

# Evolution After Gene Duplication: Models, Mechanisms, Sequences, Systems, and Organisms

CHRISTIAN ROTH<sup>1,2</sup>, SHRUTI RASTOGI<sup>1,3</sup>, LARS ARVESTAD<sup>4</sup>,  
KATHARINA DITTMAR<sup>1</sup>, SARA LIGHT<sup>2</sup>, DIANA EKMAN<sup>2</sup>,  
AND DAVID A. LIBERLES<sup>1,3\*</sup>

<sup>1</sup>*Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071*

<sup>2</sup>*Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden*

<sup>3</sup>*Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

<sup>4</sup>*School of Computer Science and Communication, Albanova University Center, Royal Institute of Technology, 10044 Stockholm, Sweden*

**ABSTRACT** Gene duplication is postulated to have played a major role in the evolution of biological novelty. Here, gene duplication is examined across levels of biological organization in an attempt to create a unified picture of the mechanistic process by which gene duplication can have played a role in generating biodiversity. Neofunctionalization and subfunctionalization have been proposed as important processes driving the retention of duplicate genes. These models have foundations in population genetic theory, which is now being refined by explicit consideration of the structural constraints placed upon genes encoding proteins through physical chemistry. Further, such models can be examined in the context of comparative genomics, where an integration of gene-level evolution and species-level evolution allows an assessment of the frequency of duplication and the fate of duplicate genes. This process, of course, is dependent upon the biochemical role that duplicated genes play in biological systems, which is in turn dependent upon the mechanism of duplication: whole genome duplication involving a co-duplication of interacting partners vs. single gene duplication. Lastly, the role that these processes may have played in driving speciation is examined. *J. Exp. Zool. (Mol. Dev. Evol.)* 308B:58–73, 2007. © 2006 Wiley-Liss, Inc.

---

**How to cite this article:** Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool. (Mol. Dev. Evol.)* 306B:58–73.

---

According to Ohno's classic view, the evolution of genes and genomes is typically conservative in the absence of gene duplication (Ohno, '70). Large-scale genome sequencing has enabled us to test and characterize this classic hypothesis on the importance of gene duplication to the evolution of novelty at multiple levels.

Whole genome duplication (WGD) can be seen as large-scale gene duplication and this large potential source of novelty can probabilistically (neither completely randomly nor deterministically) be played out in different ways in different populations, potentially leading to increased rates of speciation. However, the advantage of WGD is magnified due to the potential advantage of duplicating entire pathways, enabling pathway-level innovation and retention of interacting

partners. However, this involves an interplay between the capabilities of developing new function at the individual protein level coupled to that at the pathway level (or at least in relation to different individual proteins that may interact biochemically or genetically). This means that selection acting at the level of the individual protein (for example, in its ability to fold or catalyze a reaction) is important, as is selection

---

Grant sponsor: Formas, Swedish Agricultural Research Council; Grant sponsor: Swedish Foundation for Strategic Research; Grant sponsor: FUGE, the Norwegian Functional Genomics Platform.

\*Correspondence to: D.A. Liberles, Department of Molecular Biology, University of Wyoming, Laramie, WY 82071 USA.  
E-mail: liberles@uwyo.edu

Received 17 March 2006; Revised 3 May 2006; Accepted 4 June 2006  
Published online 12 July 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21124.

acting upon pathway-level function dependent upon the correlated properties of individual proteins.

Currently, the mechanism of retention of duplicate gene copies is an issue of debate in the literature. This plays out in several ways. One key question is the relative importance of neofunctionalization and positive selection (Ohno, '70) vs. that of subfunctionalization and purely neutral processes (Hughes, '94; Force et al., '99; Lynch and Force, 2000) and is now leading to the emergence of hybrid views (He and Zhang, 2005; Rastogi and Liberles, 2005). A second debate involves the relative importance of gene expression evolution vs. coding sequence function evolution (see Liberles, 2005 for a discussion). Ultimately, this all feeds into generating an understanding of if and how gene duplication leads to the emergence of new phenotypes and of biodiversity.

Starting from first principles, our knowledge of protein chemistry and of population genetics makes predictions that are testable. Models based upon these types of principles will be described, including their predictions. However, from genomes, models for comparative genome analysis can also predict, as well as characterize, both general and lineage-specific trends of gene duplication. Further, the most recent duplicates can give a snapshot into the evolutionary process that is being predicted by these models. An analysis of this using the Arabidopsis genome will be presented. This mechanistic picture of course ties in with processes at the whole genome level that feed through systems biology (selection on proteins for their ability to interact with other proteins and to function in coordination with other proteins) to organismal biology (Ardawatia and Liberles, submitted manuscript). The whole picture describing the importance of gene duplication to the evolution of species, given our current level of understanding, is presented.

## METHODS

A systematic analysis of recent duplication in the Arabidopsis lineage was carried out. An all-against-all BLAST search (Altschul et al., '97) with an  $e$ -value cutoff of  $10e^{-20}$  was run on the TAIR6 *Arabidopsis thaliana* genome release (Arabidopsis Genome Initiative, 2000). For each hit,  $K_a$  and  $K_s$  were calculated using a previously described estimation method (Liberles, 2001) and global PAM distances were calculated using

Darwin (Gonnet et al., 2000). Pairs were considered of recent origin if  $K_s < 0.05$  and  $PAM < 30$ . These criteria encompassed most recent duplicates (as judged by the  $K_s$  spectrum), but was not too relaxed as a threshold to include duplication events that were known to occur in other Brassica species (only one such example was found). The PAM cutoff eliminated spurious examples of genes with  $K_s = 0$ , but  $PAM > 30$ . Families were formed by single linkage clustering from all pairs. GO annotations for all significant pairs were collected from the TAIR GOSLIM-file (Berardini et al., 2004). Additionally, distance trees were calculated for all recent duplicates where homologs were identified in other Brassica species using Darwin (Gonnet et al., 2000).

## RESULTS AND DISCUSSION

### *Models for individual duplicate gene evolution*

Gene Duplication is one of the primary driving forces in the evolution of genes and genomes. The idea that gene duplication has a fundamental role in the origin of phenotypic diversity has motivated numerous proposals to explain how a new gene copy can emerge from its predecessor, and evolve a novel function. We will begin an analysis of this process at the level of the individual protein-encoding gene and move up to the level of the whole organism.

Models for the evolution and retention of duplicate gene copies differ in several fundamental respects, including the time of origin of the functional novelty, the types of mutational events that drive the crucial phase of early establishment, and the rates of molecular evolution in the two copies (equal vs. asymmetric). Ohno postulated that gene duplication creates redundant loci that are free to accumulate mutations that lead to loss or gain of function, as long as one of the copies still performs the essential ancestral task (Ohno, '70). Under his model, neofunctionalization was identified as the primary mechanism for preservation of duplicate copies of genes. Subsequently, Hughes and Lynch proposed that preservation of duplicate copies can also occur by subfunctionalization, i.e., by the accumulation of degenerative mutations causing complementary loss of subfunctions in the two members of the pair (Hughes, '94; Force et al., '99; Lynch and Force, 2000). Lynch also proposed an alternative model, called DDC or the duplication-degeneration-complementation model, according to which degenerative mutations in

regulatory elements can increase the probability of duplicate gene preservation. In this process, the usual mechanism of duplicate gene preservation is the partitioning of ancestral expression patterns rather than biochemical functions or the evolution of new functions (Force et al., '99).

