1 Understanding Gene Duplication Through Biochemistry and Population Genetics

DAVID A. LIBERLES and GRIGORY KOLESOV
Department of Molecular Biology, University of Wyoming, Laramie, Wyoming
KATHARINA DITTMAR
Department of Biological Sciences, SUNY at Buffalo, Buffalo, New York

1 INTRODUCTION

Gene duplication has emerged as an important process supporting the functional diversification of genes. Since publication of the seminal book *Evolution by Gene Duplication* by Ohno (1970), the hypothesis regarding the importance of gene duplication in the generation of evolutionary novelty has steadily gained support as we have entered the genome-sequencing era. It is through the link to functional biology that an ultimate understanding of the preservation and diversification of duplicate genes will be accomplished.

Genes can diverge in function through accumulation (fixation) of coding sequence changes, which may influence binding interactions and/or catalysis, through the evolution of splice variants, and through spatial, temporal, and concentration-level changes in the expression of the protein product. Governing these processes is an interplay among mutational opportunity, population dynamics, protein biochemistry, and systems and organismal biology. This interplay is described systematically in this chapter.

2 SYSTEMS BIOLOGY AND HIGHER-LEVEL ORGANIZATION

At the level of biological systems, two early but still relevant views suggested a role for gene duplication in constructing pathways. These views are both dependent on a new function emerging in one of the duplicates, but differ in the manner in which it occurs. One view, *patchwork evolution*, involved a conservation of catalytic activity coupled with the evolution of a new substrate after duplication (Jensen, 1976). An alternative view, *retrograde evolution*, suggested that pathways are built up backward, with product becoming substrate based on recognition of the transition state in the

Evolution After Gene Duplication, Edited by Katharina Dittmar and David Liberles

active site, with the evolution of a new catalytic activity to generate the substrate for the downstream reaction after duplication (Horowitz, 1945). In a systematic analysis in *Escherichia coli*, Light and Kraulis found some evidence for the retrograde evolution model, but found the patchwork model to be much more common, possibly because it is easier to gain new binding specificity than to evolve a new catalytic activity (Light and Kraulis, 2004). Relatedly, it has been suggested that (also in bacteria) there are secondary (moonlighting) functions where enzymes with a given catalytic activity carry it out on multiple substrates with different specificities (Copley, 2003). This nature of enzymatic activities might generally lead to quick differential optimization after duplication, especially easily if maintained with different specificities in different alleles by balancing selection before duplication. Further (as discussed in detail below), specificity is chemically and evolutionarily difficult to attain, and nonspecific binding activities may arise easily when there is no selective pressure against them. Whereas selective pressures are ultimately at the systems level, divergence occurs gene by gene and mutation by mutation. This process will be dissected.

3 MUTATIONAL DYNAMICS AND SUBSTITUTIONS

Both intramolecular and intermolecular coevolution of sites affects the probability of fixation of any individual mutation, where genetic background (the sequence at genetically interacting positions) determines the phenotype of any given mutation. The evolutionary accessibility of different mutations from a given genetic background is therefore dictated partly by the mutation rate and the frequency of multiple segregating mutations as well as the population size as a dictator of strength of selection. The same evolutionary properties affect both intramolecular and intermolecular interaction, only with differing degrees of sensitivity to mutation, due to the entropic differences between the two types of interresidue interaction. For these entropic reasons, it is easier to knock out a binding interaction than to knock out proper protein folding (although this happens, too) with a single mutation. This is because although there are a greater number of sites that influence proper folding, covalent attachment means that there will also be a greater local effective concentration of intramolecularly interacting residues requiring a lower affinity interaction to generate the same levels of bound state. If one views two residues as interacting or not interacting, the probability of interaction at any given time is dependent on their affinity for each other and how many opportunities they have to interact (their concentration about each other).

So far, we have focused on the coding properties of a gene. Gene expression is another important process that is subject to phenotypic divergence through mutation. The typical gene has approximately 12 transcription factor binding sites [the distribution of this across genomes is not well characterized, and this number is given with an approximation of six to eight base pairs (Harbison et al., 2004; Hughes and Liberles, 2007)]. The specificity of binding typically enables transcription factors to discriminate among many sites with single-base-pair mutations (Lusk and Eisen, 2008). Because of the small size of transcription factor–binding sites, site loss and de novo site evolution are reasonably common, and this is explored further below. Due to the periodicity of standard B-form DNA of about 10 bp, as well as changes in effective local concentration of transcription factors about each other and about the initiation site, it might be expected that spacing between sites is important in gene regulation, but evidence

generated so far seems to downplay the role of these effects (Shultzaberger et al., 2007), leading to a focus on the evolution of the sites themselves.

Splicing is another mechanism by which genes can diverge through mutation. There are two types of splicing, constitutive and alternative, with alternative splicing simply showing a weaker consensus to splicing regulatory sites (Churbanov et al., 2008). Like transcription factor–binding sites, splicing regulatory sequences are short and potentially subject to turnover. However, because of the lack of redundancy (unlike transcription factor–binding sites), loss in the absence of duplication may frequently be highly deleterious. It has been shown that alternative splicing itself enables a substitution burst mediated by relaxed selection on and around these regulatory sites (Xing and Lee, 2005). That gene duplication can also enable such a burst of substitution under relaxed selection suggests that gene duplication should enable enhanced rates of alternative transcript generation, and this has indeed now been demonstrated (Jin et al., 2008).

Many other molecular mechanisms can contribute to mutation-driven diversification. A far from exhaustive list would include glycosylation sites, protein splicing, and RNA editing—one only needs to think of the effects of duplication and relaxed selection on any processes generating constraint described in a molecular biology textbook.

Starting with a few examples of several of these molecular processes, we will then link mutational opportunity to evolutionary mechanism and process. The following section includes a series of examples of the fates of duplicate genes. These examples are meant to be illustrative, and we will ultimately address how general the various processes that underlie the examples actually are.

4 EVOLUTION OF ENZYME ACTIVE CENTERS AFTER DUPLICATION

Mutations in the active center(s) of an enzyme can lead to a change of its substrate or a change in its kinetics. For example, Vick and Gerlt (2007) demonstrate that a single-base-pair change leading to D-to-G substitution in the active center of the monofunctional L-Ala-D/L-Glu epimerase from *E. coli* introduced the ability to catalyze the *o*-succinylbenzoate synthase reaction while reducing the level of the original reaction (Figure 1). Four additional nucleotide substitutions led to a complete switch of specificity and kinetics to the new reaction. Consistent with the patchwork model discussed earlier, a large number of enzymes in the arginine and lysine synthetic pathways are homologous to each other (Miyazaki et al., 2001).

Mutations in the structure surrounding the active center can lead to fine-tuning the active center to different but fundamentally similar substrates. For example, residues in the active centers of Leu-tRNA synthase and Ile-tRNA synthase are mostly conserved; Leu and Ile are very close chemically. There are a number of variable residues that do not directly contact the substrate residue in the active center but, rather, shape the active center, allowing for recognition of the cognate substrate residue. Both tRNA synthases are highly similar on both the sequence and structural levels. Leu- and Ile-tRNA synthases probably arose via gene duplication (Brown and Doolittle, 1995). This demonstrates a shift in substrate specificity following gene duplication.

