to individually producing 75 units—for a total of 150—we must now be able to distinguish the fitness effects of individuals producing 100 versus 150 units in order to distinguish between the 2 models. Though the dosage model additionally predicts the adaptive fixation of the duplicate, the effects of such events on patterns of nucleotide variation are only evident for a short time after

fixation (Simonsen et al. 1995) and are therefore very hard to detect long after the duplication.

Specialization and Gene Sharing

The specialization and gene sharing models for the maintenance of duplicate genes are extremely similar to one another and differ only in the details. I will consider them to both be models of subfunctionalization, though they are sometimes grouped with models of neofunctionalization (Figure 2). This disagreement largely revolves around the ontological question of what makes up a function. For instance, should we consider paralogous hemoglobin genes to all have the same function or different functions, as the oxygen-binding affinities differ between fetal and adult duplicates? If we consider them all to have the same hemoglobin function, then they are clearly subfunctionalized relative to the single-copy ancestor. But if we think that they have taken on new and different functions, then the duplicates are neofunctionalized. (We can now see how the semantic difficulties involved in simply distinguishing among outcomes make it even more difficult to distinguish among models.) This distinction is especially relevant to the 2 models considered here as both involve adaptive improvement (or change) of the ancestral function.

The difference between the specialization and gene sharing models also revolves around the number of distinct functions carried out by the ancestral protein. As the result of specialization (a term coined by Otto and Yong 2002), "the products of the duplicated genes, although they still act upon the same substrate and use the same coenzyme, acquire kinetic properties which are markedly different from each other" (Ohno 1970, p.67). This model describes cases like the hemoglobin example given above, where a single function is refined among paralogs expressed in several tissues or developmental stages. Gene sharing, on the other hand, requires that the ancestral gene has 2 or more functions that are not independently mutable. Importantly, the multiple functions must already be present in the single-copy gene; Orgel (1977) outlined just such a model "in which the appearance of a new function in a preexisting protein precedes gene duplication." A similar model was also proposed by Jensen (1976), Piatigorsky and Wistow (1991), and Hughes (1994). Piatigorsky (2007) makes it clear that the gene sharing model is distinct from simple models of specialization because it requires proteins with multiple distinct functions (the term "gene sharing" is originally due to Piatigorsky et al. 1988). A recent example (Hittinger and Carroll 2007) shows that this model can even apply to conflicts between the regulatory requirements of genes. But other than relatively minor differences concerning the multiplicity of functions in the ancestral genes, specialization and gene sharing share many of the same predictions.

Both models have an underlying assumption that there is a conflict between the different roles played by the single-copy ancestral gene—it cannot improve one aspect of its performance without negatively affecting other aspects. This "adaptive conflict" (Lynch and Katju 2004) cannot be resolved until gene duplication allows a paralog to escape one of its roles. According to Ohno (1970, p. 67): "Once it is possible for an organism to discriminate between duplicated genes for the same enzyme and use them differentially during ontogenic development, the way is open for an organism to derive ultimate benefit from this type of gene duplication." The 2 models therefore predict that duplication is accompanied by or followed closely by favored mutations in both duplicated genes moving them closer to their new optima. One resulting signature of both specialization and gene sharing will be a pattern of positive selection on the sequences released from conflict. However, the statistical weakness of all common tests for selection still makes it unlikely that such a signature will be found.

Fortunately, the specialization and gene sharing models make some of the most distinct predictions concerning the relative functions of the resulting duplicates. Both models predict that an ancestral single-copy protein will fill multiple roles but will carry out each role worse than the specialized duplicates. Examples of specialization can be found between many pairs of gene duplicates (e.g., Wen et al. 2006), though good functional data are available only for a few examples (e.g., Des Marais and Rausher 2008). The most famous example of gene sharing comes from eye crystallins, which in single-copy form fulfills both the structural role of a lens protein and the enzymatic role of central metabolism (Piatigorsky and Wistow 1991). The duplicated copies of this ancestral gene that appear in multiple lineages have subsequently been able to become highly specialized on the 2 different functions. Another example involves the yeast galactose pathway, where an ancestral single-copy gene performing both transcriptional induction and enzymatic roles is represented by duplicated genes that individually perform each role in *S. cerevisiae* (Meyer et al. 1991). Platt et al. (2000) converted the duplicate responsible for induction into an enzymatically active protein by the substitution of 2 amino acids, demonstrating the close relationship between the 2 paralogs. Hittinger and Carroll (2007) showed that the regulatory sequences of the single-copy gene also appear to have been constrained by adaptive conflict between the inducer and enzymatic roles.