Although the DDC process is based entirely on degenerative mutations, there are at least three ways in which it may play a significant role in creative evolutionary processes. First, by stabilizing duplicate genes in the genome, the DDC process extends the time period during which genes are exposed to natural selection, thereby enhancing the chance that rare beneficial mutations to novel functions may arise (as compared to the situation under the classical model, where a gene is removed from selection once it has become nonfunctionalized). Second, the partitioning of gene expression patterns by the DDC process may reduce the pleiotropic constraints operating on single-gene loci, thereby allowing natural selection to more closely tune the duplicate members of a pair to their specific subfunctions. Third, gene duplicates that have unresolved subfunctions at the time of a reproductive isolation event provide a powerful mechanism for the development of reproductive incompatibility, i.e., speciation. This last process can also enable the emergence of novel functions in genetically or biochemically interacting genes/proteins in the genome.

Based upon population genetic theory, a statistical framework has been presented for the probability of retention of a duplicate gene in a genome (Lynch et al., 2001). In examining the influence of various aspects of gene structure, mutation rates, degree of linkage, and population size ( $N$ ) on the joint fate of a newly arisen duplicate gene pair from an ancestral locus, it was found that unless there is active selection against duplicate genes, the probability of permanent establishment of such genes is usually no less than  $1/(4N)$  (half of the neutral expectation), and it can be orders of magnitude greater if neofunctionalizing mutations are common. Relative to subfunctionalization, neofunctionalization is expected to become a progressively more important mechanism of duplicate-gene preservation in populations with increasing size. However, even in large populations, the probability of neofunctionalization scales only with the square of the selective advantage. Ultimately, protein structure (as discussed below) may play a role in amplifying the importance of positive and negative selection

through the covarian nature of the folding and fitness landscapes. Tight linkage also influences the probability of duplicate-gene preservation, by increasing the probability of subfunctionalization but decreasing the probability of neofunctionalization (with a high recombination rate doing the reverse).

The Lynch model supports the idea that deleterious mutations that can never be fixed in single-copy genes might often accumulate in a nearly neutral fashion shortly after gene duplication. In principle, the chance of accumulation of a sufficient number of such mutations in one copy might eventually lead to its complete nonfunctionalization, removing it from the eyes of natural selection, and setting it on the course of pseudogenization. By contrast, the small subset of duplicate genes that escape this fate (Lynch and Conery, 2000) might owe their initial preservation to subfunctionalization by a more balanced distribution of degenerative mutations, which can then be followed by a phase of adaptive-conflict resolution, neofunctionalization or further subfunctionalization.

Large-scale analyses, based upon the ratio of nonsynonymous to synonymous nucleotide substitution rates (Roth et al., 2005; Roth and Liberles, 2006) or MacDonald-Kreitman statistics (Fay et al., 2001) have indicated small to intermediate degrees of positive selection (adaptive substitutions) in embryophytes and in mammals, but these clearly do not represent the majority of substitutions. In such studies, there appear to be specific positions in protein-encoding genes, rather than the genes as a whole that are under positive selection (Roth and Liberles, 2006). Even examining substitution as a neutral walk through sequence in a folded protein (ignoring positive selection) has shown the substitutional process to have fairly complex dynamics (Bastolla et al., 2003). From this, it is relevant to examine population genomic phenomena, like the fates of duplicated genes, in the context of physical models of proteins under increasingly realistic conditions. A true comparative genomic benchmarking of modifications to molecular function that influence organismal fitness would require a combination of high throughput in-vitro biochemistry, genetic manipulation, and in-vivo competition experiments to properly assess where subfunctionalization and neofunctionalization have occurred. This type of analysis is just beginning on a case-by-case basis.

Analytical models often assume a random mapping between genotype and phenotype, because

the correlation among mutational effects proves to be mathematically complex. However, such correlations are crucial to understanding neutral mutations and mutational stability. To address these issues, many recent theoretical and computational efforts have focused on constructing models of sequence-structure mapping for proteins motivated by polymer physics theory (Williams et al., 2001; Deeds et al., 2003; Wroe et al., 2005). Because of the immense sizes of the systems, all these models involve significant simplifications (but some of which may be realistic). Many proteins maintain their native structures while undergoing single and double mutations at many different sites. Simple protein models have shown that proteins are surprisingly robust to site mutations, undergoing little change in structure, stability and function by significant numbers of substitutions, supporting what has been seen through homology modeling with more sophisticated models and structure determination of distant orthologs (Taverna and Goldstein, 2002a). Another vital insight is the importance of kinetic accessibility of the native structures as an evolutionary selection criterion. Structures that are more kinetically accessible are expected to be more plastic with regard to sequence and function (Taverna and Goldstein, 2002b).

A structurally constrained model based upon the neutral theory has been developed (Bastolla et al., 2003). The neutral model has two key underlying assumptions. First, most mutations in protein sequences are either disruptive and eliminated by negative selection or neutral in that they leave the protein active and their effect on fitness is much smaller than the inverse of the effective population size. Second, the rate of appearance of neutral mutations is constant throughout evolution. The neutral model predicts that the rate of fixation of neutral mutations in an evolving population is equal to the rate of their appearance, independent of the population size, because the number of appearing mutations is proportional to the population size and the probability of their fixation is inversely proportional to it. Even with a strictly neutral model, the dynamics of the mutational process are extremely complex.

Another modeling approach involves simple exact models, which address general principles of evolution as they permit the exhaustive enumeration of both sequence and structure (conformational) spaces (Wroe et al., 2005). These physical models can then be applied in large scale to make predictions about the mapping

between substitutions and molecular phenotypes seen in genomes.

The evolutionary dynamics of protein function can also be studied using models based upon protein lattices (Williams et al., 2001). From this study, population dynamics strongly influence the frequency of variously observed structures, where the need to fold into a stable structure and function (binding) can be used as a selection criterion to explain some of the properties common to all proteins. These simple models, increasingly, can be extended into more sophisticated models (for example, all-atom models that incorporate van der Waals effects, electrostatic interactions, amino acid rotamer, and other important physical principles) at a decreasing computational cost. However, the simplicity of a purely statistical potential function still allows one to sample over many types of structures, and to replicate results. Despite their approximate character, they feature a unique sequence-structure relationship akin to that of real proteins.

Since proteins are robust to site mutations and plastic in nature in that they accept mutations without destroying the fold (Taverna and Goldstein, 2002a,b; Shakhnovich et al., 2005), the development of an evolutionary model based upon population genetics theory together with the evolution of lattice (or real protein-encoding) genes based upon either purely statistical or physical (force field) energy constraints is an area of growing interest. The evolutionary constraints for proteins include that they should perform a function and they must be stable enough to perform that function reliably while resisting unfolding, aggregation, and proteolysis. Protein functionality has been modeled by designing ligands which bind to two overlapping binding sites in a functional protein with an enzymatic or binding role (Braun and Liberles, 2003; Rastogi and Liberles, 2005). In these models, genes can be duplicated and through evolution, nonfunctionalize, subfunctionalize, or neofunctionalize. For example, the evolution of duplicates of a pleiotropic ancestral gene under the subfunctionalization model can be evaluated by incorporating an idealized sequence-function mapping with enzyme-substrate binding affinity related to a hydrophobic binding pocket surrounded by a general polar surface of the protein. The measure of pleiotropic function was defined by the specificity of the enzyme in the presence of two competing substrates (Braun and Liberles, 2003). This is further incorporated in a model with explicit peptide

binding, which suggests under a set of evolutionary constraints that subfunctionalization of duplicate genes is important, but has a short half-life as a transition state to neofunctionalization (Rastogi and Liberles, 2005).

A further extension of lattice protein models has examined the retention of functionality in homologous recombinants of proteins (Xu et al., 2005). From this study, exceptional structures with many sequence options evolved quickly and tended to retain functionality, even in highly diverged recombinants; in contrast, the more common structures with fewer sequence options evolved more slowly, but the fitness of recombinants dropped off rapidly as homologous proteins diverged.