4.1 Change of DNA-Binding Specificity

Homeobox genes are homeodomain-containing transcription factors that are known as principal regulators in the formation of the animal body plan during embryo

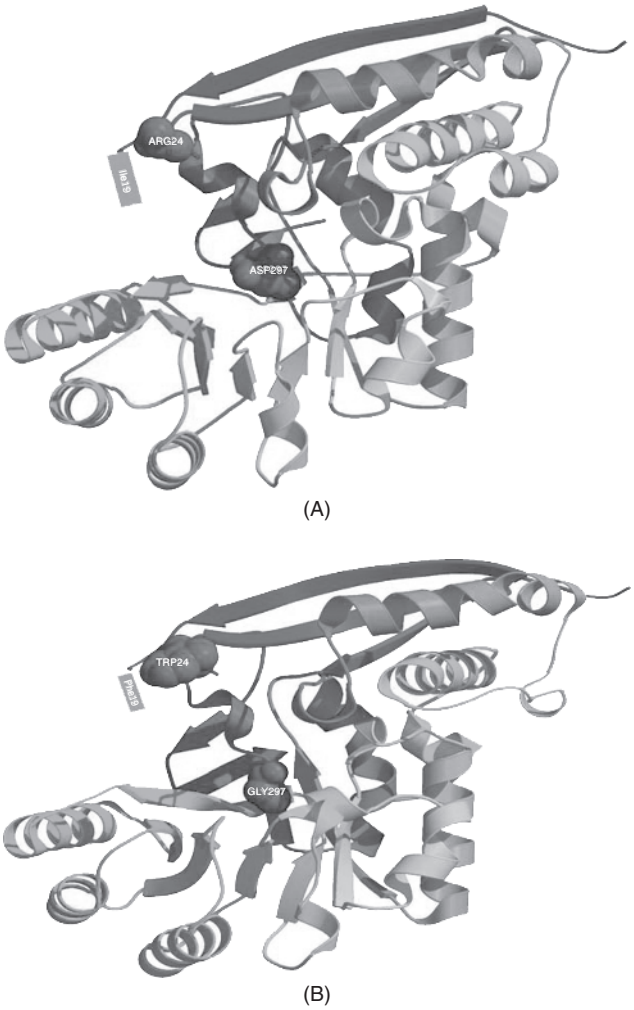

Figure 1 (A) Substitutions in the active center of the L-Ala-D/L-Glu (SP2Q filling residues). (B) Substitution of one nucleotide in codon coding for Asp297 leads to acquiring of *o*-succinylbenzoate activity. Further substitutions of Arg24 and Ile19 (located on an unresolved loop) lead to a complete switch of specificity of the reaction (PDB ID 1JPD).

development. They are often organized in homologous gene clusters such as *Hox*, *ParaHox*, and *NK* (Garcia-Fernández, 2005). Hox clusters can contain different numbers of genes in different species, where new genes in the clusters arise via duplication and loss in the course of evolution.

It has been shown that the DNA-binding specificity of *Hox* genes is controlled by a few key positions in the homeodomain. For example, substitution of Gln to Lys in position 50 of the homeodomain alters recognition from TAATCC (recognized by bicoid class hox proteins) to the TAAT(T/G)(A/G) motif recognized by the Antennapedia and Engrailed classes (Hanes and Brent, 1989; Treisman et al., 1989; Percival-Smith et al., 1990). This is shown in Figure 2. Similarly, substitutions in positions 3, 6, and 7 of the

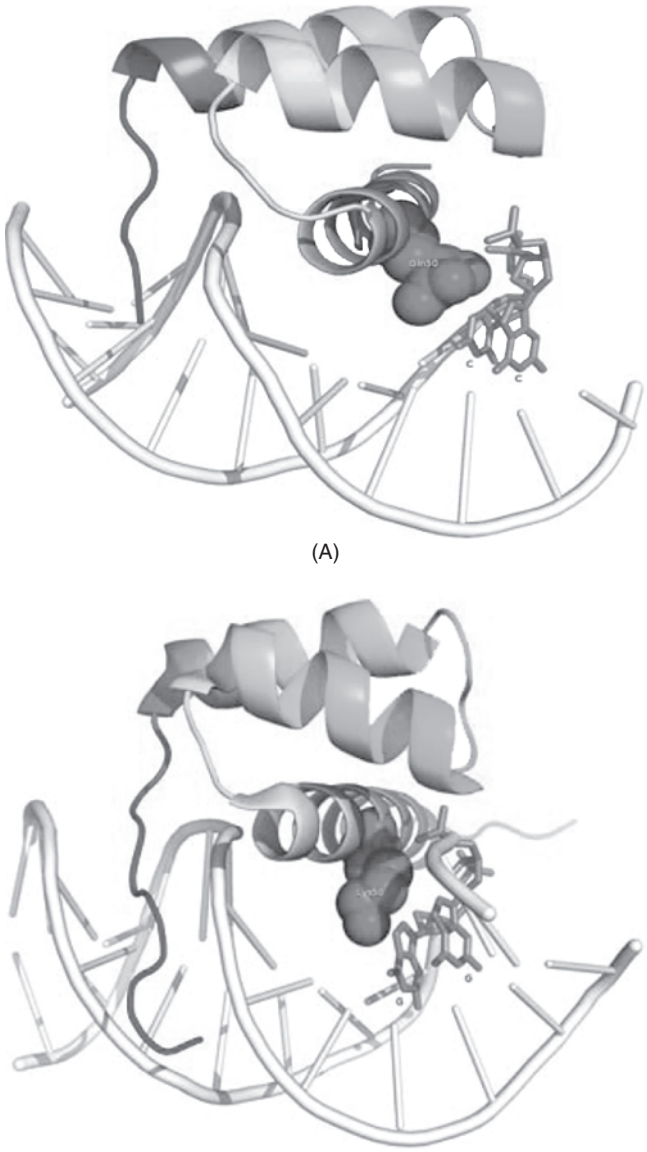

Figure 2 X-ray and NMR structures of three Hox proteins from *Drosophila melanogaster* in complex with its DNA recognition site. Mutation in crucial position 50 of the homeodomain switches the specificity to a different DNA binding site. Gln50 in Antennapedia (A) contacts GG on the antisense strand, leading to recognition of the TAATGG core motif, while Lys50 in the bicoid homolog of Antennapedia (B) recognizes TAATCC. The recognition site can be changed mutagenically (Tucker-Kellogg et al., 1997). (C) Ultrabithorax (a tandem duplicate of Antennapedia) has the same DNA-binding motif, but it developed an interaction with the Ebx homeodomain protein mediated primarily by the terminal YPWM motif of Ubx that binds to the hydrophobic pocket in Ebx (Passner et al., 1999). As a result of the interaction, the complex cooperatively recognizes the TGATTTATGG/ATAAATCA motif. The disordered flexible linker (unresolved on the structure) is shown in an extended conformation. (From Baird-Titus et al., 2006.)

