Gavin C. Conant
Candidate

Department of Biology
Department

This dissertation is approved, and it is acceptable in quality and form for publication on microfilm:

Approved by the Dissertation Committee:

, Chairperson

Accepted:

Dean, Graduate School

Date

FUNCTIONAL DIVERGENCE AT THE MOLECULAR LEVEL:

ROBUSTNESS, ASYMMETRY, AND CONVERGENCE

BY

GAVIN C. CONANT

B.S., Biology, The University of New Mexico, 1998

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biology

The University of New Mexico
Albuquerque, New Mexico

May 2004

Acknowledgements

I would like to particularly thank Andreas Wagner, my co-author in chapters 2-4 and the appendix and a constant source of advice and counsel. I would also like to thank Michael Gilchrist, Michael Fuller, and Annette Evangelisti for many helpful discussions as well as the members of my committee, Donald Natvig, Robert Miller and Laura Salter for their patience and assistance. Financial support was provided by the Department of Energy's Computational Sciences Graduate Fellowship program, administered by the Krell Institute. Further financial and computational support was provided by the University of New Mexico's High Performance Computing Center. Finally, I would like to thank my parents, Ettajane and Henry Conant, and my sister, Eleanore Conant, for their support.

Dedication

To M. N. Duty

FUNCTIONAL DIVERGENCE AT THE MOLECULAR LEVEL:

ROBUSTNESS, ASYMMETRY, AND CONVERGENCE

BY

GAVIN C. CONANT

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biology

The University of New Mexico
Albuquerque, New Mexico

May 2004

Functional divergence at the molecular level: Robustness, asymmetry, and convergence

Gavin C. Conant

B. S., Biology, The University of New Mexico, 1998
Ph. D., Biology, The University of New Mexico, 2004

Abstract

The origin of novel structures in evolution has been of interest since Darwin proposed the theory of natural selection. One important source of molecular novelty is the duplication and diversification of genetic material. Here I study the association of duplication and novelty and find all possible combinations of the two: duplication and diversification to a new function, duplication without diversification, and the appearance of structures having novel functions without duplication. First, in chapter 2, I report that duplicated genes in four eukaryotic genomes show asymmetric amino acid sequence divergence in roughly 30% of cases, an estimate substantially higher than a previous whole-genome study suggested. This result is significant because one potential cause of asymmetric divergence is the appearance of novel features in one member of the duplicate pair. In chapter 3, I consider whether duplicated genes provide mutational robustness in the nematode worm *Caenorhabditis elegans.* I find that duplicated genes do indeed show less severe phenotypic effects when their expression is prevented and that this effect is stronger for duplicates with more similar amino acid sequences or expression patterns. Finally, in chapter 4, I demonstrate that several recently discovered regulatory motifs in the transcriptional regulatory networks of the yeast *Saccharomyces cerevisiae* and the bacterium *Escherichia coli* are abundant due to convergent evolution and not ancestral duplications. Collectively, these results suggest that while duplication is one route to novel function (chapter 2), other routes are possible (as in the case of circuit motifs) and that duplication may contribute to evolution in ways other than the generation of novelty, such as the buffering against loss of gene function.

Table of Contents

Table of Figures

Table of Tables

Chapter 1: Introduction

Evolutionary novelty, gene duplication, and convergent evolution

Despite the theological difficulties raised by *The Origin of Species* (Darwin 1859), most scientists even in Darwin's time found persuasive his evidence that the modern diversity of life was the result of descent with modification from one or a few original forms (Bronowski 1973; Futuyma 1998). However, Darwin's theory of how organic evolution had occurred, namely through natural selection, gained support more slowly. At least one reason was the problem of the origins of novel features in evolution (Ruse 2003). Darwin's argument for the gradual accumulation of small changes leading to structures with novel function was felt to be unpersuasive, especially because a genetic theory of inheritance was not in place. The difficulty was that while a fully-formed complex structure such as a wing conferred obvious benefits to an organism possessing it, it was hard to see how an incipient and incomplete form of that structure could confer any benefit. At a morphological level, one method of resolving this difficulty has been to recognize that many complex biological structures were not constructed *de novo* by natural selection but were co-opted from existing structures with other functions. A classic example of this co-option is the panda's thumb, described by Gould (1980). Pandas' have evolved the ability to grasp bamboo stalks (their primary food) with a thumb-like structure in their paws. However, the "thumb" in question is not the anatomical homolog of the human digit. Rather, it is constructed from a modified radial sesamoid bone of the wrist. In constructing the panda's thumb, natural selection was able to use not just existing anatomy, but existing enzymes, developmental pathways, and tissues, making this "novel" structure out of mostly existing parts. Another more recently discovered example is the discovery of feathers on distinctly non-avian dinosaurs (Ji *et*

al. 2001; Xu *et al*. 2001). That dinosaurs which were not the direct ancestors of birds nonetheless possessed feathers indicates that feathers originally evolved for some other function (Bakker suggests warmth [1986]) and only later were co-opted for flight.

Molecular novelty and duplication

The discovery of the structure of DNA (Watson & Crick 1953) naturally led to questions regarding the origins of novelty at the fundamental levels of genes and proteins. Such novelty is perhaps most usefully thought of in terms of function: the ability of an organism to do something its ancestors could not. Such novelty can clearly occur in the context of a new structure, such as an enzyme able to catalyze a new reaction. However, as in the case of the panda's thumb, it is very likely that these "new" structures are actually modified forms of older ones. In the following pages, I will consider some of the possible pathways leading from existing structures to novel functions.