While most of the processes modeled have attempted to examine general aspects of protein evolution, lineage-specific processes affecting gene duplicates, like repeat-induced point mutation (RIP) in filamentous fungi, can have major impacts on the proteome of species, which has warranted modeling (Braun and Liberles, 2004). According to most views, this mechanism acts primarily as a conservative genomic safe guard by introducing stop codons into duplicated nucleotide sequences and thereby pre-empting dosage effects. However, it also performs nonsynonymous nucleotide substitutions. This process impacts the hydrophobic vs. polar (HP) composition of proteins, based upon the mutational process and the genetic code. To test for a genome-wide directional shift in the content of filamentous fungi proteins, a mapping between HP composition and tertiary structural phenotype was performed by using residue packing fraction and predicted thermodynamic stability, leading to the conclusion that RIP tends to generate more compact and thermodynamically stable tertiary structures.

While direct consideration of physical properties is clearly important to understanding the process of gene duplication on a genomic scale, appropriate models to do this are still being developed. From first principles models, predictions can be made about the probability of retaining duplicate genes under various sets of evolutionary conditions. Genomic-scale models are simultaneously available for measurement of parameters seen as evolution has evolved to generate the genomes that have been sequenced. The current state of the art of modeling on a genomic scale is presented in the next section. Ultimately, these two types of models may be integrated to increase the power of comparative genomics.

### *Models for comparative genomics*

Probabilistic models have been very successful in phylogeny (Yang, '97; Ronquist and Huelsenbeck, 2003) and it is desirable to be able to use ML and MCMC approaches also for tree reconciliation analysis. Hence, probabilistic models of gene family evolution are required that put gene family evolution in the context of species evolution. According to today's widespread tools for probabilistic phylogenetic analysis, any tree is as good as another and no prior beliefs on tree structure are imposed. In the context of gene families, such a model does not express what we know about gene family evolution.

Rannala and Yang ('96) departed from common practice and formulated a prior on the phylogenetic trees that would influence the likelihood of a phylogeny. Their prior was formulated as a birth–death model (Kendall, '48) that describes how taxa underwent speciation and loss and capture the fact that some trees are more likely to occur than others. However, the model does not regard the fact that there might be another underlying process shaping the phylogeny. Common intuition, on the other hand, would say that it is likely that a gene tree reflects the species tree.

Arvestad et al. (2003) formulated the first probabilistic model of gene evolution that took the species tree into account and developed algorithms for computing the probability of observing a gene tree and an associated reconciliation to a species tree  $S$ . In this model, a gene starts at the root of  $S$  and evolves downward over the edges in  $S$ . We assume that we know the speciation times in  $S$ . A gene is duplicated with a rate  $\lambda$ , in which case the gene is turned into two independent evolving copies. Gene losses occur with a rate  $\mu$  which, naturally, terminates a gene from continuing to evolve. When a gene arrives at a speciation in  $S$ , it continues to evolve in two independent copies toward each child of  $S$ . Thus, we have independent birth–death processes over the edges in  $S$ .

This gene evolution model is quite simple, but carries some powerful implications. Contrary to the parsimony approach to tree reconciliations (Goodman et al., '79; Page and Charleston, '97; Zhang, '97; Berglund-Sonnhammer et al., 2006), alternative reconciliations are modeled and can be taken into account in for example orthology analysis. An interesting application of the model is that a set of gene trees could be used to infer  $S$  and its edge times, thus dating speciations

without worrying about substitution rates. Moreover, the fact that duplication and loss rates of genes are explicitly modeled allows a more direct estimation of these parameters than has been carried out previously (Lynch and Conery, 2000; Cotton and Page, 2005).

Similar questions have been addressed without looking at more than gene family size in different genomes (Gu and Zhang, 2004; Reed and Hughes, 2004; Hahn et al., 2005). Such a model can be used to predict gene count in ancestral species without modeling the actual phylogeny. While giving up on phylogeny is sacrificing some analytic power, the simplification has made it possible to also include models of lateral transfer or gene innovation (Novozhilov et al., 2005; Csuros and Miklos, 2006) and a separation of gene and genome duplications (Maere et al., 2005).

The models above are lacking an integral part of gene evolution: sequence analysis. However, it has been shown how to integrate the gene evolution model with standard models of sequence evolution (Arvestad et al., 2004). A first advantage is obvious, in that a gene tree is no longer necessary for analysis. In fact, this model allows for estimating the gene tree from sequence data while allowing a species tree to affect the estimate. Thus, instead of figuring out if an estimated gene tree agrees well with a species tree, the reconciliation is an integrated part of the model from the start.

A second powerful advantage is that an integrated model can give orthology probabilities for gene pairs without actually deciding on a gene tree. By integrating over all possible gene trees and reconciliations, orthology probabilities can be calculated. While such an integral is impractical to calculate directly, MCMC methods (Gilks et al., '96) offer a natural framework in which the problem can be addressed.

One of the main advantages with probabilistic models is that the probabilistic framework makes integration and extension of models straightforward, at least in a mathematical sense. Computationally and algorithmically, integration is not necessarily straightforward. The integrated model poses a computational challenge and is today only really applicable on smaller datasets including up to 20 genes, but there is room for both algorithmic and implementational improvements to the method.

More interesting opportunities lie in extending models to capture natural evolution with more mechanistic detail. For instance, it is currently

required to know all of the genes in a family for each species analyzed for the model to give reasonable results. With low-coverage genome sequencing being common, it is to be expected that some genes are missed. By modeling gene sampling as a function of genome coverage, biases in parameter estimations could be offset.

Further, models that explicitly consider the biochemistry and physical chemistry of the duplication and loss process as well as of the sequence evolution of proteins that fold and function according to physical parameters as described in the previous section (also see for example, Dutheil et al., 2005; Stern and Pupko, 2006; or an approach that uses methods from Berglund et al., 2005 to model direct physical interactions using either purely statistical models or force fields) are a logical extension of current models.

Additionally, birth and death models have typically examined a single birth and death parameter for each gene family (as described above), or examined processes across gene families to examine the relative influence of rare large-scale duplication vs. more common smaller duplication events (see for example Dehal and Boore, 2005). In the future, the integration of lineage-specific methods (see Hahn et al., 2005) to detect family and lineage-specific innovation with both sequence-based approaches and gene-tree to species-tree coupling will enable a simultaneous analysis of duplication and sequence divergence in the context of the tree of life. This will then ultimately enable a further linking of the duplication process with existing methods for the analysis of other processes in gene family evolution in the context of species trees (see Roth et al., 2005), ultimately enabling a greater understanding of what makes each species unique (see for example Francino, 2005 for an interesting hypothesis).

As modeled and detected using the above methods, both WGD and individual gene duplication have been important in shaping genomes. In the context of pathway and signaling evolution, the capabilities of WGD are greater because of the capability of simultaneously duplicating interacting partners. However, in most lineages, such WGD events seem to have been rarer than single gene duplication events (as normalized on a per gene per unit time basis). The next section gives an overview of the types of gene duplication that are actually observed, including the functional consequences, using the Arabidopsis genome as an example.

### ***Large- and small-scale duplication in individual lineages***

In plant evolutionary history, genome duplications have been relatively common. Specifically focusing on Arabidopsis, the species has experienced at least three ancient polyploidy events (Vision et al., 2000; Bowers et al., 2003). In analyzing Arabidopsis and other plant genomes, the most likely conclusion is that most angiosperms are to some extent polyploid (Soltis, 2005). By contrast, most vertebrate lineages have undergone only one or two large-scale genome duplication events in their ~500 million year history (Postlethwait et al., '98; Gu et al., 2002; McLysaght et al., 2002), and extant polyploids are rare (Gallardo et al., '99; Soltis and Soltis, '99).

The extra WGD in plants has been linked to an observation that grass genomes have fewer highly conserved regulatory elements that are lost in complementary ways between duplicate gene pairs in comparison with mammals to propose that subfunctionalization has been rampant in plant genomes (Lockton and Gaut, 2005). A further, more detailed analysis is needed, with explicit consideration of expression patterns and functions of the ancestral state, to differentiate between the various models of gene duplication fate.