Neofunctionalization ("The Creation of a New Gene from a Redundant Duplicate of an Old Gene" [Ohno 1970])

As mentioned earlier, neofunctionalization is often characterized as Ohno's only proposed mechanism for the maintenance of gene duplicates. We have seen that in actual fact Ohno considered a wide range of mechanisms that may have contributed to increased numbers of genes. Yet even he recognized that there is a qualitative difference between neofunctionalization and other outcomes (Ohno 1970, p. 72): "[neofunctionalization] is different because it contributed to the creation of new gene loci which acquired previously nonexistent functions." Other than the de novo evolution of new genes (e.g., Long et al. 2003; Levine et al. 2006) or the evolution of new functions in existing single-copy genes—which requires either the loss of ancestral functions or the ability for gene sharing—there is no other way to evolve new

functions. Despite the obvious importance of neofunctionalization for both molecular and organismal "evolutions," the role of neofunctionalization over the past decade has been minimized. This must be due partly to a perceived lack of examples of neofunctionalization and partly to a fascination with other outcomes maintaining duplicates. Below I describe 2 highly similar models for neofunctionalization and the predictions of each; I also review the evidence for each of them.

Dyhkhuizen–Hartl and Adaptation

The 2 most common models for neofunctionalization differ only in the role played by adaptive natural selection. There has been some disagreement over the role played by adaptation in neofunctionalization, as it is variously portrayed either as completely absent (Hughes 1994) or as the distinguishing feature of this mechanism (Clement et al. 2006). One problem is that Ohno did not specify what role, if any, he thought adaptive evolution would play over the long term. He makes it clear that a duplication event allows 1 of the 2 loci to be free from the constraints of natural selection to maintain the ancestral function, but he appears to be agnostic about the role natural selection would play in fixing the mutations that accumulate in the redundant copy (whichever that happens to be): "Natural selection would ignore the redundant locus, and thus, it is free to accumulate a series of *forbidden* mutations. . .As a result, the polypeptide chain specified by it might finally acquire a function which is quite different from that assigned to the original gene [p. 72]." The question is whether that first forbidden mutation or any of the subsequent mutations provide a fitness advantage to the organism and are consequently fixed by positive selection.

The Dykhuizen–Hartl model—named by Kimura (1983) after 2 papers that had nothing to do with gene duplication (Dykhuizen and Hartl 1980; Hartl and Dykhuizen 1981)—proposes that none of the mutations at the redundant locus are fixed by selection. Instead, mutations accumulate due to drift and at some later point in time there is a change in environment such that the new version of the duplicated gene is advantageous to the organism. Kimura (1983) thought that such neutral mutations have "latent potential for selection which can be realized under the appropriate conditions." The important feature of this model is that none of the newly arising mutations at the redundant locus ever have a fitness advantage over another segregating allele before they are fixed and the environment, or some aspect of genetic background, changes.

The adaptation model proposes that neofunctionalization occurs by the adaptive fixation of mutations at one of the duplicated loci. What should be clear from the above discussion is that the model does not specify whether the first "forbidden mutation" is fixed by selection or whether only subsequent mutations are. It may be the case that several neutral mutations are required to move a protein to a part of sequence space in which it can access new functions, after which one or dozens more mutations are all adaptively fixed because they refine the new function. Or it may be that new functions are readily accessible through only a single mutation, in which case even the first such change may be fixed by selection. Depending on how many mutations are required and what fraction of these are advantageous, the adaptation model may be characterized by a large number of nonsynonymous mutations (if the basis of the new function is in the protein) or a large number of regulatory mutations (if the new functional is tied to gene expression).