Changes in the genetic material are the raw material from which novelty is built. Biochemically, these changes include single base pair mutations, insertion or deletion of stretches of DNA, and recombination between sequences. These events can produce novelty in several ways. For instance, any one of these events could change a transcription factor binding site and as a result alter the timing or location of a gene's expression. This type of change may be responsible for many of the high-level morphological changes seen during the evolution of the vertebrate line. For example, the vertebrate genes *Hoxa-11* and *Hoxa-13* show non-overlapping expression domains in tetrapod limbs (Fromental-Ramain *et al*. 1996; Haack & Gruss 1993; Nelson *et al*. 1996). However, at least in derived teleost lines (zebrafish), they show overlapping expression in the fins (Sordino *et al*. 1995). One may hypothesize that the partitioning of where these

two transcription factors were expressed allowed each to target different genes in its own region of expression. Such targeting, in turn, may have allowed the evolution of limbs such as the human arm, with distinct arm, wrist and hand bones, from the less differentiated ancestral fin (Chiu *et al*. 2000) .

Here, I will consider another important source of genetic novelty: the duplication of genetic material. An important difference between duplication and the process of evolution through expression change which I just discussed is that duplication avoids the problem of possibly losing an important existing function in order to evolve a new one. In the example above, molecular changes precluded retaining the ancestral fin during the evolution of limbs. Obviously, such a mode of evolution is only possible in cases where loss of the ancestral state is permissible. In the case of gene duplication, on the other hand, it is possible to retain the ancestral function (in one copy of the gene, for instance) while at the same time allowing a second gene copy to be modified by selection.

Haldane (1933) made early suggestions of the potential importance of gene duplication, a topic later considered in depth by Ohno in his book *Evolution by Gene Duplication* (1970). The duplication of single genes was not the only type of duplication discussed by Ohno. In fact, he believed that whole genome duplication (polyploidization) was more likely to give rise to novel functions, arguing that a single gene duplication would often be detrimental because of the resulting doubling of that gene's expression (Ohno 1970). Duplication in fact covers a multitude of possibilities from single gene duplication through regional duplication to polyploidization. One can even consider the possibility of the duplication of entire genetic pathways. Consider a pathway consisting of two genes. If the first gene is duplicated and that duplication

becomes fixed in the population due to genetic drift (see below), it is possible (although unlikely) that the second gene could later be independently duplicated. On its own, this first duplication might not confer any benefit to the organism. However, the advent of the second duplicate pair, might, by completing the duplication of the pathway, confer a selective benefit that would drive that duplication to fixation. I mention this possibility not because it is an important process in evolution (indeed, chapter 4 below will suggest its rarity) but because it constitutes a hypothesis that must be considered when inferring the origin of a recurring pathway.

Although duplication (from single genes up to whole genomes) is now seen as a key ingredient in the generation of novelty at the molecular level (for review see Holland 1999; Lundin 1999), duplication need not result in genes acquiring novel functions. In order to understand how molecular changes (the fixing of mutations in evolutionary time) affect the fate of duplicate genes, it is first necessary to discuss briefly some key results of the theory of molecular evolution (Kimura 1983).

Molecular evolution, the neutral theory, and selection

The study of molecular evolution was revolutionized by the availability of protein and DNA sequence data (Harris *et al*. 1956). One fact these data made clear was that the common ancestry of biological species has left those species with genes of related structure and function, genes which are referred to as being *homologous*. Because these homologous genes are not identical between species, scientists sought to explain how the differences had appeared. Before the structure of the hereditary material was known, Fisher and Wright had derived models describing how alleles of a gene which did not effect an organism's fitness would change in frequency in a population over time (Fisher

1930; Wright 1931). Such alleles are referred to as *neutral* alleles and the evolutionary force that changes them over time is known as *genetic drift*. It was soon realized that many changes in the sequence of homologous genes could be described by this process of drift (beneficial changes in sequence may of course increase in frequency due to selection: there is still some debate as to the relative importance of drift versus selection in sequence evolution, Li 1997).

Drift occurs because gene frequencies in a population fluctuate over time due to the population's finite size. If one waits long enough these fluctuations will either result in the loss of a given allele from the population or in that allele replacing all other alleles in the population, at which point the allele is said to have been *fixed*. Motoo Kimura rigorously described the mathematics of base-pair substitution through drift (Kimura 1983). His neutral theory of molecular evolution made several key points. Among them: a) the rate of fixation of neutral alleles in a population is independent of population size, and b) mutations whose deleterious effects are slight relative to the population size behave effectively as neutral mutations.

Gene duplication and molecular evolution

Just as with an allele of a gene, gene duplicates can appear and disappear from populations. It is therefore important to understand the forces that preserve duplicate genes and those that tend to remove them. Probably the major force eliminating duplicates is genetic drift. It is possible to introduce a mutation into a duplicated gene that makes that copy of the gene non-functional. Because the other copy of the gene maintains its function, this new *null* mutation does not change the fitness of the organism, at least to a first approximation (Li 1980; Nei & Roychoudhury 1973). The null mutation

may thus drift to fixation, eliminating the duplication. It is generally believed that null mutations are more common than beneficial mutations, suggesting that most duplicate genes will be *silenced* before they evolve novel functions (Li 1980; Nei & Roychoudhury 1973).

Analyses of full genomes have shown that duplicate genes are very common in eukaryotes (Conant & Wagner 2002—see Appendix; Lynch & Conery 2000; Rubin *et al*. 2000) and have been preserved over long periods (Bisbee *et al*. 1977; Ferris & Whitt 1977; Hughes & Hughes 1993). Because the beneficial substitutions that preserve duplicates under the above model of diversification or degeneration are rare, the model suggests that duplicate genes should themselves be rather rare. The fact that, on the contrary, duplicate genes are observed to be very abundant suggests that this model is incomplete. The question then becomes one of identifying other forces which preserve duplicated genes.