The more recent duplications in the Arabidopsis genome come from short segmental and single gene duplication events (mostly tandem duplications). This process, characterized in Table 1, created many of the large gene families. When analyzing these families, different molecular and cellular functions seem to be duplicated and retained at different rates compared to the genome as a whole, given a random chance of duplication and retention (see Fig. 1). Mitochondrial carrier proteins and heat shock transcription

factors show low retention rates of duplicates in general. MYB transcription factors on the other hand are more often retained after segmental duplications. High tandem duplication can also be detected for the defense-related gene families germin and major latex protein, as well as the chlorophyll *a-b* binding proteins (Cannon et al., 2004).

Lynch and Conery (2000) calculated a birth rate for *A. thaliana* duplicates of 0.002/gene/MY or about 60 genes per million years. The half live of a gene duplicate was estimated to be about 17.3 MY in the same study. Using these numbers we can expect 250 *A. thaliana* specific duplicated genes compared to its closest relative *A. lyrata* (~5.2 MY; Koch et al., 2000). As described below, the observed number is significantly higher.

A common form of gene duplicates is tandemly duplicated genes, created by unequal crossing-over. A first example for such events is the *A. thaliana* trypsin inhibitor locus ATTI with six tandem duplicated genes (Clauss and Mitchell-Olds, 2004). Several factors (linkage disequilibrium, low level of diversity between ATTI1 and ATTI2, and a presence/absence polymorphism for ATTI5) led to a rejection of a homogeneous neutral model of evolution for these duplicates. These duplicates were also systematically differently expressed when presented with a herbivore attack.

Another example of functional divergence in a tandem duplicated locus is in the glucosinolate biosynthesis pathway. The two 2-oxoglutarate-dependent dioxygenases AOP2 and AOP3 catalyze two different reactions, creating either alkenyl (GS-ALK) or hydroxyalkyl (GS-OHP) glucosinolates. *A. thaliana* ecotypes can be categorized into GS-ALK (expressing AOP2), GS-OHP (expressing AOP3), or methylsulfinylalkyl (precursor) accumulating (no functional expression) plants (Kliebenstein et al., 2001).

Two additional recently duplicated genes (PreP1 and PreP2) have been described biochemically (Bhushan et al., 2005). These two metalloproteases degrade small peptides in mitochondria and chloroplasts. The two sequences have a PAM distance of 14 and a  $K_s$  of 0.417. Using age estimates provided by Blanc and Wolfe (2004), an age for the duplication of about 15 million years is obtained, which places it close to the divergence of the Brassica and Arabidopsis lineages. These two genes show differential expression in a tissue-specific manner. In addition, they seem to have a different enzymatic specificity (Bhushan et al., 2005).

TABLE 1. The distribution of recently duplicated genes on the five Arabidopsis thaliana chromosomes is shown

	# of genes	# of families	# of tandem duplicates
Chromosome 1	135	71	34
Chromosome 2	92	41	8
Chromosome 3	113	56	13
Chromosome 4	157	44	87
Chromosome 5	138	77	45

The first column gives the number of genes from a chromosome that are involved in a duplicate pair, column 2 contains the number of families that are formed by these genes, and column 3 shows how many of the gene pairs are tandem duplicates (separated by less than 20 loci).

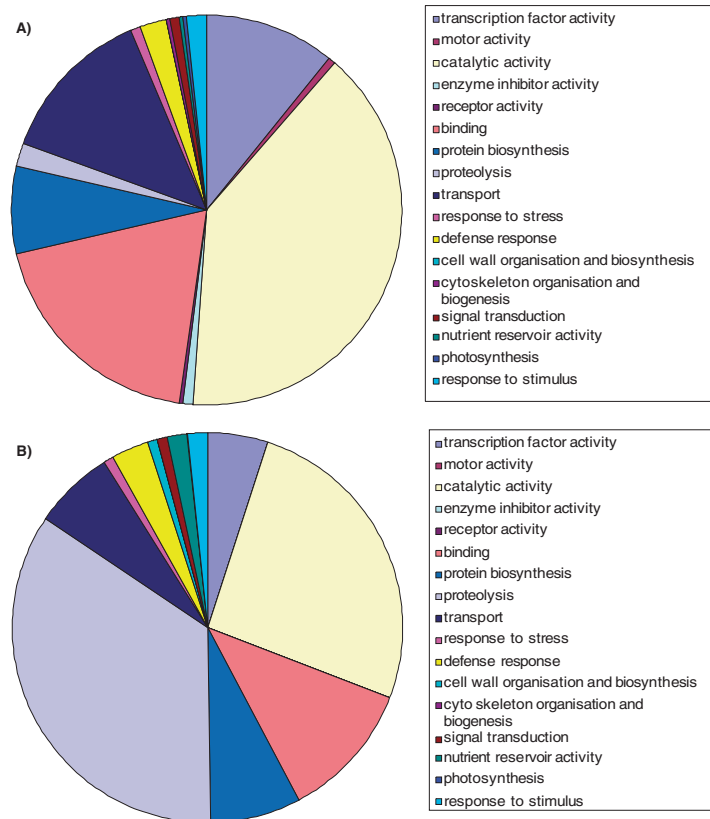


Fig. 1. (a) The whole genome of *Arabidopsis thaliana* (27,142 genes) was compared to the GO database (Berardini et al., 2004) and genes with unknown functions were excluded (11,529 genes). The distribution of known functions in the entire genome is shown. (b) The recent gene duplicates identified with  $K_s < 0.05$  and PAM < 30 from *A. thaliana* were compared to the GO database (Berardini et al., 2004) and genes with unknown function were excluded (323 genes). The distribution of remaining recent duplicates is shown, where (as described in the text and in comparison with Fig. 1a), there is some deviation from the GO classifications of the genome in general.

A genome-wide analysis of recent gene duplications in the *A. thaliana* genome revealed 446 gene pairs with  $K_s < 0.05$  and a PAM distance of less than 30. These genes formed 189 families containing 187 tandem (separated by less than 20 loci) duplicates. The largest family of tandem duplicated genes is on chromosome 4 and consists of 13 genes containing a protein kinase domain, but of otherwise unknown function. The distribution of the genes per chromosome is shown in Table 1. For 68% (303) of these duplicated genes, the molecular function is unknown, a significant over-representation compared to the corresponding fraction in the whole genome (38%; 10,417 out of 27,503 GO annotations). In addition, proteins involved in proteolysis are over-represented in the duplication dataset (9.6% compared to 1.2% in the whole genome). In examining substitution patterns, 127 pairs had a  $K_a$  and a  $K_s$  of 0, indicating very recent duplication (or gene conversion). An additional eight pairs had  $K_a = 0$  and  $K_s > 0$ ,

while 34 had  $K_s = 0$ ,  $K_a > 0$ . For 122 pairs,  $K_a$  was larger than  $K_s$ . Additional work is needed to elucidate the fate of the duplicated genes, especially for the pairs with  $K_a/K_s > 1$  as well as the pairs with  $K_a/K_s = 0$ . As seen in Figure 2, most recently duplicated genes become pseudogenes and are observed with decreasing frequency at increasing  $K_s$  (this result is qualitatively consistent with the observations of Lynch and Conery, 2000). To confirm that these were actually lineage-specific duplicates, sequences were compared with EST and gene datasets from other Brassica species and the relative branching order of *Arabidopsis* duplicates and Brassica homologs was confirmed phylogenetically. This analysis identified only one out-paralog (a duplication event that was not specific to *Arabidopsis*), with the remainder appearing to be lineage-specific.