Whether high rates of evolution following a duplication event are due to positive selection or a relaxation of negative selection has been an ongoing question for many years (Goodman et al. 1975; Kimura 1981; Li and Gojobori 1983). The most unambiguous evidence for adaptive evolution is an excess of nonsynonymous mutations per nonsynonymous site to synonymous mutations per synonymous site, $K_A/K_S > 1$ (also known as K_N/K_S, d_N/d_S, d_R/d_S, and ω). This test is also one of the most stringent tests of selection and requires a large number of nonsynonymous substitutions to be significant. Nonetheless, studies of the early evolution of gene duplicates have found a significant proportion of young paralogs with $K_A/K_S > 1$, even if these cases are not always explicitly acknowledged (see, e.g., Figure 1 in Lynch and Conery 2000; Figure 3 in Zhang et al. 2003; Figure 1 in Kondrashov et al. 2002). As an alternative to K_A/K_S, one may use a test of selection that takes advantage of the power gained by using polymorphism data. These tests have been used to look at levels of diversity in new gene duplicates (Yi and Charlesworth 2000; Moore and Purugganan 2003) or to compare nonsynonymous to synonymous ratios between polymorphisms and fixed differences (Long and Langley 1993; King 1998; Cirera and Aguade 1998; Betran and Long 2003; Holloway and Begun 2004; Thornton and Long 2005; Arguello et al. 2006). This latter test was suggested by McDonald and Kreitman (1991) and has been applied to many duplicated gene pairs, either by comparing polymorphism at a recent gene duplicate to divergence from its paralog (King 1998; Jones et al. 2005; Thornton and Long 2005; Arguello et al. 2006) or by comparing polymorphism at an older gene duplicate to divergence from its ortholog in a closely related species (Betran and Long 2003; Holloway and Begun 2004; Matzkin 2004). However, applying the McDonald–Kreitman test to duplicated genes violates one of the major assumptions of this test: namely, that the neutral substitution rate is constant (Jones et al. 2005; Thornton and Long 2005; Arguello et al. 2006). Because of this violation, application of the McDonald–Kreitman test to duplicated genes can be positively misleading. For instance, if there is little constraint early in the history of a duplicate (say $K_A/K_S = 1$), then many nonsynonymous fixed differences will accumulate because they are neutral. If selection then gets stronger in the recent past—as after duplicates find new functions—then the ratio of nonsynonymous to synonymous polymorphism will provide a much different picture than that of divergence (see Figure 3). This can lead to a rejection of the neutral hypothesis and an interpretation of positive selection in the history of the gene duplicate, simply because there is an excess of fixed nonsynonymous differences relative to

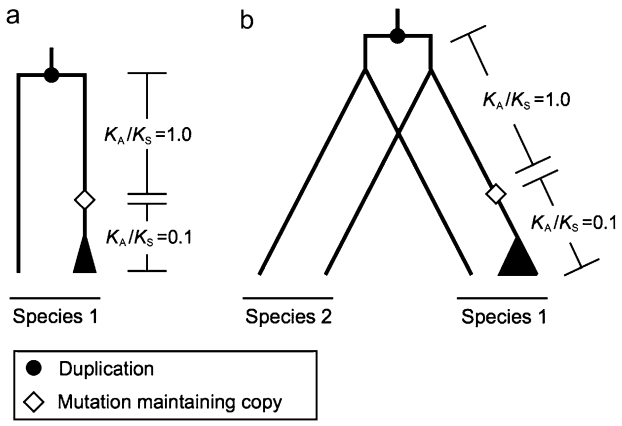

Figure 3. The McDonald–Kreitman test can be positively misleading when applied to duplicated genes. Two scenarios for carrying out the McDonald–Kreitman test are shown. In (**a**) 2 paralogs from the same genome are compared, with polymorphism collected from only one. In (**b**) 2 orthologs are compared, where the duplication event closely preceded speciation between the 2 species; polymorphism is again only collected from one. In both scenarios, a mutation occurs that maintains the duplicated gene—this mutation can lead to neofunctionalization or subfunctionalization. After the duplicate is maintained, selective constraint on the gene increases so that the ratio of nonsynonymous/synonymous mutations (K_A/K_S) is lower in the present.

nonsynonymous polymorphisms. But positive selection on the coding sequence has not necessarily occurred as environmental change (e.g., the Dykhuizen–Hartl model) or even a single advantageous regulatory change may have been responsible for the new function and the consequent change in selective pressure.