There are in fact at least three potential forces that can prevent the degeneration of duplicated genes. The first is a requirement to maintain high dosages of a gene: one example comes from the ribosomal DNA genes, where multiple copies are maintained in order to allow sufficient ribosome production (Li 1997). The second possibility is that duplicate genes may be maintained because their redundancy protects the organism against detrimental mutations in either copy of the gene. Although intuitively appealing, this hypothesis of redundancy for mutational robustness is somewhat problematic because the benefit conferred on an organism by such robustness is slight (Cooke *et al*. 1997; Nowak *et al*. 1997; Wagner 1999; Wagner 2000c). Nonetheless, selection can maintain such redundancy given large population sizes (Wagner 1999; Wagner 2000c).

The third and perhaps most important possibility for the preservation of duplicate genes is functional divergence.

This final possibility will depend heavily on the particulars of the function of the gene undergoing divergence and the population in which this divergence occurs. Distinguishing duplicates preserved as a result of diversification from duplicates preserved for the other two reasons can be challenging. One approach to making these distinctions is to compare the rate of synonymous and non-synonymous changes in the sequences in question. Synonymous substitutions are base changes that yield another codon for the same amino acid—such changes do not alter the protein being coded for. Non-synonymous changes, on the other hand, do change the protein being coded for. The comparison of these two rates is a measure of the selective forces at work in the sequences. If drift is the dominant force in the divergence of the two sequences, then K_a, the number of non-synonymous substitutions per non-synonymous site, cannot exceed K_s, the number of synonymous substitutions per synonymous site (Li 1997). If, however, natural selection is the dominant force, then K_a can exceed K_s because the selectively-driven nonsynonymous substitutions occur more rapidly than do synonymous substitutions resulting from drift (Li 1997). Despite the prevalence of duplicated genes and the plausibility of the hypothesis that at least some of these duplicates have diverged in function through natural selection, whole-genome studies have identified only a handful of gene pairs where $K_a > K_s$ (Kondrashov *et al*. 2002; Lynch & Conery 2000).

The number of detected cases of directional selection in duplicates likely under-estimates the prevalence of functional divergence for several reasons. The "subfunctionalization" model of Force and coauthors (1999) describes mechanism by

which duplicates can diverge in function without directional selection. This model recognizes the importance of the fact that genes often have multiple biological roles. Force and collaborators argued that some duplicate genes may diverge by specializing in a subset of the ancestral functions. Importantly, this type of divergence occurs *neutrally* through partial losses of function in each duplicate due to the accumulation of degenerative substitutions. Lynch and Force (2000) have shown that this model generates duplicate preservation probabilities that are within the range seen in genomes studied. A second reason for the rarity of detected directional selection is the nature of the estimates used to infer it. Because K_a and K_s are averages taken over an entire gene, the signal from directional selection in a small region of the gene may be swamped by genetic drift in the rest of the sequence and selection hence undetectable using the K_a/K_s ratio.

Molecular Evolution: New perspectives from novel experimental techniques

In attempting to understand the relationship of gene duplication and novelty, we have the advantage of being able to draw upon many sources and types of data, most of them quite new. They include full genome sequences (for example Goffeau *et al*. 1996; Wood *et al*. 2002), DNA microarray expression studies (Gasch *et al*. 2000; Spellman *et al*. 1998), whole-genome gene knock-out experiments (Giaever *et al*. 2002; Steinmetz *et al*. 2002; Winzeler *et al*. 1999) and protein-interaction compendia (such as compiled by Ito *et al*. 2001; Uetz *et al*. 2000). Such datasets can be used to assess the degree to which protein-protein interactions constrain the evolution of gene duplicates (Fraser *et al*. 2002; Hahn *et al*. 2004), identify duplicate pairs where expression or interaction is

evolving asymmetrically (Wagner 2002), or assess the redundancy of duplicate genes (Gu *et al*. 2003; Wagner 2000b).

Of course each of these experimental protocols has its own limitations. Full genomic DNA sequences are highly accurate: the genomics company Celera estimates that the sequence error rate (as opposed to assembly errors) in their human genome sequence is less than 0.1% (Venter *et al*. 2001), but it is well-known that raw sequence difference is an imperfect measure of functional divergence. For instance, the proteolytic enzyme trypsin, which cleaves proteins after lysine or arginine residues, can be converted to the substrate specificity of chymotrypsin (which cleaves after large hydrophobic residues) by a single amino acid change (Hedstrom *et al*. 1994). On the other hand, many of the active sites of prokaryotic DNA polymerase I can be mutated without detectable effect on the activity of the enzyme (Patel & Loeb 2000). Thus, one must be careful not to overstate the importance of raw sequence differences in determining functional differences. mRNA expression assays from microarrays, on the other hand, tend to produce data points with large error bounds even under ideal conditions (Brown *et al*. 2001; Schuchhardt *et al*. 2000). Similarly, the errors associated with some protein-protein interaction data are amply demonstrated by the fact that two analyses using similar experimental protocols yielded protein interaction networks that overlapped by only about 20% (Ito *et al*. 2001).

Significant as these limitations are, these types of data offer even greater advantages in studying duplication and divergence. The first advantage is the ability to consider duplication on the level of the whole genome, which prevents any possible biases in the selection of gene families for analysis. Secondly, expression and protein

interaction data are indicators of functional divergence which are comparatively independent of sequence metrics (see figure 10 and Wagner 2001) and are appropriate for trying to determine the prevalence of functional divergence. And finally, the ability to individually eliminate the expression of each gene in the genome is a powerful tool for studying functional redundancy in duplicate genes (see chapter 3).