The picture of large-scale and lineage-specific gene duplication in *Arabidopsis* leads to a general question of how gene duplication feeds into

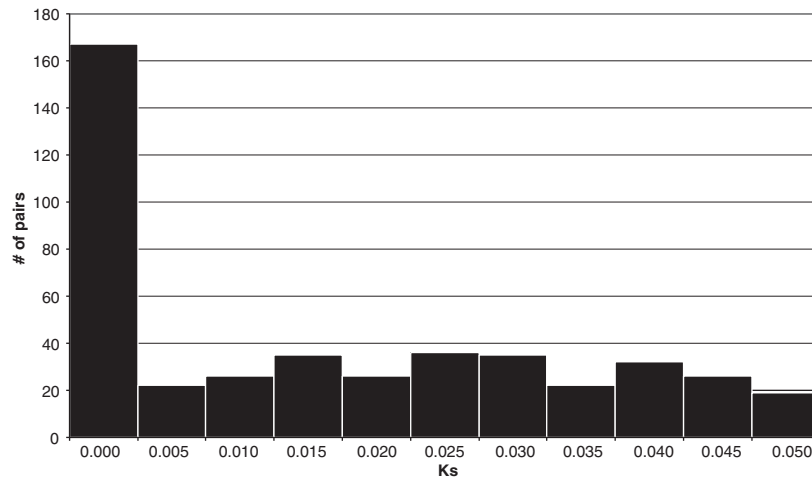


Fig. 2. A histogram shows the  $K_s$  (as a proxy for time) distribution of recent duplicates in *Arabidopsis thaliana*. Most recent duplicates appear to become pseudogenes by  $K_s = 0.05$ .

existing metabolic, signaling, and transcriptional networks in determining which duplicates are retained, as will be discussed in the next section. This process is, of course, different for single gene vs. large-scale duplication events.

### ***Metabolic networks and duplication***

Gene duplication was suggested as a primary source of new enzymatic activities, as one of the first theories regarding the evolution of metabolic pathways, often referred to as retrograde evolution, was proposed by Horowitz ('45). The theory was primarily addressing the problem of how one enzyme could provide a selective advantage to an organism before the whole pathway had been completed. It states that during evolution, pathways assembled backwards compared to the direction of the pathway, in response to substrate depletion, through gene duplication (Horowitz, '65). Consequently, enzymes which are close to each other in the pathways should, according to this scenario, be paralogous.

Patchwork evolution (Jensen, '76; Lazcano and Miller, '96) was later suggested as an alternative to retrograde evolution. This model states that the number of enzymes was initially small, which promoted broad substrate specificities to encompass as many diverse functions as possible. As the number of enzymes grew, through gene duplication, enzyme specificity increased, for example, as a result of single domain insertions via neofunctionalization (Bjorklund et al., 2005). Ultimately either subfunctionalization (increases in specificity) or neofunctionalization (evolving specificity

to bind to new substrates) can be consistent with the general framework of the patchwork evolution model. In addition, Jensen suggested that pathways may evolve *en bloc*, i.e., genes whose gene products catalyze reaction chains are duplicated *en bloc* and evolve towards increased substrate specificities. And indeed, several reaction chains in different pathways are highly similar, including for example reaction chains in the citric acid cycle and in lysine biosynthesis.

Recent studies have shown that there is little support for the retrograde evolution model while Jensen's patchwork evolution model has substantially more support (Alves et al., 2002; Rison et al., 2002; Light and Kraulis, 2004). Interestingly, an analogous argument to retrograde evolution can be successfully applied to the evolution of vertebrate steroid receptors (Thornton, 2001). Gene duplicates may rapidly be turned into pseudogenes unless there is selection for their products. However, new specificities do emerge. An intriguing question is if proteins co-evolve with their new ligands or if new specificities can arise in the absence of ligands. According to Thornton's ligand exploitation process, the terminal ligand in the biosynthetic pathway of steroids in vertebrates is the first one for which a receptor evolved, namely an estrogen receptor. Subsequently, the family of steroid receptors appears to have evolved new ligand specificities in order to accommodate the intermediate products in the estrogen synthesis pathway for an increased specificity in signaling (Thornton, 2001).

Ultimately, the importance of gene duplication in the evolution of the metabolic network of *E. coli*

may be small compared to the impact of horizontal gene transfer (HGT) (Pal et al., 2005). A recent study shows that the metabolic network of *E. coli* evolves by incorporating horizontally transferred genes which are predominantly involved in transportation (Pal et al., 2005), and another study indicates that the horizontally transferred enzymes are, on average, highly connected in the metabolic network (Light et al., 2005). The presumed large effective population size of ancestral *E. coli* might render pseudogenization more likely than subfunctionalization according to the model of Lynch (Force et al., 2001), but the population genetic basis for a preference of lateral gene transfer over gene duplication in organisms of large effective population size remains unclear. Further, the effect of the operon structure of bacteria on the DDC process remains to be modeled.

### ***Protein–protein interaction networks (PPINs) and duplication***

The PPIN of *S. cerevisiae* is susceptible to mutations on proteins with many interactions (hubs) (Albert et al., 2000; Jeong et al., 2001). Since hub proteins appear to be pivotal for the robustness of the PPIN, it is conceivable that the *S. cerevisiae* genome may contain more genetically redundant duplicates of the hubs compared to other proteins. In fact, evolutionary theory has predicted a set of conditions under which selection for duplicates as a buffering effect against future deleterious mutation may act, but this has not been clearly established as an important mechanism for duplicate gene retention. In contrast, gene duplications may cause an imbalance in the concentration of the components of protein–protein complexes which might be deleterious (Veitia, 2002; Papp et al., 2003). A neutral argument that pseudogenization is most probable when there is no selective pressure to retain the gene based upon a lack of binding partners can also be presented.

However, recent studies show that the duplication rate of hub proteins is similar to that of other proteins (Ekman et al., 2006; Prachumwat and Li, 2006). Interestingly, static hubs (Han et al., 2004), many of which are at the cores of highly conserved protein complexes, have few paralogs which originate from relatively recent duplications (Ekman et al., 2006). The cause for the apparent decrease of hub duplicates is unclear, but it is possible that the dosage sensitivity of static, well-

connected (party) hubs has increased in comparison to other proteins through evolution. Additionally, this may reveal a timing issue, where the duplication of hub proteins is related to the frequency of large-scale duplication events as opposed to smaller-scale lineage-specific events, where interacting partners are duplicated simultaneously.

Interestingly, a recent study by Pereira-Leal and Teichmann (2005) indicates that the complete or partial module duplication, i.e., duplication of protein complexes, has occurred on several occasions during the evolution of *S. cerevisiae*. With the exception of these complexes, the large numbers of interactions of hub proteins cannot be explained by interactions with many paralogous proteins (Ekman et al., 2006). In fact, after duplication, interactions are rapidly lost in an asymmetric manner, where one of the genes loses most of the original interactions (Wagner, 2001, 2002). The asymmetry is prominent in hubs, where divergence results in loss of subfunctions in the duplicated gene, whereas duplicates of proteins with few interaction partners are more likely to gain new functions (Zhang et al., 2005).

### ***Transcriptional networks and duplication***

At a transcriptional level, regulatory interactions consist of a transcription factor, which binds to its target gene through a DNA binding site. One single transcription factor may control several target genes and one target gene can be regulated by many different transcription factors.