The weak power of tests for selection means that it will be hard to distinguish between the Dykhuizen–Hartl and adaptation models for neofunctionalization. However, even if these tests are significant, the adaptation model initially appears to be indistinguishable from subfunctionalization models that predict patterns of positive selection, such as specialization or gene sharing (Figure 2). Additionally, as pointed out by Force et al. (1999), neofunctionalization must be accompanied by the loss of an ancestral function—otherwise there is no pressure to maintain the unchanged paralog. In these cases, functional experiments may have to be combined with sequence analyses to first distinguish among outcomes before distinguishing among models. One pattern of sequence evolution that has been put forward to distinguish among outcomes without the need for functional tests is asymmetrical rates of evolution among paralogs. Previous studies of whole genomes have found that between 5% and 30% of all paralogous genes evolve asymmetrically, with one copy evolving faster than the other (Kondrashov et al. 2002; Zhang et al. 2002, 2003; Conant and Wagner 2003; Kellis et al. 2004; Chain and Evans 2006), though statistical power to detect asymmetry appears to be extremely low (Lynch and Katju 2004). Asymmetry in evolutionary rates is expected under both models of neofunctionalization, with 1 of the 2 duplicates evolving a new function, whereas the other is constrained to carry out the ancestral function. But the interpretation of this pattern as evidence for neofunctionalization has been challenged by He and Zhang (2005), who pointed out that the pattern would also be expected under subfunctionalization models if there is asymmetry in the extent of protein sequence devoted to each of the subfunctions.

A number of examples of neofunctionalization were pointed out by Ohno (e.g., chymotrypsin and muscle actin), and many more examples have accumulated since the publication of his book. Some recent examples include pancreatic ribonucleases in leaf-eating monkeys (Zhang et al. 1998), primate opsins (Yokoyama and Yokoyama 1996), plant lectins (Van Damme et al. 2007), snake phospholipases (Lynch 2007b), vertebrate retinoic acid receptors (Escriva et al. 2006), plant methylthioalkylmalate synthases (Benderoth et al. 2006), vertebrate *Myb* genes (Davidson et al. 2005), primate chorionic gonadotropin (Maston and Ruvolo 2002), primate glutamate dehydrogenase (Burki and Kaessmann 2004), and chordate aldehyde oxidases (Rodriguez-Trelles et al. 2003). All these examples require researchers to study both the functions of the duplicated genes and the function of a single-copy gene in an outgroup to determine that there has been neofunctionalization, usually with the unstated assumption that the single-copy gene has not lost any functions since the most recent common ancestor. However, one class of genes makes it easy to identify neofunctionalization because there is usually no need to test the function of single-copy orthologs—"chimeric" or fusion genes (Long 2000). Chimeric genes are generally formed by the duplication of a single gene, which then co-opts either exons from an unrelated neighboring gene or flanking noncoding DNA that can be used as new coding sequence; often the duplicated gene is a retrotransposed duplicate. Because of the vast structural differences between chimeric genes and the parental genes contributing to their origin, it is usually assumed that they have evolved new functions. As unlikely as this scenario seems, a large number of very interesting examples of chimeric genes have been found with novel functions (Long and Langley 1993; He et al. 1996; Begun 1997; Nurminsky et al. 1998; Finta and Zaphiropoulos 2000; Rogalla et al. 2000; Thomson et al. 2000; Courseaux and Nahon 2001; Elrouby and Bureau 2001; Wang et al. 2002; Paulding et al. 2003; Ciccarelli et al. 2005; Jones and Begun 2005; Jones et al. 2005; Arguello et al. 2006; Cordaux et al. 2006). The fact that many of these new chimeric genes have been found to evolve under positive selection strongly implies that the adaptation model of neofunctionalization is responsible for their maintenance.

Discussion

Despite an amazing amount of new data on the ubiquity of gene duplication among all organisms, it should be clear

from the above discussion that the field is not close to having answers to its major questions. Both conceptual and semantic disagreements abound, leading to wildly varying interpretations of similar data as well as to overgeneralized conclusions from a small set of clear examples. Whereas I have attempted to introduce a clearer framework for interpreting data from studies of gene duplication, no doubt many researchers will disagree with both the definitions and distinctions presented here. However, without at least a common set of terms for the models and outcomes proposed for the maintenance of duplicated genes, we cannot begin to ask which are more prevalent in nature.