Overview

Incorporating the above types of data and recalling the various possible fates of duplicated genetic material, I have studied the relationship between duplication and the appearance of functional novelty. My three studies provide evidence for all possible combinations of duplication and divergence. First, in chapter 2, I will focus on detecting asymmetric divergence of amino acid sequence in duplicate genes. Such asymmetries can arise for a number of reasons, including directional selection and the subfunctionalization process proposed by Force and coauthors (1999). As a result, asymmetries are at least a potential indicator of functional divergence. In that chapter, I thus describe evidence for the divergence of duplicate genes: in other words duplication with diversification. In the next chapter (3), I examine the degree of functional overlap remaining in the duplicates of *Caenorhabditis elegans*. Functional overlap in duplicate genes is an important alternative hypothesis when studying functional divergence and constitutes an example of duplication without diversification. This functional overlap is also interesting in and of itself, as it helps researchers understand why organisms show robustness to changes both in their environments and in their own genetic material. Finally, in chapter 4, I present a study of a somewhat different sort where I consider the question of duplication in higher-level genomic structures, namely motifs in the transcriptional regulatory networks of the

yeast *Saccharomyces cerevisiae* and the bacterium *Escherichia coli*. These circuit motifs are very common in the two networks, raising the question of whether they too evolved through duplication or whether they rather evolved convergently due to desirable properties. I show that these circuits do not appear to have arisen via duplication. They thus provide a case of diversification without duplication. Taken collectively, the three problems studied here demonstrate the contingent nature of (molecular) evolution and the often *ad hoc* way in which existing structures are adapted to new needs.

Chapter 2: Asymmetric Sequence Divergence of Duplicated Genes

This chapter has previously appeared in substantially the same form as: Conant, G. C. and Wagner, A. (2003) "Asymmetric sequence divergence of duplicate genes, *Genome Research*, **13(9)**: 2052-2058.

Abstract

Much like humans, gene duplicates may be created equal, but they do not stay that way for long. We here show for four completely sequenced genomes that 20-30% of duplicate gene pairs show asymmetric evolution in the amino acid sequence of their protein products. That is, one of the duplicates evolves much faster than the other. The greater this asymmetry, the greater the ratio K_a/K_s of amino acid substitutions (K_a) to silent substitutions (K_s) in a gene pair. This indicates that most asymmetric divergence may be caused by relaxed selective constraints on one of the duplicates. However, we also find some candidate duplicates where positive (directional) selection of beneficial mutations ($K_a/K_s>1$) may play a role in asymmetric divergence. Our analysis rests on a codon-based model of molecular evolution that allows a test for asymmetric divergence in K_a. The method is also more sensitive in detecting positive selection ($K_a/K_s >1$) than models relying only on pairwise gene comparisons.

Introduction

Much work on the evolution of gene duplication since Ohno (1970) has focused on how gene duplicates diverge both in sequence and function. Although most substitutions in a duplicate are selectively neutral immediately after duplication (Li 1980; Nei & Roychoudhury 1973), this period of neutrality may be ended by a variety of events: nucleotide substitutions that affect protein expression, localization, or dimerization (Force *et al*. 1999; Gibson & Spring 1998) can lead to increasing functional and sequence divergence of gene duplicates and thus to increased selective constraints on both genes. Functional divergence often occurs rapidly, although this is not always the case. For

instance Langkjær *et. al.* have presented evidence suggesting that genes dating from an ancient whole-genome duplication in the yeast *S. cerevisiae* may not have diversified until well after the duplication event (Langkjær *et al*. 2003).

Mounting evidence indicates that gene duplicates can assume unequal roles in divergence. A study by one of us suggests that gene function, as indicated by protein interactions and gene expression patterns, diverges asymmetrically for many gene duplicates in the yeast *Saccharomyces cerevisiae* (Wagner 2002). Other pertinent evidence comes from sequence divergence. Some of this evidence is based on detailed studies of individual genes. For example, Li and Tsoi found that mammalian lactate dehydrogenase C evolved more rapidly than lactate dehydrogenase A (Li & Tsoi 2002). A large-scale study by Kondrashov and collaborators (Kondrashov *et al*. 2002) analyzed 39 genomes from eubacteria, archaea and eukaryotes and found a small number of cases of asymmetric divergence among 101 analyzed duplicate gene pairs. In contrast to this study, where the incidence of asymmetric divergence was less than 5 percent, Van de Peer and collaborators (Van de Peer *et al*. 2001) found that fully half of 26 duplicate gene pairs in zebrafish showed evidence of asymmetric divergence. Using a more sensitive amino-acid based method to detect asymmetry, Dermitzakis and Clark found that roughly 50% of 12 mammalian transcription factor paralogs showed evidence of asymmetric evolution (Dermitzakis & Clark 2001). The functional significance of such asymmetric divergence is still unclear, although some existing evolutionary models might contribute to an explanation. For example, it has been argued that some form of evolutionary asymmetry is required for functional diversification of duplicates (Krakauer & Nowak 1999), and that asymmetric functional divergence might reflect selection for mutational

robustness (Wagner 2002). However, before one can seriously pursue evolutionary models explaining asymmetry, it is necessary to establish its incidence, because existing work has not yielded a final picture.