(C)

Figure 2 (*Continued*)

N-terminus of the homeodomain alter the specificity toward the nucleotide in position 2 of the motif TTATGG → TAATGG (Ekker et al., 1994; Noyes et al., 2008).

The evolution of homeotic genes in *Hox*-like clusters demonstrates how gene duplication followed by a single or a few mutations can create new functions that have dramatic effects on the phenotype (in the case of *Hox* genes, the number of body segments, limbs, etc.). An example of the rearrangement of *Hox* genes and their regulatory elements is shown in Figure 3.

4.2 Change of Binding Interface and Interaction Partners

Most proteins do not act alone but, instead, interact with other proteins. This is another mode of potential divergence for duplicated genes. Protein–protein interactions are in most cases highly specific and form complex protein interaction networks that execute metabolic functions and make up regulatory, signal transduction, and intercellular circuits. Mutations in the protein–protein binding interfaces have a significant impact on the function of the protein and the network in which it participates. A bacterial example will be used to illustrate this ubiquitous nature of biological systems.

Two-component systems are the most common signal transduction systems in bacteria and are responsible for the sensing and adaptation of the bacteria to a variety of environments, nutrients, and stresses. The family of two-component systems is comprised of homologous proteins. The high degree of binding specificity in proteins making up a two-component system allows for virtually no detrimental crosstalk in a bacterial cell, which can contain up to 200 different two-component systems. The

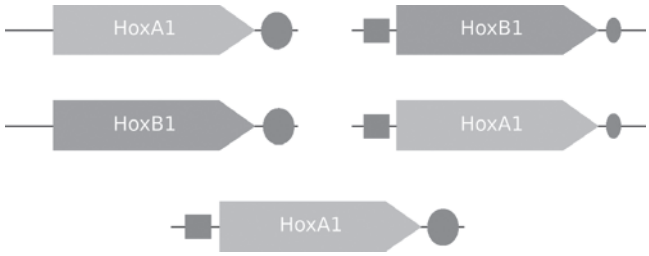

Figure 3 *HoxA1* and *HoxB1* genes in mammals. (a) Wild type. *HoxA1* has a fully functioning 3′retinoic acid response element (RARE, circle). *HoxB1* has both a *Hox1* autoregulatory element (ARE, square) and a RARE, with the functionality of the latter severely reduced. (b) Swapping coding regions of the gene in mice produce a normal phenotype. (c) Combing the fully functional ARE and RARE elements around one of the genes and knocking out another also produces a normal phenotype. This is believed to be an ancestral form that preceded WGD and subsequently subfuctionalized so that *HoxA1* lost ARE and *HoxB1* retained it but deteriorated its RARE. (From Trdvik and Capecchi, 2006.)

two components are typically a membrane-localized sensor with kinase activity and a transcription factor that is phosphorylated. It has been shown that by substituting only three residues in the kinase, the specificity of the kinase can be switched completely to another two-component signal transduction pathway, thus drastically changing the signal transduction logic (Skerker et al., 2008) (Figure 4).

Another interesting feature of the archetypical two-component system is that similar to *Hox* clusters, genes coding for its components are clustered on the bacterial chromosome, comprising an operon. That allows for duplication and subsequent divergence of the entire system cooperatively.

4.3 Change of Regulatory Elements That Control Gene Expression

Changes in the way that genes are regulated affect the timing, level, and tissue specificity of gene expression. It has been shown in the case of the *HoxA1* and *HoxB1* genes that swapping their coding regions has no detrimental effect in mouse development (Tvrdik and Capecchi, 2006) (Figure 3). Moreover, combining elements of the regulatory region of both duplicates into one delivers a fully functional gene that carries out the function of both genes, resulting in normal mice. This work demonstrates an apparent case of historical subfunctionalization of regulatory regions. As discussed later, this appears to be common in the evolutionary trajectories of duplicated genes.

4.4 Instantaneous Change of Regulation of Duplicate Copy

The duplication event itself can radically alter the way in which gene expression is regulated. For example, in considering the fate of X-linked gene *Utp14*, Bradley et al. (2004) found a retrocopy (*Utp14b*) integrated in the intron of autosomal gene *Acsl3* on mouse chromosome 1 (Figure 5). The presence of the retrocopy is essential for proper spermatogenesis in mice. The retrocopy is regulated by the promoter of the host gene, and unlike *Utp14*, is not affected by X inactivation during spermatogenesis. Thus, as a result of the retrotransposition event, the second copy switched its promoter, moved

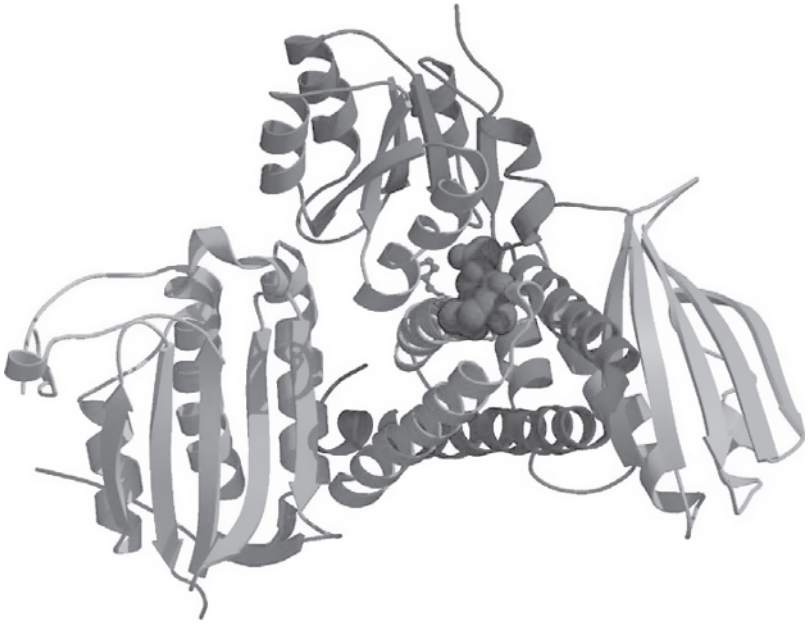

Figure 4 Change of binding interaction in the two-component system. Shown in red blobs are three amino acid substitutions described by Skerker et al., (2008) that completely switch from EnvZ histidine-kinase to OmpR HK signal transduction. The HK homodimer is on the bottom (green and blue domains), and the response regulator domain (brown, top) was computationally docked to the 2C2A HK structure by Marina et al. (2005). The phosphotransfer histidine is shown in magenta. (*See insert for color representation of the figure*.)

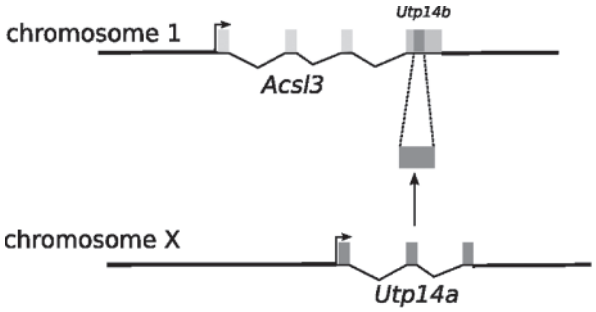

Figure 5 Retrotransposition of *Utp14* from the mouse X-chromosome onto the untranslated exon of gene *Acsl3* located on chromosome 1 is depicted.

to a different chromosome with different regulation of chromatin packing, and lost all introns. This can be considered a case of regulatory neofunctionalization.