Duplication has played a key role in the evolution of interactions in the regulatory networks. Both transcription factors and the target genes have been extensively duplicated. Following duplication, interactions may be inherited from the ancestor to the duplicate or new interactions may be gained through divergence. According to a study by Teichmann and Babu (2004), approximately half of the interactions have been gained through sequence divergence after duplication, whereas one-third of the interactions have been inherited from the ancestral gene. Only a minority (~10%) of the interactions between transcription factors and target genes consist of genes that lack homologs. Inherited interactions after duplication of the target gene comprise approximately 20% of the interactions in both *E. coli* and yeast. On the other hand, conserved interactions of transcription factor duplicates are more common in yeast (22%) than in *E. coli* (10%), perhaps because

yeast have more genes that are regulated by multiple transcription factors, or perhaps because the action of positive selection is stronger in *E. coli*. Simultaneous duplications of transcription factor and target gene, which is probably facilitated by having adjacent locations on the chromosome, is also responsible for a minor number of inherited interactions. However, duplication alone is not sufficient to explain the large numbers of target genes for some transcription factors. In fact, transcription factor binding sites can evolve easily from random DNA and the development of better models that combine sequence evolution and the specificity of transcription-factor DNA interactions will expand our understanding of this type of regulatory evolution (Berg et al., 2004).

Given the process by which large-scale gene duplication leads to the evolution of new metabolic and signaling pathways, new protein-protein interactions, and new transcriptional networks, it clearly has the capability to present a foundation for the evolution of biological novelty. This novelty potential can be exploited in different ways in different lineages, dependent upon the time taken to exploit or resolve the redundancy from duplication, potentially driving biodiversity. The next section will explore the impact that such duplications may have had on speciation rates and the evolution of biodiversity on Earth.

### ***WGD and organismal evolution***

Radiations abound throughout the evolutionary history of organisms and involved such crucial transitions as that from single-celled organisms to complex multicellular animals and plants. The points that have long sparked the interest of researchers are the pivotal radiations at the base of the Metazoa, including the Chordata-Vertebrata split.

One of the early hypotheses brought forward by Ohno ('70) was that two rounds of WGD occurred at the cusp of vertebrate evolution (Urochordata—Craniata), leading to the relatively large size and complexity of the vertebrate genome (2R hypothesis). This basic tenet is still hotly debated, and observations made with recently available genome mining tools have resulted in conflicting evidence. The accumulation of complete genomic sequences clearly showed that significant portions of the vertebrate genome consist of duplicated gene loci, thus making de novo origins of genes a much more rare event. Also, the now refuted speculation that because some gene families have four representa-

tives in vertebrates (e.g., *Hox* clusters) but just one in invertebrates, that this would be general seemed to support this hypothesis (Garcia-Fernandez and Holland, '94; Popovici et al., 2001), but has subsequently proven to be an overly simplistic view. Studies comparing the complete human and *Drosophila* genome sequences though revealed that less than 5% of homologous genes follow this rule (Friedman and Hughes, 2003).

Recently, Dehal and Boore (2005) showed by determination of the evolutionary history of all gene families of tunicate (invertebrate), fish, mouse, and human (vertebrate) and by analysis of the relative position of the resulting paralogs in the vertebrate genome (prior to the fish-tetrapod split), that there is evidence of two distinct WGD events early in vertebrate evolution (Urochordata—Craniata). According to current evolutionary theory, the invertebrate tunicate (*Ciona intestinalis*) is a close basal chordate relative of vertebrates (Lit), and thus is better suited for comparison than the more distant *Drosophila* (Dehal et al., 2002). Dehal and Boore again found evidence that the simple notion of quadrupling of certain gene families of single invertebrate origin in vertebrates is not supported, and that early observations to that respect are anecdotal. However, based on the combined evidence of gene families, phylogenetic trees and genomic map position conclusive evidence for the 2R hypothesis is presented, supporting earlier studies (Lundin, '93; Spring, '97; Meyer and Schartl, '99; Wang and Gu, 2000; Taylor et al., 2001; Larhammar et al., 2002; McLysaght et al., 2002; Panopoulou et al., 2003 among others). After the generation of gene clusters containing all genes descended by shared single ancestry, the authors inferred evolutionary relationships of each gene cluster. Then, gene trees were compared to organismal trees, allowing for dating each duplication in relation to organismal lineage divergence. By further examination of the human genomic map position of only those genes derived from a duplication event at the base of vertebrates, a clear image of tetra-paralogs emerged. Due to the fact that some pairs of those sets extend over longer regions than others, the authors speculate that two rounds of auto-tetraploidization happened, as opposed to a single octoploidy. Furthermore, they rather conservatively allude to the effect of WGD on subsequent organismal radiations and rapid and extensive evolutionary change. While the authors do not deny the fact that WGDs drive macroevolutionary change, they

feel unsure as to what extent. Interestingly, they observe a general massive gene loss after duplication, indicating that few genes might have been involved in the increased vertebrate complexity (this rapid loss and neofunctionalization of a few genes is predicted by many of the models presented earlier, see Rastogi and Liberles, 2005). Additionally, they speculate about the potential loss of constraint on certain genes having played some role in the evolution of vertebrate complexity beginning.

Earlier studies conducted by Panopoulou et al. (2003), based on a comparison of gene catalogs from amphioxus (*Branchiostoma floridae*, even closer related to vertebrates than *Ciona*) to over 3,400 single-copy genes orthologs of *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *C. intestinalis* show an increased gene duplication activity after the separation of amphioxus and vertebrate lineages, supporting the 2R hypothesis. Contrarily to most studies, Panopoulou et al. (2003) actually use molecular clock approaches to estimate the duplication time of certain segments. Based on a large number of amphioxus genes, the cephalochordate–vertebrate split was placed at 647–654 MYA. The majority of the human orthologs dated via this method were duplicated between 300 and 680 MYA (mean 488 MY). Clearly, the accuracy of molecular clocks is still debatable, and currently estimated phylum divergences vary around a mean of 800 million years (Wray et al., '96; Nikoh et al., '97; Ayala et al., '98; Valentine et al., '99). However, a crucial evolutionary event for early metazoan history lies within the estimated timeframe of the 2R WGDs.

During the Cambrian period, there was a frenzy of evolutionary innovation, marking a sharp transition in the fossil record with the appearance of numerous metazoan body plans, including the ones of living fauna. Accurate dating methods indicate that the so-called *Cambrian Explosion* happened during a very short period—from ca. 530 to 520 MYA, spanning only 1.7% of the duration of the animal fossil record (Bowring and Erwin, '98). Early metazoans were already found in the Neoproterozoic, and bilateral traces date back to ca. 610 MYA (Hoffman et al., '90; Brasier and McIlroy, '98). Valentine et al. ('99) however emphasize, that based on fossil, phylogenetic, and morphological evidence an estimated metazoan origin at 700 MYA may be possible. By the end of the 10 million years of Cambrian explosion, all but one of the phyla with fossilizable skeletons had appeared. The earliest unmista-

kenly metazoan body fossils were recovered from phosphorite deposits on the Yangtze in China, dating to ca 570 MYA (Xiao et al., '98).

Valentine et al. ('96, '99) point out that minimum *Hox* gene array studies (analysis using a parsimony reconstruction of *Hox* gene presence and absence) suggest that much of the basic gene-regulatory machinery required to set up metazoan body plans was in place significantly before the explosion. This might have been related to WGD events at the base of metazoan development, and would give support to the claim that the “explosion” itself was of limited biological significance, as all important, preparative events in metazoan evolution had occurred earlier.

Additionally, this scenario lends further support to the “punctuated equilibrium” observation made by Gould and Eldredge ('77) and the founder effect model of speciation proposed by Mayr ('63). Taken all above-mentioned evidence into account, it would also support a traceable scenario, where WGDs are followed by radiation and the emergence of novel life forms and strategies.