Previous work on asymmetric sequence divergence relies on relative rate tests between two duplicates and an outgroup gene, using either nucleotide or amino acid substitutions (Kondrashov *et al*. 2002; Li & Tsoi 2002; Van de Peer *et al*. 2001). Nucleotide-based tests cannot distinguish between silent substitutions and amino acid replacement substitutions. The presence of (often neutral) silent substitutions may obscure any signal of asymmetry, which mostly derives from replacement substitutions. Amino acid-based models, on the other hand, have problems with correctly determining outgroup genes, which is necessary to calibrate divergence estimates. Specifically, if two duplicates have diverged asymmetrically, one of the duplicates may have become more divergent than a true outgroup gene. For this reason, amino acid based methods also tend to underestimate the number of gene pairs with asymmetric divergence. These shortcomings prompted us to use a codon-based model of evolution that distinguishes between silent substitutions and amino acid-changing substitutions when testing for asymmetric protein sequence divergence.

Codon-based models of sequence evolution can address questions in both phylogenetics and molecular evolution (for discussion, see Lewis 2001; Liò & Goldman 1998). Such models estimate both synonymous divergence (K_s) and nonsynonymous divergence (K_a) between genes. For the purpose of the present study, we use the model of Muse and Gaut (1994) (A very similar model is described by Goldman & Yang 1994). This model allows each branch of a phylogenetic tree to have its own value of K_s and K_a.

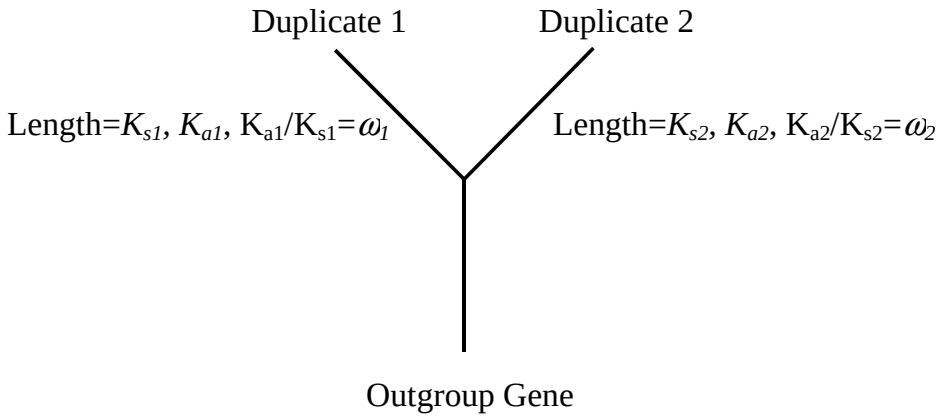


Figure 1: Schematic representation of our model tree. Two duplicates are presumed to have diverged from an outgroup gene. Each of the three branches of this tree is allowed to have its own K_s and K_a values.

To study asymmetric divergence, we apply the model to gene duplicates from the fully sequenced genomes of the yeasts *Saccharomyces cerevisiae* (Goffeau *et al*. 1996) and *Schizosaccharomyces pombe* (Wood *et al*. 2002), the nematode worm *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), and the fruit fly *D. melanogaster* (Adams *et al*. 2000).

Methods

Model of sequence evolution

Following Muse and Gaut (1994) we have applied a codon model to three-taxa trees containing two duplicate genes and an outgroup gene. The model allows both the length of the tree branches (t) and K_a/K_s to differ on each of the three branches (see figure 1). This allows for the possibility that duplicate genes evolve independently of each other.

We tested for statistically significant differences in the rate of evolution between gene duplicates by constraining the duplicates' rate of amino acid divergence to be equal ($K_{a1}=K_{a2}$) (see figure 1). By comparing the likelihood (Felsenstein 1981) of observing

the data under the constraint of symmetry to the likelihood of the unconstrained model, we could evaluate whether the degree of asymmetric divergence was statistically significant (the random nature of molecular evolution means that even two symmetrically evolving sequences will almost never show exactly identical divergence values). If the two likelihoods are very similar, that argues that any asymmetry present in the sequences is due only to stochastic effects and that the symmetrical hypothesis is reasonable. Larger differences in likelihood indicate that the sequences are unlikely to have evolved symmetrically. To judge the significance of the likelihood differences, we used a likelihood ratio test. This test compares twice the log of the ratio of the likelihoods between the two models (the likelihood ratio statistic) to a chi-square distribution with 1 degree of freedom (Goldman 1993). The chi-square distribution is known to be the distribution of this statistic at the limit of infinite data (Sokal & Rohlf 1995).

To ascertain the validity of the chi-square distribution for our data, we applied parametric bootstrapping (Hillis *et al*. 1996) to one asymmetrically diverged gene pair from each of the four genomes we studied. Specifically, for each of these four gene triplets, we simulated the evolution of 100 gene triplets using the maximum likelihood parameter estimates from our example triplet under the assumption that $K_{a1}=K_{a2}$ (that is, assuming symmetric divergence). For each of these 100 simulated triplets we then obtained maximum likelihood estimates of K_a and K_s under both the unconstrained model and the latter model where $K_{a1}=K_{a2}$. The result of this procedure was a distribution of log-likelihood ratios that could be compared to a chi-squared distribution. For all four sets of simulations a chi-square goodness-of-fit test indicated that the distribution of likelihood differences is consistent with a chi-square distribution (results not shown).

Computational limitations prohibited taking a similar approach as our primary significance test.

Selection of gene duplicates

At first sight, it might seem most sensible to choose outgroup genes from a separate genome. However, this approach faces two serious obstacles: 1) currently available outgroup genomes are evolutionarily distant, showing saturation in synonymous sites for many genes; 2) it is often impossible to differentiate recent gene duplications in the test genome from the loss of ancient duplicates in the outgroup genome. In a recent paper, Kondrashov and collaborators (2002) attempted to avoid this problem by analyzing duplicates which were closer to each other in amino acid sequence than either was to the outgroup. This conservative approach can potentially lead to underestimation of the number of asymmetrically diverged gene pairs because some asymmetric pairs may violate this requirement.