4.5 How General Do We Expect the Examples Above To Be?

We have seen a collection of examples where different molecular mechanisms interplay with different evolutionary processes. Starting from the initial duplication event,

through mutational opportunities to affect different molecular processes to different mechanisms on a population and evolutionary scale, we systematically evaluate the potential for duplicate gene retention.

Initial gene duplication events occur in a single individual. The rate of fixation of duplicates within a population depends on the effective population size and the degree of selection. Some treat the initial events as neutral (Force et al., 1999; Lynch and Force, 2000) with some case-specific positive selection (see below; Perry et al., 2007), whereas others view duplication events as deleterious (Wagner, 2005). Given that trisomy in humans is lethal except for chromosome 21 and the sex chromosomes (and these cases are associated with reduced fitness), duplication of a subset of genes is clearly deleterious in humans. Different lineages in the tree of life show different propensities to tolerate gene duplication, and mammals of small effective population size differ from plants of small effective population size in this regard. Even in plants, different genes and gene functions are retained differently after duplication events (Hanada et al., 2008), although this analysis does not yet sort out the role for selection in the initial duplication event, and further work is needed.

All of the processes above are described as single events in a species. In actuality, these events occur in a single individual and are then subjected to population-level processes simultaneous to the process of divergence. Genes that are born identical or that have not diverged in a mechanism that affects fitness will be born in proportion to effective population size and will be fixed in inverse proportion to effective population size. Once a fitness advantage is gained (where the probability of advantageous mutation is proportional to effective population size), the probability of fixation is inversely proportional to less than the effective population size. The degree by which the probability of fixation based on effective population size is modulated is dependent on the strength of selection. It is therefore expected that selective processes are more common in organisms of large effective population size. Selection can also be driven by mutation rate, with higher mutation rates providing a greater sampling of changes to access those of adaptive effect.

Whole-genome duplications (WGDs) have an added complexity in sexual organisms. Perhaps the rarity of whole-genome duplication events is that successful reproduction is dependent on two individuals with whole-genome duplications finding each other and mating, coupled to the interplay of population genetics involving the relative fitness of offspring with a whole-genome duplication compared to individuals without a whole-genome duplication. This scenario is dependent on the cessation of gene flow between the two subpopulations.

Moving on to the initial duplication event at the molecular level, a wide variety of processes can lead to gene duplication. At the grossest level, whole-genome duplication results in duplication of every gene in the genome. Under this process, each gene is identical upon arrival, in terms of both coding sequence and regulation. Further, every interacting partner is duplicated together with the gene itself, resulting in a doubling of the interactome. The next level down involves other large-scale (e.g., whole-chromosome) duplication events, where the gene is identical in coding sequence and regulation but does not necessarily have any or all of its interacting partners duplicated. This distinction is important for some of the underlying mechanisms for duplicate gene retention, as we will see.

Other mechanisms involve duplication of a single gene at a time without interacting partners, but otherwise also involve differences. Tandem duplication is mediated

by recombination, break-and-repair processes, or polymerase error. Tandemly duplicated genes are probably identical in coding sequence and regulatory elements, but have a chance of missing a terminal domain and distal regulatory elements. Genes duplicated by DNA-level transposition will probably be identical in coding sequence, again with the chance of missing a terminal domain, but will probably be born in a new gene expression environment. There is a chance of retention of proximal expression elements. Genes duplicated by retrotransposition will be born identical in coding sequence except for the lack of introns (eliminating the possibility of splicing-level divergence). These genes will be born in a new gene expression environment. If the new environment does not result in expression of the gene, the duplicate that was created is dead upon arrival.

5 MUTATIONAL OPPORTUNITIES AFTER DUPLICATE GENE BIRTH

Whereas the birth process itself may introduce changes to the gene that result in functional modification, subsequent to birth, random changes occur independently in each copy that lead to divergence. The opportunity to effect functional change through either gain or loss of function without creating a nonfunctional gene in either duplicate is expected to be proportional to the number of sites where such changes can possibly happen. The easiest events to envision are loss of a transcription factor binding site and loss of a binding site from the protein. The average gene has 12 transcription factor–binding sites of typical length six to eight base pairs, where one or two mutations in a site will alter or knock out transcription factor binding. The average protein interacts with one to three other proteins under a power-law distribution (Luscombe et al., 2002). The size and nature of an interaction interface ranges from two to five amino acids for modifying enzymes (Puntervoll et al., 2003), with larger sites for transient and obligatory interactions. For transient interactions, the average recognition site is widely variable in size and typically has shown a significant energetic contribution, from 12 to 15 amino acids (Chakrabarti and Janin, 2002), but fewer in other studies (Bogan and Thorn, 1998). Each amino acid site corresponds to potential changes in roughly 2.5 nucleotide positions (from the genetic code). Larger interaction interfaces, including among obligate interactions, tend to be driven by hydrophobic interactions, while smaller interfaces, including transient interaction interfaces, are more driven by electrostatic interactions (Bradford et al., 2006). The role of the remaining residues not contributing to the binding affinity is thought largely to be to exclude solvent (Bogan and Thorn, 1998). These residues are less constrained evolutionarily and do not affect specificity (Caffrey et al., 2004; Guharoy and Chakrabarti, 2005). Even among the binding interface residues that contribute to the binding affinity, the degree of amino acid sensitivity between similar amino acids is unclear. It has been suggested that a small subset of electrostatic residues may drive specificity in a sea of hydrophobic interactions driving affinity (Pechmann et al., 2009). Changes in untranslated regions can also affect mRNA stability, but have not been factored into the view described above. This quick back-of-the-envelope calculation with a few unknowns shows that it should be easier to change a gene expression profile than a binding profile, but not overwhelmingly so (roughly, one- to tenfold more likely). In fact, evidence suggests that subfunctionalization of gene expression is typically the first thing to happen, but followed subsequently by change (potentially neofunctionalization as well as subfunctionalization) in protein

function (He and Zhang, 2005). However, the back-of-the-envelope calculation above, based on the mean, will be sensitive to the underlying distributions, with many fold- and gene-specific effects (see below). Additionally, the affinity of the transcription factor to a regulatory region can be determined by the enrichment of different motives rather than by singular sites (Badis et al., 2009).

While the foregoing estimations deal with loss or modification of existing binding sites, surface regions can evolve new binding interactions that were not present in the ancestor. This has been suggested for leptin in primates in the absence of duplication (Gaucher et al., 2003), but represents a mechanism that should be even more accessible to duplicate genes. Although it is generally thought that binding interactions will evolve more easily than catalytic activities, many binding interfaces include residues that are buried in pockets or exposed only upon conformational shifts in binding.