Another hypothesized WGD, initially based on the observation of a duplication of *Hox* gene clusters in zebrafish, relates to the origin of fishes (Amores et al., '98). Taylor et al. (2003) tested this hypothesis further, which postulates a WGD after the divergence of ray-finned and lobe-finned fishes, but before teleost radiation. In their study, different phylogenetic analyses were used to produce topologies identifying zebrafish duplicates and their orthologs in other fish species. Genes were then mapped by using a radiation hybrid panel and checked for synteny. Their results indicate a large number of fish genes being duplicated before the divergence of zebrafish and pufferfish ancestors. Although based on individual gene studies, leaving open the possibility of multiple independent gene duplications, the author feels that single gene duplication events are not to be expected to produce multiple, multigene blocks of paralogy. Models described in the second section can be expanded to give a probabilistic view of the correlated timing of these events. Based on the results of Taylor et al. (2003), and the fact that the Euteleostei are a species-rich clade, including 22,000 extant taxa, they speculate about a potential causal link of WGD and speciation, without alluding to the supposed timeframe of the speciation event(s).

Donoghue and Purnell (2005) take the fish example to challenge the notion of a connection between WGD and bursts of character acquisition,

on the grounds of undersampling of taxonomically and anatomically intermediate fossils. According to them, the observed pattern of sudden morphological change and evolutionary bursts is due to sampling error. In fact more inclusive sampling of fossil taxa (e.g., 11 intercalate clades between teleosts and their nearest living relative) shows no support for linking gen(om)e duplications and the evolution of vertebrate complexity. This is an interesting, although very Darwinian observation, and the mere presence of additional, now extinct morphologically intermediate taxa does not challenge the possibility of their rapid occurrence, nor the occurrence of new body forms or radiations of those morphologically intermediate fossils. The fact that morphological change does not always develop abruptly also does not interfere with the possibility that gen(om)e duplications may lay at the base of species-rich clades. Fossil evidence, although an additional source of information will not solve this discussion.

The real problem in all arguments supporting or countering the connection between gen(om)e duplication and species radiation is that no studies have considered to correlate rates of duplicated gene evolution and rates of speciation (and any potential lag time between WGD events and phenotypic innovation). Under the scenario of gene duplication, changed (e.g., accelerated) rates of genetic evolution for the “surviving” neo- or subfunctionalized genes should be expected. Lately, several studies demonstrated a connection between speciation net rates, and genetic change (Barraclough and Savolainen, 2001; Webster et al., 2003), indicating that rapid genetic evolution results may result in higher speciation rates.

The continued surveying has in its essence led to three putatively confirmed WGD throughout the history of the Deuterostomia. Although none of the results listed above, if considered alone, provide full support for a connection of WGD coupled with following radiation, when combined with a variety of evidence, support for this connection can be observed throughout the evolutionary history of life on earth (Holland et al., '94; Ruddle et al., '94; Sidow, '96; Stellwag, '99; Depew et al., 2002; Aburomia et al., 2003; Wagner et al., 2003).

## CONCLUSION

The development of increasingly sophisticated integrative models that examine the process of gene duplication, with explicit consideration of

evolutionary processes at the species level coupled to a structurally constrained molecular-level analysis, are improving our understanding of how genes and genomes evolve and how this leads to the evolution of new function. Comparative genomic analysis is enabling these models to be tested in large scale and is increasingly being coupled to in-vitro biochemistry and ancestral sequence reconstruction to understand the process at that level of detail. Integrating this knowledge enables an understanding of systems biology from an evolutionary perspective, which can feed into an understanding of how organisms evolve to utilize novel ecological niches. At the species level, there appears to be a correlation between WGD events and speciation rate, providing a potential mechanistic link to the generation of Eukaryotic biodiversity on Earth.

## ACKNOWLEDGMENTS

We are grateful to Xun Gu for inviting this article and to Arne Elofsson for enabling two of his Ph.D. students to spend time working on this article.

## LITERATURE CITED

- Aburomia R, Khaner O, Sidow A. 2003. Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail. *J Struct Funct Genom* 3:45–52.
- Albert R, Jeong H, Barabasi AL. 2000. Error and attack tolerance of complex networks. *Nature* 406:378–382.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
- Alves R, Chaleil RA, Sternberg MJ. 2002. Evolution of enzymes in metabolism: a network perspective. *J Mol Biol* 320:751–770.
- Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Lagneland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711–1714.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
- Arvestad L, Berglund AC, Lagergren J, Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19:i7–i15.
- Arvestad L, Berglund AC, Lagergren J, Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB* 2004:326–335.
- Ayala FJ, Rzhetsky A, Ayala FJ. 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proc Natl Acad Sci, USA* 95:606–611.
- Barraclough TG, Savolainen V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55:677–683.

- Bastolla U, Porto M, Roman HE, Vendruscolo M. 2003. Statistical properties of neutral evolution. *J Mol Evol* 57: S103–S119.
- Berardini TZ, Mundodi S, Reiser R, Huala E, Garcia-Hernandez M, Zhang P, Mueller LM, Yoon J, Doyle A, Lander G, Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY. 2004. Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol* 135:1–11.
- Berg J, Willmann S, Lassig M. 2004. Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.
- Berglund AC, Wallner B, Elofsson A, Liberles DA. 2005. Tertiary windowing to detect positive diversifying selection. *J Mol Evol* 60:499–504.
- Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA. 2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*, in press.
- Bhushan S, Stahl A, Nilsson S, Lefebvre B, Seki M, Roth C, McWilliam D, Wright SJ, Liberles DA, Shinozaki K, Bruce BD, Boutry M, Glaser E. 2005. Catalysis, subcellular localization, expression and evolution of the targeting peptides degrading protease, AtPreP2. *Plant Cell Physiol* 46:985–996.
- Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A. 2005. Domain rearrangement in protein evolution. *J Mol Biol* 353:911–923.
- Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.
- Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.
- Bowring SA, Erwin DH. 1998. A new look at evolutionary rates in deep time: uniting paleontology and high-precision geochronology. *GSA Today* 8:1–8.
- Brasier MD, McLroy D. 1998. *Neonereites uniserialis* from c. 600 Ma year old rocks in western Scotland and the emergence of animals. *J Geol Soc Lond* 155:5–12.
- Braun FN, Liberles DA. 2003. Retention of enzyme gene duplicates by subfunctionalization. *Int J Biol Macromol* 33: 19–22.
- Braun FN, Liberles DA. 2004. Repeat-modulated population genetics effects in fungal proteins. *J Mol Evol* 59: 97–102.
- Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol* 4:10.
- Clauss MJ, Mitchell-Olds T. 2004. Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. *Genetics* 166:1419–1436.
- Cotton JA, Page RDM. 2005. Rates and patterns of gene duplication and loss in the human genome. *Proc R Soc Lond B Biol Sci* 272:277–283.
- Csuros M, Miklos I. 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *RECOMB 2006*, 206–220.
- Deeds EJ, Dokholyan NV, Shakhnovich EI. 2003. Protein evolution within a structural space. *Biophys J* 86: 2962–2972.
- Dehal P, Boore J. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLOS Biol* 3:e314.
- Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KE, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Dettler C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 13:2157–2167.
- Depew MJ, Lufkin T, Rubenstein JL. 2002. Specification of jaw subdivisions by *Dlx* genes. *Science* 298:381–385.
- Donoghue PCJ, Purnell MA. 2005. Genome duplication, extinction and vertebrate evolution. *Trends Ecol Evol* 20: 312–319.
- Dutheil J, Pupko T, Jean-Marie A, Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* 22:1919–1928.
- Ekman D, Light S, Bjorklund AK, Elofsson A. 2006. What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*. *Genome Biology* 7:RY5.
- Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.
- 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.
- Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nat Genet* 37:573–577.
- Friedman R, Hughes AL. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol* 20:154–161.
- Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Kohler N. 1999. Discovery of tetraploidy in a mammal. *Nature* 401:341.
- Garcia-Fernandez J, Holland PWH. 1994. Archetypal organization of the amphioxus *HOX* gene cluster. *Nature* 370: 563–566.
- Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Monte Carlo Markov chain methods in practice. London: Chapman & Hall.
- Gonnet GH, Hallett MT, Korostensky C, Bernardin L. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* 16:101–103.
- Goodman M, Cselusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1997. Fitting the gene lineage into its species tree lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool* 28:132–168.
- Gould SJ, Eldredge N. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* 3: 115–151.
- Gu X, Zhang H. 2004. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol* 21:1401–1408.