For these reasons, we pursued a within-genome approach in our four genomes (the baker's yeast *Saccharomyces cerevisiae*, the fission yeast *Schizosaccharomyces pombe*, the nematode worm *Caenorhabditis elegans*, and the fruit fly *Drosophila melanogaster)*. We identified triplets of genes closest to each other in synonymous divergence, K_s, using our whole genome analysis tool (Conant & Wagner 2002, reproduced here as the appendix). We considered the two closet members of the triplet (in terms of K_s) to be the duplicates, while the third gene constituted the outgroup. When faced with multiple outgroup choices (in gene families of more than three genes), we chose the closest outgroup gene, because outgroups with shorter branch lengths yield more trustworthy divergence estimates (Muse & Weir 1992).

We excluded gene triplets where (i) the outgroup gene showed less than 40% amino acid identity to the other two genes, (ii) any genes differed in length by more than 20%, (iii) members of a triplet were alternatively spliced version of the same gene, and (iv) member genes showed saturation in synonymous divergence (K_s). We determined saturation in K_s with a heuristic test: saturation was inferred if there was no decrease in the likelihood of observing the sequence data when the divergence (K_s value) for the sequence was increased beyond the maximum likelihood estimate (Hahn *et al*. 2004).

The complex phylogenies of large gene families make determining duplication orders difficult, leading us to exclude gene families of 9 or more members from analysis.

Assessment of asymmetry in duplicates

We aligned triplets using Clustalw (Thompson *et al*. 1994), removed gap characters, and calculated the likelihood of observing these alignments under two evolutionary models: (i) an unconstrained model (distinct K_a and K_s values); (ii) a model where the duplicates were constrained to have $K_{a1}=K_{a2}$. Nucleotide frequencies were estimated from the sequence alignments. Cases where pairwise K_s estimates had incorrectly identified the outgroup were corrected manually (5 triplets in fruit fly and 7 triplets in worm). We also excluded from analysis triplets with highly diverged outgroups ($K_{so}>4$ or $K_{ao}>1$), because longer outgroup branches decrease sensitivity to asymmetries. To obtain maximum likelihood estimates, branch lengths were optimized by the method of Yang (2000) for the unconstrained model; all other parameters and all parameters in the constrained model were estimated using Powell's routine (Press *et al*. 1992). In the remaining gene triplets, a likelihood ratio statistic greater than 3.85 (chi-squared $\mathbf{P}\leq 0.05$) between the two models indicated asymmetrical amino acid divergence. Analysis of all identified asymmetric

pairs (**P**≤0.05) with a model that allowed each codon position to have its own nucleotide frequencies did not affect our conclusions (results not shown).

Significance of observed patterns of asymmetry

Using a **P**=0.05 significance cutoff for repeated statistical tests can lead to elevated type I errors (false positives). Although this problem can be avoided with a Bonferroni correction (Sokal & Rohlf 1995: adjusts the P-value of individual tests to give a desired "family" error rate), such corrections reduce the power of individual tests. For our purposes, it is less important to minimize false positives than to discover whether the number of apparently asymmetrically diverged genes in a genome can be explained by chance. We therefore took a different approach to assess false positives. With a significance cutoff of **P**=0.05, we would expect 5% of the individual triplet tests to falsely reject the null-hypothesis of symmetric divergence. We used a binomial distribution with parameter p=0.05 to ask: "How likely would it be to observe the actual number of asymmetric triplets due solely to false positives?"

Functional distribution of asymmetric pairs

We used public databases for annotations: the Saccharomyces Genome Database (baker's yeast: Cherry *et al*. 1998), the *S. pombe* genome sequence (fission yeast: Wood *et al*. 2002), Flybase (fruit fly: The FlyBase Consortium 2002), and WormBase (nematode: Stein *et al*. 2001).

Analysis of expression profiles

Using microarray expression data from baker's yeast (Gasch *et al*. 2000) and worm (Kim *et al*. 2001), we asked whether there was a statistical association between sequence

asymmetry and a) expression divergence and b) asymmetry of expression divergence. In the baker's yeast data (time-series data for 11 experimental conditions), we used $\log_2$-transformed ratios of fluorescence intensities at previously described time-points where (on average) maximal induction or repression was seen (see Wagner 2002). For worm, we again used $\log_2$-transformed ratios, accepting only gene pairs where at least 100 matching microarray data points were available. Data were normalized by the number of experiments per pair. For both organisms, we treated as constant all data points with less than 2-fold expression change. Requiring 4-fold expression change excluded too many data points in yeast to permit analysis but gave similar results in the worm (not shown). To avoid microarray cross-reactivity between recent duplicates, we excluded pairs with $K_{s1}+K_{s2}<0.1$.

To answer part (a) of the above question, we calculated the absolute value of the difference in transformed ratios between our duplicate pair, summed over all conditions. We compared this net expression deviation to the normalized absolute difference in amino acid divergence K_a between the duplicates, given by

$$\left|\frac{K_{a1}-K_{a2}}{K_{a1}+K_{a2}}\right| \qquad (1)$$

For part (b), we counted the number of experimental conditions where a gene was over- or under expressed by at least 2-fold, a crude indicator of the number of conditions under which each duplicate has a significant change in expression. If expression patterns have diverged asymmetrically, one gene will show expression change in a significantly greater number of conditions that the other (Wagner 2002). We compared the difference in the number of changed conditions to the normalized difference in K_a (equation 1).