6 EVOLUTIONARY MECHANISMS

We have mentioned several possible evolutionary mechanisms acting upon available mutations at different levels. Next, we examine these mechanisms systematically; they are summarized in Figure 6.

6.1 Pseudogenization

Pseudogenization, the most common fate for duplicate genes, arises from the random neutral accumulation of mutations, most of which are deleterious. Eventually, the gene no longer functions. For the products of small-scale duplication (SSD) events, a fraction of genes will be born without the expression elements necessary to have a function that confers fitness. The same applies to genes born missing terminal domains.

6.2 Subfunctionalization

Subfunctionalization is a mechanism that involves a combination of neutral mechanisms and negative selection to relax the redundancy of duplicate copies via complementary loss of functional attributes between the duplicates. The functions of a protein, whether expression domains, binding interactions, alternative splice forms, or other features, are viewed modularly, with evolutionary dynamics characterized by mutational opportunities for loss of different modules. Genes that have more regulatory regions, including those that regulate development, will be more prone to subfunctionalization. Because this mechanism does not involve positive selection, it has been viewed as more important in smaller effective population-size lineages. Some products of tandem duplication or DNA-level transposition will be born subfunctionalized.

6.3 Neofunctionalization

Neofunctionalization involves the development of new functions. This can include the development of de novo transcription factor–binding sites, the modification of existing sites to change the specificity, affinity, or kinetics, the modification or gain of binding interactions, the modification or gain of splice regulatory elements, and a number of other events. The frequency of neofunctionalization depends on the frequency of

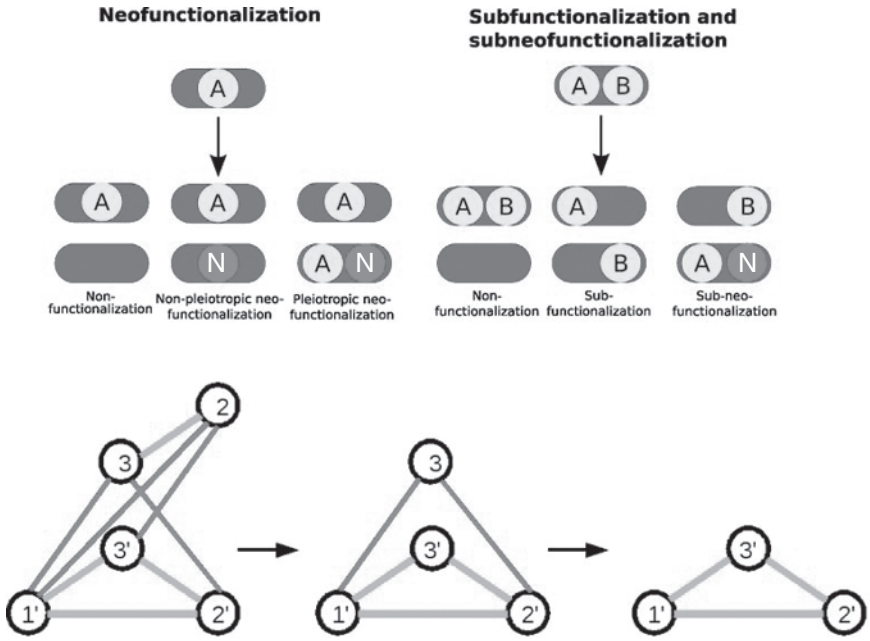

Figure 6 Schematic depicting the processes of neofunctionalization, subfunctionalization, and dosage compensation. Neofunctionalization can occur either pleiotropically or nonpleiotropically, depending on whether the new function occurs in a region that also carries out the original function. Subfunctionalization can occur alone or together with neofunctionalization. In the bottom panel, the decay of binding interactions driven by changes in stoichiometry is shown.

neofunctionalizing events. Because of the complexity of interacting mutations not only within but between genes, neofunctionalization rates may show a time lag and are more complex than the simple rate of instantaneously beneficial mutations within a population. Further, a new function at the molecular level does not necessarily implicate a selectable advantage and positive selection. Some new molecular functions will be evolutionarily neutral.

Timing of Neofunctionalization The classical model (which still may be the most common) suggests that when neofunctionalization occurs, the relaxation of selective constraints associated with gene duplication paves the way. However, there is some evidence for increased substitution preceding duplication events leading to duplicate gene retention (Johnston et al., 2007). There are several mechanisms that have been characterized associated with this. One mechanism is fixation of selectively balanced alleles (Sato et al., 2001), where alleles that benefit the heterozygous individual individually are fixed at different loci. Another mechanism involves enzymes that catalyze side reactions, where duplication allows subfunctionalization of the main reaction and side reaction and optimization of the side reaction without pleiotropic constraint.

Selection for Increased Dosage as a Form of Neofunctionalization In addition to changes to transcriptional (and translational) regulatory regions, gene duplication can be

a mechanism to increase the dosage of a gene, where increased dosage is beneficial. An example of this that has been suggested in the human population is salivary amylase I, which apparently varies in copy number in correlation with starch consumption (Perry et al., 2007).

Dosage Compensation Duplicating a gene that instantaneously leads to a doubling of expression is potentially deleterious for several reasons (Wagner, 2005; Drummond and Wilke, 2008). Beyond any deleterious effects due to the cost of expression or mistranslation (or gain of low-affinity interactions at higher protein concentrations), it is thought that stoichiometric imbalance is deleterious (Aury et al., 2006). Thus, when two or more interacting partners are duplicated, there is expected to be a selective pressure to retain such duplicates together in a genome for long evolutionary periods. Loss of one of the copies or down-regulated expression of one copy is then expected to lead to positive selection for the loss of interacting partner duplicates (or down-regulated expression) [see Hughes et al. 2007 for a discussion]. Subfunctionalizing mutations are expected to be deleterious and also lead to loss of interacting partners to restore stoichiometric balance in interactions. Additionally, subfunctionalizing interactions have the potential to cause dominant negative effects in genes retained through dosage compensation. Thus, as we discuss subsequently, dosage compensation is expected to yield very different evolutionary signals from those generated by neofunctionalization and subfunctionalization.

Selection for Genetic Redundancy Another mechanism that has been proposed for the retention of duplicate genes is that of serving as a backup copy and, as interactions diverge, playing a role in providing genetic redundancy to generate a more robust system. Under this mechanism, duplicated genes play a buffering role as backup copies for future mutation. The expectation of this mechanism is strong negative selection on coding sequence and function, and it does not explain the burst of substitutions that are typically observed after a duplication event. Further, it has been argued that although the most robust systems are those in chordates, the small effective population sizes and low mutation rates in chordates would not provide strong enough selection for such a weak secondary selection type of mechanism (Forster et al., 2006; Elena et al., 2007).

6.4 Interplay Between Mechanisms

Clearly, these mechanisms are not all mutually exclusive, although some clearly are. For example, subfunctionalization of binding interactions or transcriptional domains would not be compatible with dosage compensation as a mechanism. However, if one views neofunctionalizing changes as rare, any mechanism that increases, even temporarily, the retention time of a duplicate gene has the potential to serve as a transition state for neofunctionalization. This has been established most clearly for the interplay between subfunctionalization and neofunctionalization.