This review has stressed the overlapping predictions offered by the different models for the maintenance of duplicated genes and therefore the difficulty in distinguishing among them. Part of the problem is due to the extensive functional data one must collect to distinguish among outcomes, and part is due to the general weakness of all tests for selection in detecting adaptive evolution. Even with information about the extent of gene expression or the evolutionary forces acting on gene duplicates, however, these data can be orthogonal to questions about the relevant models. A pattern of positive selection does not necessarily imply the adaptation model of neofunctionalization, and complementary patterns of gene expression between paralogs do not necessarily imply the DDC model of subfunctionalization. With the exception of unique cases such as fusion genes (neofunctionalization) or fission genes (subfunctionalization)—or unique patterns of sequence evolution (Dermitzakis and Clark 2001)—both functional and evolutionary data must be obtained to distinguish among outcomes and models.

In addition to the specific predictions each model makes about the evolution of individual pairs of duplicated genes, several studies have attempted to take a more global view to distinguish among models. For instance, Lynch et al. (2001) and Walsh (2003) presented results from theoretical models that examined the role of population size in the maintenance of gene duplicates. Both studies found that maintenance due to the DDC model of subfunctionalization increases in probability with decreasing population size and that models incorporating advantageous mutations are more likely to maintain duplicates in large populations. In 2 studies that used the predictions of these theoretical results, Shiu et al. (2006) and Lynch and Conery (2003) compared the retention of gene duplicates in species with larger population sizes (mice and prokaryotes, respectively) against species with smaller population sizes (humans and eukaryotes, respectively). Consistent with a model of adaptive evolution, Shiu et al. (2006) found a higher rate of retention in mice. Consistent with the DDC model, Lynch and Conery (2003) found that eukaryotes maintain larger numbers of duplicates than do prokaryotes and thus that subfunctionalization is more important. These results may not only be due to the DDC model as the differences between prokaryotes and eukaryotes are not limited to population size: "the evolution of multicellularity undoubtedly posed some new selective challenges that were met through neofunctionalization" (Lynch and Conery 2003). Because of the very different methodologies used by these 2 studies, it is difficult to know all the reasons for the differences in results. However, Shiu and colleagues also point out that the original theory is actually not a contrast between subfunctionalization and neofunctionalization per se but rather a comparison of a model with no natural selection (DDC) against models with adaptive natural selection due to new mutations (i.e., dosage, specialization, gene sharing, or adaptation). This makes the results from such studies even harder to fit into the framework of the sub- versus neofunctionalization debate.

It must also be remembered that the main questions of this field revolve around how duplicate genes can be maintained permanently. Data collected scores of My after a duplication event are interesting but not of direct relevance to the question at hand. A number of recent papers have proposed new terminology for the series of events that occur in the lifetime of a gene duplicate (e.g., "subneofunctionalization"; He and Zhang 2005; Rastogi and Liberles 2005). But once a duplicate is maintained in the genome—that is, when it is necessary for organismal function—it will evolve exactly like a single-copy gene. As single-copy genes also undergo adaptive evolution, we might refer to this process as neofunctionalization as well, though it would seem to unduly burden us with meaningless terms. It may simply be more informative to focus our studies on duplicated genes in the first few million years of their existence.

For the present, it is still unclear as to whether neofunctionalization, subfunctionalization, or gene conservation is the most common outcome, and the importance of individual models for each outcome is far from being known. It may also be the case that the most prevalent models differ among duplicates generated via different molecular mechanisms. Ohno clearly thought that polyploidy would have effects on evolution distinct from single-gene mechanisms of duplication: "It would not be surprising if during the course of vertebrate evolution these 2 means were used alternatively [p. 98]." There may be further distinctions still—for instance, retroposition and duplicative transposition are more likely to create duplicates in novel chromosomal environments, which may facilitate neofunctionalization over other outcomes. Additional functional and evolutionary data, as well as novel data on polymorphic gene duplicates (e.g., Sebat et al. 2004; Redon et al. 2006; Kidd et al. 2008), should allow us to fill in many of the gaps in our knowledge of duplicated genes. Whatever the outcome of these future studies, it is clear that Susumu Ohno anticipated them long before the rest of us.