Asymmetric divergence and selective constraints

To determine whether asymmetry in amino acid divergence, K_a, was correlated with relaxed selective constraints, we calculated the correlation between the absolute value of the normalized difference in K_a, (equation 1), and the absolute value of the normalized difference in selective constraint, measured by:

$$\left| \frac{K_{a1}/K_{s1} - K_{a2}/K_{s2}}{K_{a1}/K_{s1} + K_{a2}/K_{s2}} \right| \qquad (2)$$

Fission yeast was excluded from this analysis because of its small sample size.

Cases of positive selection ($K_a/K_s > 1.0$)

We tested whether observed values of $K_a/K_s > 1.0$ were significantly different from one with another likelihood ratio test. Here the constrained model has $K_{ai}/K_{si}=1$ if duplicate *i* has $K_{ai}/K_{si}>1$ (i=1,2). As above, we used a chi-square distribution with 1 degree of freedom to test the significance of the observed difference in likelihood. We compared the triplet-based method of identifying cases of $K_a/K_s > 1.0$ to conventional pairwise methods, calculating pairwise significances also using a likelihood ratio test of $K_a/K_s >1.0$.

Results

Figure 2 shows a simple measure of asymmetric divergence: the absolute difference $|K_{a1}-K_{a2}|$ of the number of amino acid replacement substitutions per site for each gene pair analyzed, plotted against that pair's statistical significance **P** (as $-\log_{10}$ **P**). As described in the methods, we tested the hypothesis that the number of pairs with asymmetric K_a

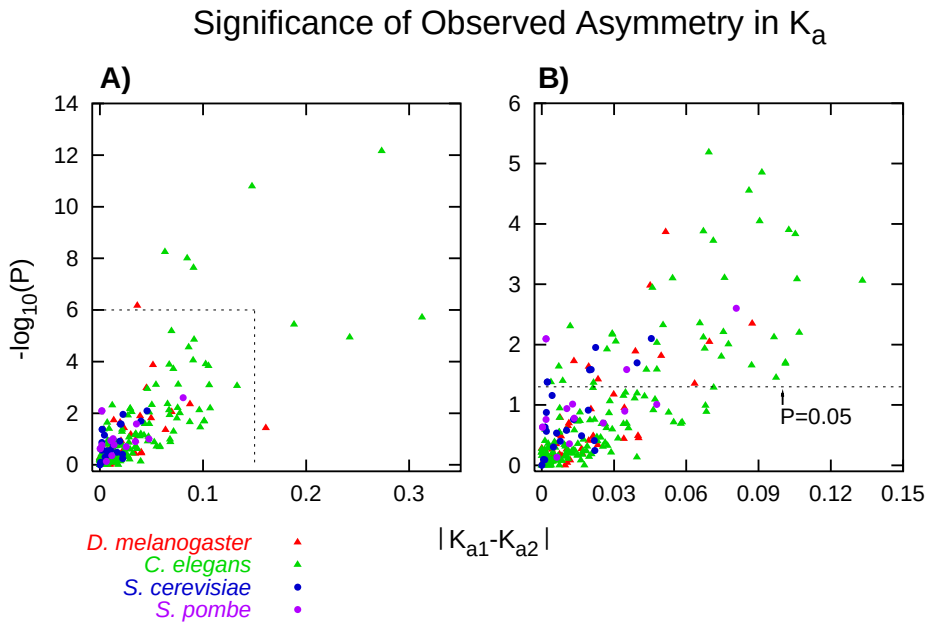


Figure 2: Significance of observed differences in K_a. On the x-axis is plotted the absolute value of the difference in K_a value between two duplicates. The y-axis gives the negative logarithm (to base 10) of the P value for that pair. **A)** All asymmetric duplicate pairs shown. **B)** Only the square region marked in panel A is shown. The dashed line in B shows the significance level of P=0.05.

values could be explained by the 5% error rate of our individual hypothesis tests. For all four genomes, we must reject this null hypothesis (baker's yeast: **P**=7x10^{-5}, fission yeast: **P**=0.0042, fruit fly: **P**=2x10^{-8}, worm: **P**=6x10^{-15}). Clearly, many gene pairs diverge asymmetrically. A full list of the identified asymmetric pairs is available from our website (http://www.unm.edu/~compbio/Supplemental_Data/Sequence_Asymm/). Below we discuss, species by species, the number of asymmetric pairs and highlight a few examples.

Saccharomyces cerevisiae. In baker's yeast we identified 22 gene triplets with unsaturated K_s, six of which (27%) showed asymmetry in K_a. An example is the gene pair encoding the alcohol dehydrogenase enzymes ADH3 and ADH1. ADH3 showed an amino acid divergence K_a nearly twice that for ADH1 (K_{a1}=0.101, K_{a2}=0.056, **P**=0.008). Interestingly, ADH3 is localized in the mitochondrial matrix (Pilgrim & Young 1987),

whereas ADH1 is found in the cytosol (Fraenkel 1982). The alcohol dehydrogenase genes ADH5 and ADH2 also showed asymmetric divergence, but their subcellular localization is unknown. In addition to the alcohol dehydrogenases, the acid phosphatases PHO3 and PHO5, as well as the pyruvate decarboxylases PDC1 and PDC5 showed asymmetric divergence.