7 EXPECTATIONS FOR RETENTION PROFILE AND FOR SUBSTITUTION PROFILE

The different mechanisms present different profiles expected for time-dependent retention probabilities and time-dependent substitution (dN) and selective pressure (dN/dS)

probabilities. Although relaxation of selective constraint and positive selection can be difficult to differentiate, the expectation from both the neofunctionalization and subfunctionalization models is a burst of substitution after duplication and a declining death rate with time (Figure 7). The substitution process will probably include greater levels of substitution when the events occur in the coding sequence than when they occur transcriptionally. The retention process is typically characterized by a Weibull distribution for neofunctionalization and an exponential distribution for periods between 0.02 and 0.15 *dS* units, followed by a concavely declining hazard function after this point, with an initial waiting time for complementary loss events that appears like neutral loss (Hughes and Liberles, 2007).

In contrast, mechanisms that involve retention of the coding sequences will impose immediate negative selection and will not show a burst of substitution. The dosage compensation mechanism will show immediately high retention rates followed by cooperative loss driven by positive selection once one interacting partner is lost. The loss dynamics of the genetic robustness model are less clear but will probably show retention of duplicated genes where loss is more deleterious at higher rates.

It is clear that the dynamics are slightly different between WGD events and SSD events (Maere et al., 2005; Blomme et al., 2006; Hughes and Liberles, 2008). It is unclear at this stage if this is due to the underlying mutational process or to other features of WGDs. In both cases, there does appear to have been a burst of substitution immediately following duplication, consistent with the subfunctionalization and neofunctionalization mechanisms. Following SSD, the retention pattern is clearly Weibull-like in mammalian genomes. Model-based gene family analysis will enable a more detailed description of underlying processes in different families (see Chapter 10).

One pattern that has emerged is that subfunctionalization and/or dosage compensation might be relatively more important in chordates for whole-genome duplication events (Blomme et al., 2006; Hughes et al., 2007), whereas neofunctionalization is relatively more important in chordates for smaller-scale stochastic events (Hughes and Liberles, 2008). Further initial subfunctionalization events result in genes that eventually neofunctionalize (He and Zhang, 2005).

To complement comparative genomic data analysis, lattice and framework modeling systems have been developed to understand both the time-dependent retention profiles observed under different evolutionary mechanisms and the time-dependent *dN*/*dS* ratios observed in different protein regions during different evolutionary mechanisms. These frameworks will enable creation of better models, consistent with different evolutionary scenarios, which can then be tested in gene families and genome-specific data.

8 ROLE OF PROTEIN FUNCTION AND PROTEIN FOLD

It has previously been reported that some protein functions, especially those that function extracellularly, evolve particularly rapidly after a gene duplication event, whereas other functions, such as those with various immune functions, evolve particularly rapidly after a speciation event (Seoighe et al., 2003). In addition to these protein function-specific differences, it has been observed that different protein folds present different dynamics and relative propensities to subfunctionalize and neofunctionalize. It is expected that a protein with a larger surface area (and necessarily a smaller surface

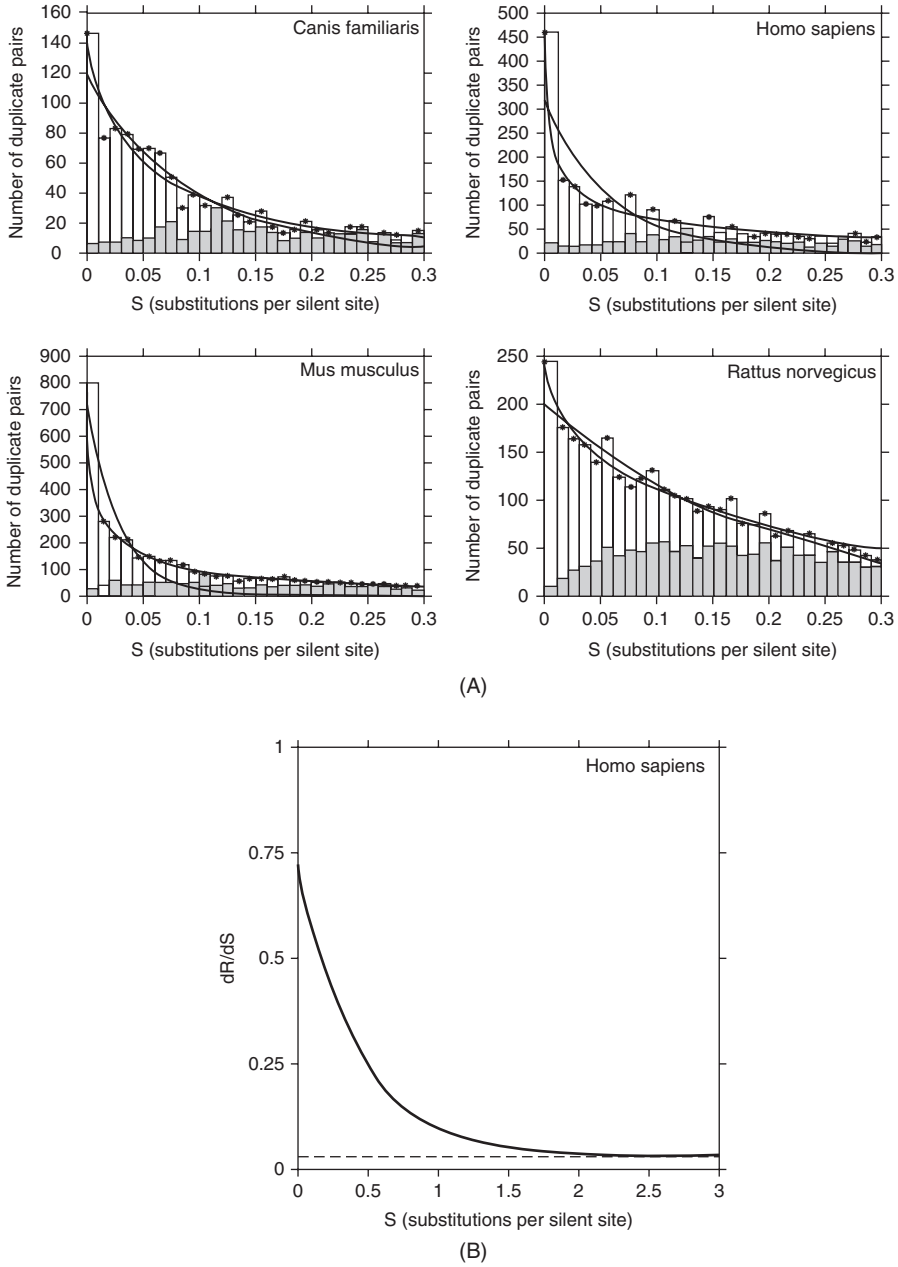

Figure 7 (A) Time (dS)-dependent decay of retained duplicate genes as fit by an exponential and a Weibull distribution in four mammalian genomes. The gray bars show the growth of gene pairs under negative selection with increasing time. (B) In the human genome, the time (dS)-dependent decay of dN/dS for duplicate genes from a relaxed level of selection to an orthologous substitution rate is modeled. (From Hughes and Liberles, 2007.)

area/volume ratio) will have a greater opportunity to evolve new binding functions on its surface. Because chordate proteins tend to be larger than bacterial proteins, this is one possible explanation for the unexpectedly high rates of neofunctionalization in chordates of small effective population size. A possible explanation from the other perspective is that if neofunctionalization is an important process for duplicate gene retention, the folds that are more likely to neofunctionalize rather than nonfunctionalize after gene duplication will be enriched more in species of small effective population size than in species of large effective population size, where a less evolvable fold will also readily neofunctionalize. A prediction of this hypothesis is that chordate genomes will be enriched with the most evolvable folds from a natural distribution, even after correcting for differences in protein surface area.