- Gu X, Wang YF, Gu JY. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet* 31: 205–209.
- Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* 15: 1153–1160.
- Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430:88–93.
- He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1174.
- Hoffman HJ, Narbonne GM, Aitken JD. 1990. Ediacaran remains from intertillite beds in northwestern Canada. *Geology* 18:1199–1202.
- Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* S:125–133.
- Horowitz NH. 1945. On the evolution of biochemical syntheses. *Proc Natl Acad Sci, USA* 31:153–157.
- Horowitz NH. 1965. Evolving genes and proteins. In: Bryson V, Vogel HJ, editors. *The evolution biochemical synthesis—retrospect and prospect*. New York: Academic Press. p 15–23.
- Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 256:119–124.
- Jensen RA. 1976. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425.
- Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411:41–42.
- Kendall DG. 1948. On the generalized “birth-and-death” process. *Ann Math Stat* 19:1–15.
- Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T. 2001. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell* 13:681–693.
- Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol* 17:1483–1498.
- Larhammar D, Lundin LG, Hallbook F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res* 12:1910–1920.
- Lazcano A, Miller SL. 1996. The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell* 85:793–798.
- Liberles DA. 2001. Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol Biol Evol* 18:2040–2047.
- Liberles DA. 2005. Datasets for evolutionary comparative genomics. *Genome Biol* 6:117.
- Light S, Kraulis P. 2004. Network analysis of metabolic enzyme evolution in *Escherichia coli*. *BMC Bioinform* 5:15.
- Light S, Kraulis P, Elofsson A. 2005. Preferential attachment in evolution of metabolic networks. *BMC Genom* 6:159.
- Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogous evolution. *Trends Genet* 21:60–65.
- Lundin LG. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* 16:1–19.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
- Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473.
- Lynch M, O’Hely M, Walsh B, Force A. 2001. The probability of preservation of newly arisen gene duplicate. *Genetics* 159:1789–1804.
- 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.
- Mayr E. 1963. *Animal species and evolution*. Cambridge, MA: Harvard University Press.
- McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet* 31: 200–204.
- Meyer A, Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol* 11: 699–704.
- Nikoh N, Iwabe N, Kuma K-I, Ohno M, Sugiyama T, Watanabe Y, Yasui K, Zhang S-C, Hori K, Shimura Y, Miyata T. 1997. An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. *J Mol Evol* 45:97–106.
- Novozhilov AS, Karev GP, Koonin EV. 2005. Mathematical modeling of evolution of horizontally transferred genes. *Mol Biol Evol* 22:1721–1732.
- Ohno, S. 1970. *Evolution by gene duplication*. New York: Springer-Verlag.
- Page RD, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol* 7:231–240.
- Pal C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37:1372–1375.
- Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res* 13:1056–1066.
- Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424: 194–197.
- Pereira-Leal J, Teichmann SA. 2005. Novel specificities emerge by stepwise duplication of functional modules. *Genome Res* 15:552–559.
- Popovici C, Leveugle M, Birnbaum D, Coulier F. 2001. Homeobox gene clusters and the human paralogy map. *FEBS Lett* 491:237–242.
- Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A, Brownlie A, Donovan A, Egan ES, Force A, Gong Z, Goutel C, Fritz A, Kelsh R, Knapik E, Liao E, Paw B, Ransom D, Singer A, Thomson M, Abduljabbar TS, Yelick P, Beier D, Joly JS, Larhammar D, Rosa F, Westerfield M, Zon LI, Johnson SL, Talbot WS. 1998. Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* 18:345–349.
- Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol* 23:30–39.
- Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol* 43:304–311.

- Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated gene as a transition state to neofunctionalization. *BMC Evol Biol* 5:28.
- Reed WJ, Hughes BD. 2004. A model explaining the size distribution of gene and protein families. *Math Biosci* 189: 97–102.
- Rison SC, Teichmann SA, Thornton JM. 2002. Homology, pathway distance, and chromosomal localisation of the small molecule metabolism enzymes in *E. coli*. *J Mol Biol* 318: 911–932.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Roth C, Liberles DA. 2006. A systematic search for positive selection in Embryophytes (higher plants). *BMC Plant Biology* 6:12.
- Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA. 2005. The adaptive evolution database (TAED): a phylogeny based tool for comparative genomics. *Nucleic Acids Res* 33:D495–D497.
- Ruddle FH, Bentley KL, Murtha MT, Risch N. 1994. Gene loss and gain in the evolution of the vertebrates. *Development* S:155–161.
- Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. 2005. Protein structure and evolutionary history determine sequence space topology. *Genome Res* 15:385–392.
- Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev* 6:715–722.
- Soltis PS. 2005. Ancient and recent polyploidy in angiosperms. *New Phytol* 166:5–8.
- Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol Evol* 14:348–352.
- Spring J. 1997. Vertebrate evolution by interspecific hybridization—are we polyploid? *FEBS Lett* 400:2–8.
- Stellwag EJ. 1999. Hox gene duplication in fish. *Sem Cell Dev Biol* 10:531–540.
- Stern A, Pupko T. 2006. An evolutionary space-time model with varying among-site dependencies. *Mol Biol Evol* 23: 392–400.
- Taverna DM, Goldstein RA. 2002a. Why are proteins so robust to site mutations? *J Mol Biol* 315:479–484.
- Taverna DM, Goldstein RA. 2002b. Why proteins are marginally stable? *Proteins* 46:105–109.
- Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B Biol Sci* 356:1661–1679.
- Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome Duplication, a trait shared by 22,000 species of ray finned fish. *Genome Res* 13:382–390.
- Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. *Nat Genet* 36:492–496.
- Thornton JW. 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proc Natl Acad Sci, USA* 98:5671–5676.
- Valentine JW, Erwin DH, Jablonski D. 1996. Developmental evolution of metazoan bodyplans: the fossil evidence. *Dev Biol* 173:373–381.
- Valentine JW, Jablonski D, Erwin D. 1999. Fossils, molecules and embryos: new perspectives on the Cambrian. *Development* 126:851–859.
- Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* 24:175–184.
- Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114–2117.
- Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18:1283–1292.
- Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol* 19:1760–1768.
- Wagner GP, Amemiya C, Ruddle F. 2003. Hox cluster duplications and the opportunity for evolutionary novelties. *Proc Natl Acad Sci, USA* 100:14603–14606.
- Wang Y, Gu X. 2000. Evolutionary patterns of gene families generated in the early stage of vertebrates. *J Mol Evol* 51: 88–96.
- Webster AJ, Payne RJH, Pagel M. 2003. Molecular phylogenies link rates of evolution on speciation. *Science* 301:478.
- Williams PD, Pollock DD, Goldstein RA. 2001. Evolution of functionality in lattice proteins. *J Mol Graph Model* 19: 150–156.
- Wray GA, Levinton JS, Shapiro LH. 1996. Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* 274:568–573.
- Wroe R, Bornberg-Bauer E, Chan HS. 2005. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophys J* 88:118–131.
- Xiao S, Zhang Y, Knoll AH. 1998. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. *Nature* 391:553–558.
- Xu YO, Hall RW, Goldstein RA and Pollock DD. 2005. Divergence, recombination and retention of functionality during protein evolution. *Hum Genomics* 2:158–167.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13:555–556.
- Zhang L. 1997. On a Mirkin–Muchnik–Smith conjecture for comparing molecular phylogenies. *J Comp Biol* 4: 177–187.
- Zhang Z, Luo ZW, Kishino H, Kearsey MJ. 2005. Divergence pattern of duplicate genes in protein–protein interactions follows the power law. *Mol Biol Evol* 22:501–505.