Funding

National Science Foundation (DBI-0543586).

Acknowledgments

Special thanks to M. Lynch, C. McGrath, and J. Demuth for many discussions of the ideas presented here and to C. McGrath, A. Holloway, M. Han, C. Casola, R. Meisel, and P. Nista for comments on the manuscript.

References

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 408:796–815.

Arguello JR, Chen Y, Yang S, Wang W, Long M. 2006. Origination of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila*. PLoS Genet. 2:745–754.

Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science 297:1003–1007.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73:823–834.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current Human Genome Project assembly. Genome Res. 11:1005–1017.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol. 25:2445–2455.

Begun DJ. 1997. Origin and evolution of a new gene descended from *alcohol dehydrogenase* in Drosophila. Genetics 145:375–382.

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J. 2006. Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci USA. 103:9118–9123.

Betran E, Long M. 2003. *Dntf-2r*, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics 164:977–988.

Birchler JA, Levin DM. 1991. Directed synthesis of a segmental chromosomal transposition: an approach to the study of chromosomes lethal to the gametophyte generation of maize. Genetics 127:609–618.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell. 19:395–402.

Bourguet D, Raymond M, Bisset J, Pasteur N, Arpagaus M. 1996. Duplication of the Ace.1 locus in *Culex pipiens* mosquitoes from the Caribbean. Biochem Genet. 34:351–362.

Brosius J. 1991. Retroposons—seeds of evolution. Science 251:753.

Burki F, Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet. 36:1061–1063.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.

Chain FJJ, Evans BJ. 2006. Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. PLoS Genet. 2:478–490.

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res. 15:343–351.

Cirera S, Aguade M. 1998. Molecular evolution of a duplication: the sex-peptide (*Acp70A*) gene region of *Drosophila subobscura* and *Drosophila madeirensis*. Mol Biol Evol. 15:988–996.

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

Clement Y, Tavares R, Marais GAB. 2006. Does lack of recombination enhance asymmetric evolution among duplicates genes? Insights from the *Drosophila melanogaster* genome. Gene. 385:89–95.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. 13:2052–2058.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 9:938–950.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA. 103:8101–8106.

Courseaux A, Nahon JL. 2001. Birth of two chimeric genes in the *Hominidae* lineage. Science 291:1293–1297.

Davidson CJ, Tirouvanziam R, Herzenberg LA, Lipsick JS. 2005. Functional evolution of the vertebrate *Myb* gene family: b-*Myb*, but neither A-*Myb* nor C-*Myb*, complements Drosophila *Myb* in hemocytes. Genetics 169:215–229.

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW. 2006. The evolution of mammalian gene families. PLoS ONE 1:e85.

Dermitzakis ET, Clark AG. 2001. Differential selection after duplication in mammalian developmental genes. Mol Biol Evol. 18:557–562.

Derr LK, Strathern JN. 1993. A role for reverse transcripts in gene conversion. Nature 361:170–173.

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454:762–765.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J Mol Evol. 55: 14–23.

Dykhuizen D, Hartl DL. 1980. Selective neutrality of 6pgd allozymes in *Escherichia coli* and the effects of genetic background. Genetics 96:801–817.

Elrouby N, Bureau TE. 2001. A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. J Biol Chem. 276:41963–41968.

Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, Laudet V. 2006. Neofunctionalization in vertebrates: the example of retinoic acid receptors. PLoS Genet. 2:955–965.

Finta C, Zaphiropoulos PG. 2000. The human cytochrome P450 3A locus. Gene evolution by capture of downstream exons. Gene. 260:13–23.

Fiston-Lavier AS, Anxolabehere D, Quesneville H. 2007. A model of segmental duplication formation in *Drosophila melanogaster*. Genome Res. 17:1458–1470.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Friedman R, Hughes AL. 2004. Two patterns of genome organization in mammals: the chromosomal distribution of duplicate genes in human and mouse. Mol Biol Evol. 21:1008–1013.