Ultimately, protein function and protein fold are intertwined, as the protein fold delimits or determines the accessible functions for a protein. From this, disentangling what selection is acting on becomes difficult, and both are clearly important.

9 SPECIES-SPECIFIC DETAILS

Evolution also shows lineage-specific characteristics. Some of this will be due to changes in the underlying process. For example, a lineage-specific change in effective population size or a loss of a DNA repair enzyme will, respectively, alter the relative importance of selection vs. drift on a lineage and will increase the mutation rate (Ota and Penny, 2003). These factors will affect the relative likelihood of different fates for genes duplicated on that lineage. Additionally, lineage-specific selection driven by changes at other loci as well as differences in the environment leading to differential selective pressures on the organism and resulting adaptation will create gene family-specific effects on specific lineages. This is seen in massive lineage-specific expansions or contractions of particularly gene families. The olfactory receptors are now a classic example of this (Glusman et al., 2001).

10 CONCLUSIONS AND LARGER-SCALE EFFECTS

Ultimately, the interplay between population genetic dynamics and biochemistry dictates the fate of duplicate genes. This interplay occurs over many levels of biological organization, and we have tried to integrate this into a larger understanding of the patterns of duplication that we observe in genomes today as well as in their reconstructed history.

Beyond genomes, duplication appears to affect speciation rates and the derivative clade-specific biodiversity. Based on fossil records showing the often rather sudden appearance of morphological variation, Mayr (1963) proposed the founder effect model of speciation, which later contributed to Eldredge and Gould's (1972) development of the punctuated equilibrium theory. Again, these ideas were postulated before the availability of sequence, or whole-genome data, and therefore relied largely on two concepts: (1) the idea that phenotypic variation is essentially an expression of underlying reproductive isolation (biological species concept) or of an independent evolutionary trajectory of a unit of organisms (other species concepts), and (2) the observation that in an evolutionary trajectory, many novel phenotypes seem to occur relatively fast and

without intermediate forms (e.g., the Cambrian explosion). This could not be readily explained by the gradual evolutionary processes proposed by Darwin and others, for this would mean a slow shifting of populations from one equilibrium point to another, eventually launching subpopulations onto their own evolutionary path through a series of "intermediate" stages. Discovery of the imprint of whole-genome duplications in the sequences of a variety of organisms seemed to provide plausible molecular mechanisms for a burst of innovation (i.e., new phenotypes), and a number of authors have speculated about a link between whole-gen(om)e duplication and radiation/speciation [see Roth et al. (2007) for a review]. The reasons for such speculation evolve primarily around the ideas of increased cladogenesis and gen(om)e diversification rates after duplication events. Yet only a few studies have actually correlated rates of speciation with rates of gen(om)e evolution after duplication.

The crux of calculations that include speciation events lies in the very complicated nature of defining a species. Presently, the majority of recognized species is still based on a very narrow definition of "phenotype," which is clearly influenced by our own perceptional biases. While the addition of subsets of molecular sequence data may increase our resolution to distinguish similarities and differences, it may also obscure relationships, due to the opportunistic and disjunct nature of our sampling (on organismal and molecular levels). Thus, it is very likely that we underestimate the number and evolutionary time frames of organismal units on their own evolutionary trajectory. By extension this will obscure the role of ge(nom)e duplication as a process of speciation. For example, recent studies on polyploid plants found that there is definitively a substantial contribution of polyploidy to cladogenesis, yet the authors acknowledge that phylogenetic uncertainties may render this result too conservative (Wood et al., 2009).

Therefore, we may better be served to evaluate the role of gen(om)e duplication on the "first responder" level of populations, taking the dynamics of genealogical history into account. These dynamics are naturally influenced by repetitive population size variation and fragmentation at different spatiotemporal scales, and although these parameters cannot be observed directly on an evolutionary scale, they can be included in models of evolution.

From the standpoint of population dynamics, and depending on reproductive strategies, each duplication event may affect one individual or a local set of siblings of a population. Assuming that the bearer(s) of duplication are reproductively fit (i.e. the duplication event is selectively neutral, advantageous, or only slightly deleterious), either one or both potential parents from that subpopulation may carry the duplicated set of genes. However, duplication may introduce immediate reproductive barriers in a subpopulation, thus decreasing the likelihood of an F_1 from mixed parental ploidy or gene number (Kelleher et al., 2007). Furthermore, it would mean that inbreeding of siblings with duplicated genes is likely when population connectivity is high (low dispersal). In this particular case, each subsequent perpetuation of the duplicated genetic line may then best be understood and modeled as a founder effect scenario. However, far too few studies exist regarding the actual potential of reproductive isolation of individuals after duplications. It stands to reason that there are varying degrees of "severity" in duplications of genetic material, ranging from internal gene duplications to whole-genome duplications.

The cessation of gene flow between segments of a population does not eliminate their competition in a population-like paradigm. If the initial duplicates are identical (as

in the extreme case of whole-genome duplication), the phenotype of the organism may initially be similar. In this scenario, even if the individual(s) with duplication events are reproductively isolated, they run the risk of elimination from the "population."

Organismal groups seem to have a vastly different tolerance to polypoloidization and duplication events. For instance, while many vertebrate groups are certainly paleopolyploid, they have largely returned to diploidy, and higher vertebrates (e.g., humans) seem to be particularly negatively affected by duplications. Plants, on the other hand, appear to perpetuate polyploidy on a much more frequent basis. It is currently not well understood if the process of diploidization is due to changes in master chromosome-pairing genes or through a more general loss of pairing ability between homeologs due to loss of genes (Semon and Wolfe, 2007). Although these are certainly molecular mechanisms to lose polyploidy, the time frame of the loss should again be a function of population size, and cannot be divorced from the organism's life history parameters. For example, in slowly reproducing organisms with few offspring, the likelihood of losing selectively neutral duplicates is higher than in an organism with many offspring (which may all carry the duplication).

Although still in its infancy (especially for nonmodel organisms), the simultaneous study of genomewide variation within and between species (population genomics) may reveal new mechanisms influencing the faith of genomic duplications. Particularly interesting is the additional insight into the variability of noncoding sequence across individuals, reducing the bias associated with population genetic analyses based on targeted protein-coding genes (Begun et al., 2007). Additionally, and in line with the issues mentioned previously, individuals of populations are the first responders to duplications, and such comprehensive data may allow for a better understanding of the fitness effects associated with different levels of duplications.

Acknowledgments

We wish to thank Johan Grahnen for providing Figure 6. This work was supported by an NIH-INBRE award to University of Wyoming.

REFERENCES

Aury J, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al. 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324:1720–1723.

Baird-Titus JM, Clark-Baldwin K, Dave V, Caperelli CA, Ma J, Rance M. 2006. The solution structure of the native K50 bicoid homeodomain bound to the consensus TAATCC DNA-binding site. J Mol Biol 356:1137–1151.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: whole genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 millionyears of vertebrate evolution. Genome Biol 7:R43.

Bogan AA, Thorn KS. 1998. Anatomy of hot spots in protein interfaces. J Mol Biol 280:1–9.

Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. 2006. Insights into protein-protein interfaces using a Bayesian network prediction method. J Mol Biol 362:365–386.

Bradley J, Baltus A, Skaletsky H, Royce-Tolland M, Dewar K, Page DC. 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. Nat Genet 36:872–876.

Brown JR, Doolittle WF. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Nat Acad Sci USA 92:2441–2445.

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. 2004. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13:190–202.

Chakrabarti P, Janin J. 2002. Dissecting protein–protein recognition sites. Proteins 47:334–343.

Churbanov A, Winters-Hilt S, Koonin EV, and Rogozin IB. 2008. Accumulation of GC donor splice signals in mammals. Biol Direct 3:30.

Copley SD. 2003. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr Opin Chem Biol 7:265–272.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Ekker SC, Jackson DG, von Kessler DP, Sun BI, Young KE, Beachy PA. 1994. The degree of variation in DNA sequence recognition among four *Drosophila* homeotic proteins. EMBO J 13:3551–3560.

Eldredge N, Gould SJ. 1972. Punctuated equilibria: an alternative to phyletic gradualism In Schopf TJM (ed.), *Models in Paleobiology*. San Francisco: Freeman Cooper, pp. 82–115.

Elena SF, Wilke CO, Ofria C, Lenski RE. 2007. Effects of population size and mutation rate on the evolution of mutational robustness. Evolution 61:666–674.

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

Forster R, Adami C, Wilke CO. 2006. Selection for mutational robustness in finite populations. J Theor Biol 243:181–190.

Garcia-Fernández J 2005. The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6:881–892.

Gaucher EA, Miyamoto MM, Benner SA. 2003. Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein. Genetics 163:1549–1553.

Glusman G, Yanai I, Rubin I, Lancet D. 2001. The complete human olfactory subgenome. Genome Res 11:685–702.

Guharoy M, Chakrabarti P. 2005. Conservation and relative importance of residues across protein–protein interfaces. Proc Natl Acad Sci USA 102:15447–15452.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148:993–1003.

Hanes SD, Brent R. 1989. DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. Cell 57:1275–1283.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, Danford TW, Hannett NM, Tagne J, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104.

He X, Zhang J. 2005. Gene complexity and gene duplicability. Curr Biol 15:1016–1021.

Horowitz NH. 1945. On the evolution of biochemical syntheses. Proc Natl Acad Sci USA 31:153–157.

Hughes T, Liberles D. 2007. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol 65:574–588.

Hughes T, Liberles DA. 2008. Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. J Mol Evol 67:343–357.

Hughes T, Ekman D, Ardawatia H, Elofsson A, Liberles DA. 2007. Evaluating dosage compensation as a cause of duplicate gene retention in *Paramecium tetraurelia*. Genome Biol 8:213.

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

Jin L, Kryukov K, Clemente JC, Komiyama T, Suzuki Y, Imanishi T, Ikeo K, Gojobori T. 2008. The evolutionary relationship between gene duplication and alternative splicing. Gene 427:19–31.

Johnston CR, O′Dushlaine C, Fitzpatrick DA, Edwards RJ, Shields DC. 2007. Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneoulsy with gene duplication. Mol Biol Evol 24:315–323.

Kelleher ES, Swanson WJ, Markow TA. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet 3:e148.

Light S, Kraulis P. 2004. Network analysis of metabolic enzyme evolution in *Escherichia coli*. BMC Bioinf 5:15.

Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M. 2002. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3:40.

Lusk RW, Eisen MB. 2008. Use of an evolutionary model to provide evidence for a wide heterogeneity of required affinities between transcription factors and their binding sites in yeast. *Pacific Symposium on Biocomputing*, pp. 489–500.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

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.

Marina A, Waldburger CD, Hendrickson WA. 2005. Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein. EMBO J 24:4247–4259.

Mayr E. 1963. *Animal Species and Evolution*. Cambridge, MA: Harvard University Press.

Miyazaki J, Kobashi N, Nishiyama M, Yamane H. 2001. Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through alpha-aminoadipate. J Bacteriol 183:5067–5073.

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133:1277–1289.

Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer-Verlag.

Ota R, Penny D. 2003. Estimating changes in mutational mechanisms of evolution. J Mol Evol 57(Suppl 1):S233–S240.

Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK. 1999. Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. Nature 397:714–719.

Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. 2009. Physicochemical principles that regulated the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci U S A 106:10159–10164.

Percival-Smith A, Müller M, Affolter M, Gehring WJ. 1990. The interaction with DNA of wild-type and mutant fushi tarazu homeodomains. EMBO J 9:3967–3974.

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.

Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, et al. 2003. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31:3625–3630.

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. 2007. Evolution after gene duplication: models, mechanisms, systems, and organisms. J Exp Zool B Mol Dev Evol 308:58–73.

Sato A, Mayer MW, Tichy H, Grant PR, Grant BR, Klein J. 2001. Evolution of Mhc class II B genes in Darwin′s finches and their closest relatives: birth of a new gene. Immunogenetics 53:792–801.

Semon M, Wolfe KH. 2007. Reciprocal gene loss between tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet 23:16–20.

Seoighe C, Johnston CR, Shields DC.*.baty 2003. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. Mol Biol Evol 20:484–490.

Shultzaberger RK, Chiang DY, Moses AM, Eisen MB. 2007. Determining physical constraints in transcriptional initiation complexes using DNA sequence analysis. PLoS ONE 2:e1199.

Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT. 2008. Rewiring the specificity of two-component signal transduction systems. Cell 133:1043–1054.

Treisman J, Gönczy P, Vashishtha M, Harris E, Desplan C. 1989. A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell 59:553–562.

Tucker-Kellogg L, Rould MA, Chambers KA, Ades SE, Sauer RT, Pabo CO. 1997. Engrailed (Gln50→Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. Structure 5:1047–1054.

Tvrdik P, Capecchi MR. 2006. Reversal of *Hox1* gene subfunctionalization in the mouse. Dev Cell 11:239–250.

Vick JE, Gerlt JA. 2007. Evolutionary potential of (beta/alpha)8-barrels: stepwise evolution of a "new" reaction in the enolase superfamily. Biochemistry 46:14589–14597.

Wagner A. 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365–1374.

Wood T, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg L. 2009. The frequency of polyploidy speciation in vascular plants. Proc Natl Acad Sci USA 106:13875–13879.

Xing Y, Lee CJ. 2005. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genet 1:e34.