Goodman M, Moore GW, Matsuda G. 1975. Darwinian evolution in genealogy of hemoglobin Nature 253:603–608.

Hahn MW, Demuth JP, Han S-G. 2007a. Accelerated rate of gene gain and loss in primates. Genetics 177:1941–1949.

Hahn MW, Han MV, Han S-G. 2007b. Gene family evolution across 12 *Drosophila*genomes. PLoS Genet. 3:e197.

Hartl DL, Dykhuizen DE. 1981. Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*. Proc Natl Acad Sci USA. 78:6344–6348.

He SC, Abad AR, Gelvin SB, Mackenzie SA. 1996. A cytoplasmic male sterility-associated mitochondrial protein causes pollen disruption in transgenic tobacco. Proc Natl Acad Sci USA. 93:11763–11768.

He XL, Zhang JZ. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157–1164.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17:1837–1849.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681.

Holloway AK, Begun DJ. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in *Drosophila*. Mol Biol Evol. 21:1625–1628.

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

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. 14:1870–1879.

Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol. 30:409–425.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet. 39:1361–1368.

Jones CD, Begun DJ. 2005. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci USA. 102:11373–11378.

Jones CD, Custer AW, Begun DJ. 2005. Origin and evolution of a chimeric fusion gene in *Drosophila subobscura, D. madeirensis* and *D. guanche*. Genetics 170:207–219.

Jun J, Ryvkin P, Hemphill E, Mandoiu I, Nelson C. 2008. Estimating the relative contributions of new genes from retrotransposition and segmental duplication events during mammalian evolution. RECOMB-CG 2008. LNBI. 5267:40–54.

Katju V, Lynch M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165:1793–1803.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624.

Kidd J, Cooper G, Donahue W, Hayden H, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64.

Kimura M. 1981. Was globin evolution very rapid in its early stages: a dubious case against the rate-constancy hypothesis. J Mol Evol. 17:110–113.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge (UK): Cambridge University Press.

King LM. 1998. The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. Genetics 148:305–315.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. J Theor Biol. 239:141–151.

Kondrashov FA, Rogozin I, Wolf Y, Koonin E. 2002. Selection in the evolution of gene duplications. Genome Biol. 3:research0008.1–research0008.9.

Labbe P, Berthomieu A, Berticat C, Alout H, Raymond M, Lenormand T, Weill M. 2007. Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito *Culex pipiens*. Mol Biol Evol. 24:1056–1067.

Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell. 131:1235–1247.

Lenormand T, Guillemaud T, Bourguet D, Raymond M. 1998. Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. Evolution 52:1705–1712.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci USA. 103:9935–9939.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.

Li WH, Gojobori T. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol. 1:94–108.

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 437:94–100.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.

Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science. 260:91–95.

Long MY. 2000. A new function evolved from gene fusion. Genome Res. 10:1655–1657.

Lynch M. 2007a. The origins of genome architecture. Sunderland (MA): Sinauer Associates, Inc.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet. 20:544–549.

Lynch M, O'Hely M, Walsh JB, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804.

Lynch VJ. 2007b. Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. BMC Evol Biol. 7:2.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA. 102:5454–5459.

Maston GA, Ruvolo M. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol. 19:320–335.

Matzkin LM. 2004. Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. Mol Biol Evol. 21:276–285.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 652–654.

Meyer J, Walkerjonah A, Hollenberg CP. 1991. Galactokinase encoded by *GAL1* is a bifunctional protein required for induction of the *GAL* genes in *Kluyveromyces lactis* and is able to suppress the *gal3* phenotype in *Saccharomyces cerevisiae*. Mol Cell Biol. 11:5454–5461.

Michelmore RW, Meyers BC. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 8:1113–1130.

Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. Proc Natl Acad Sci USA. 100:15682–15687.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature 396:572–575.

O'Hely M. 2006. A diffusion approach to approximating preservation probabilities for gene duplicates. J Math Biol. 53:215–230.

Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.

Orgel LE. 1977. Gene-duplication and origin of proteins with novel functions. J Theor Biol. 67:773.

Otto SP, Yong P. 2002. The evolution of gene duplicates. In: Dunlap J, Wu C-T, editors. Homology effects. San Diego (CA): Academic Press. p. 451–483.

Pan D, Zhang L. 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biol. 8:R158.

Paques F, Haber JE. 1999. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev. 63:349–404.

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis, Oryza, Saccharomyces* and *Tetraodon*. Trends Genet. 22:597–602.

Paulding CA, Ruvolo M, Haber DA. 2003. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. Proc Natl Acad Sci USA. 100:2507–2511.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 39:1256–1260.

Piatigorsky J. 2007. Gene sharing and protein evolution. Cambridge (MA): Harvard University Press.

Piatigorsky J, O'Brien WE, Norman BL, Kalumuck K, Wistow GJ, Borras T, Nickerson JM, Wawrousek EF. 1988. Gene sharing by δ-crystallin and argininosuccinate lyase. Proc Natl Acad Sci USA. 85:3479–3483.

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

Platt A, Ross HC, Hankin S, Reece RJ. 2000. The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase. Proc Natl Acad Sci USA. 97:3154–3159.

Proulx SR, Phillips PC. 2006. Allelic divergence precedes and promotes gene duplication. Evolution 60:881–892.

Pyne S, Skiena S, Futcher B. 2005. Copy correction and concerted evolution in the conservation of yeast genes. Genetics 170:1501–1513.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol. 5:28.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen WW, et al. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Rodriguez-Trelles F, Tarrio R, Ayala FJ. 2003. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the *xanthine dehydrogenase* gene. Proc Natl Acad Sci USA. 100:13413–13417.

Rogalla P, Kazmierczak B, Flohr AM, Hauke S, Bullerdiek J. 2000. Back to the roots of a new exon—the molecular archaeology of a *SP100* splice variant. Genomics 63:117–122.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. Nature Genetics 39:1461–1468.

Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. Nat Rev Genet. 3:65–72.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. Science 305:525–528.

Semon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc Natl Acad Sci USA. 105:8333–8338.

She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone MF, Rocchi M, Program NCS, Green ED, et al. 2006. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res. 16:576–583.

Shiu S-H, Byrnes JK, Pan R, Zhang P, Li W-H. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci USA. 103:2232–2236.

Shoja V, Zhang LQ. 2006. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol Biol Evol. 23:2134–2141.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429.

Spofford JB. 1969. Heterosis and evolution of duplications. Am Nat. 103:407–432.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. J Mol Evol. 49:169–181.

Storici F, Bebenek K, Kunkel TA, Gordenin DA, Resnick MA. 2007. RNA-templated DNA repair. Nature 447:338–341.

Sugino RP, Innan H. 2006. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. Trends Genet. 22:642–644.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet. 38:615–643.

Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M, et al. 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene. Genome Res. 10:1743–1756.

Thornton K, Long M. 2005. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol Biol Evol. 22:273–284.

Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol. 8:R50.

Van Damme EJM, Culerrier R, Barre A, Alvarez R, Rouge P, Peumans WJ. 2007. A novel family of lectins evolutionarily related to class V chitinases: an example of neofunctionalization in legumes. Plant Physiol. 144:662–672.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA. 103:3220–3225.

Walsh B. 2003. Population-genetic models of the fates of duplicate genes. Genetica. 118:279–294.

Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc Natl Acad Sci USA. 99:4448–4453.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61.

Wen ZM, Rupasinghe S, Niu GD, Berenbaum MR, Schuler MA. 2006. CYP6B1 and CYP6B3 of the black swallowtail (*Papilio polyxenes*): adaptive evolution through subfunctionalization. Mol Biol Evol. 23:2434–2443.

Yi S, Charlesworth B. 2000. A selective sweep associated with a recent gene transposition in *Drosophila miranda*. Genetics 156:1753–1763.

Yokoyama S, Yokoyama R. 1996. Adaptive evolution of photoreceptors and visual pigments in vertebrates. Annu Rev Ecol Syst. 27:543–567.

Zhang JZ. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.

Zhang JZ, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 95:3708–3713.

Zhang LQ, Vision TJ, Gaut BS. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol Biol Evol. 19:1464–1473.

Zhang P, Gu ZL, Li WH. 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol. 4:R56.

Corresponding Editor: Michael Lynch