Schizosaccharomyces pombe. In fission yeast, we identified 14 unsaturated triplets, 3 (21%) of which showed asymmetry in K_a. One especially clear-cut case of asymmetry regards the putative aminotransferase genes 19076066 and 19111920. Here, the outgroup (gene 19114182) is very distant from the duplicates (unsaturated K_{so} of 3.338), making it especially unlikely that the observed asymmetry is a result of incorrect outgroup selection. Asymmetry in K_a is highly significant for these two gene pairs, with K_{a2} nearly 80% greater than K_{a1} (K_{a1}=0.101, K_{a2}=0.181, **P**=0.003). The other asymmetric pairs were putative lysophospholipases and the retrotransposons *Tf2-11* and *Tf2-12*.

Drosophila melanogaster. We identified a total of 44 unsaturated triplets in the fruit fly, of which 13 (30%) showed evidence of asymmetric divergence. Among the asymmetrically diverged gene pairs with known function are the heat shock proteins Hsp-70Aa and Hsp-70-3, the beta tubulins 60D and 56D, as well as cytochrome P450 genes Cyp313a1 and Cyp313a2. The genes in the beta tubulin triplet all have different tissue-specific expression. β-tubulin 56D is the predominant isoform, while β-tubulin 60D is expressed in various larval, pupal, and adult cells (Hoyle & Raff 1990; Kimbel *et al*. 1989). The outgroup for these two genes, β-tubulin 85D, is an isoform specific to the male germ line (Fackenthal *et al*. 1995; Fackenthal *et al*. 1993).

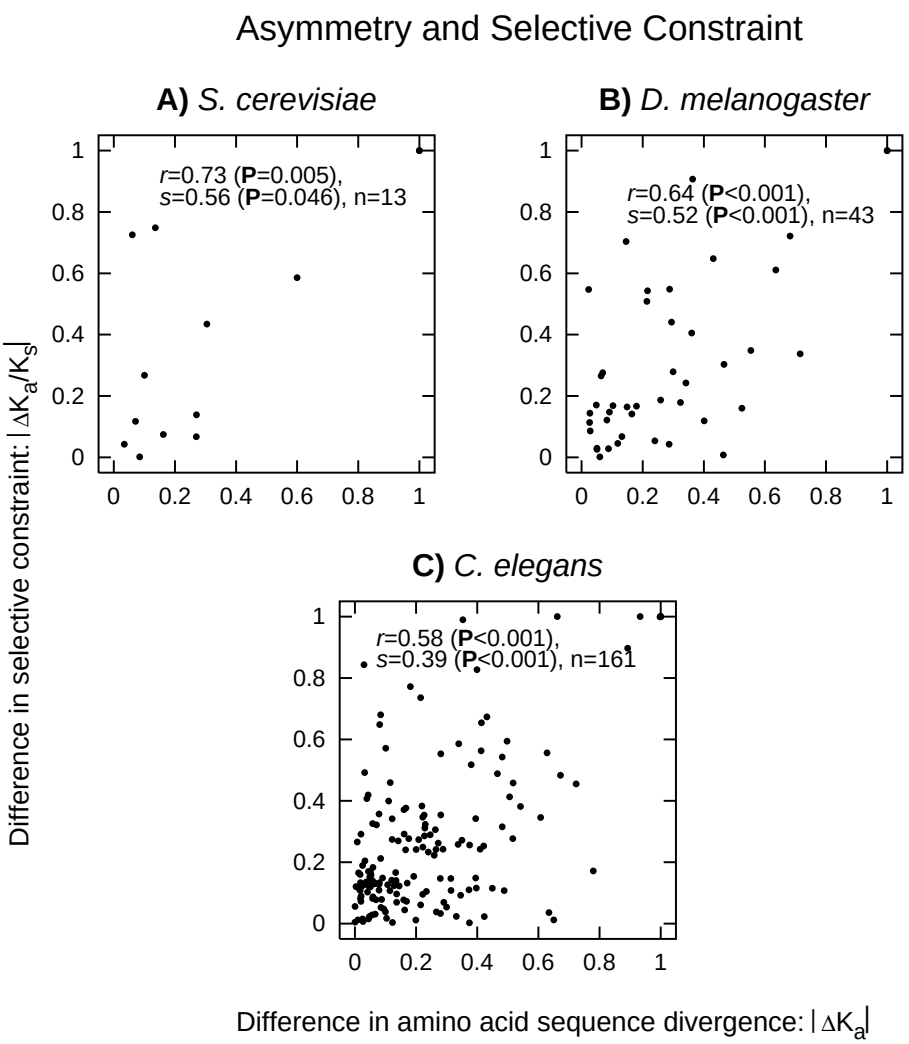


Figure 3: Correlation of the normalized difference in K_a between two duplicates and the difference in the selective constraint for those two duplicates. The variables on the axes labels are $|\Delta K_a| = |(K_{a1} - K_{a2})/(K_{a1} + K_{a2})|$ and $|\Delta K_a/K_s| = |(K_{a1}/K_{s1} - K_{a2}/K_{s2})/(K_{a1}/K_{s1} + K_{a2}/K_{s2})|$, respectively. **A)** *S. cerevisiae*. **B)** *D. melanogaster*. **C)** *C. elegans.*

An extreme example of asymmetric divergence in the fruit fly regards the *LysB* and *LysE* genes (outgroup *LysD*). The two genes show similar K_s values (K_{s1}=0.054, K_{s2}=0.057), but distinct K_a values (K_{a1}=0, K_{a2}=0.013). All three genes belong to the lysozyme D gene family. This gene family is expressed in the larval midgut (Daffre *et al*. 1994; Kylsten *et al*. 1992) and its members have chitinase activity (Regel *et al*. 1998), an interesting parallel to the asymmetrically diverged chitinase genes of *C. elegans* (see below).

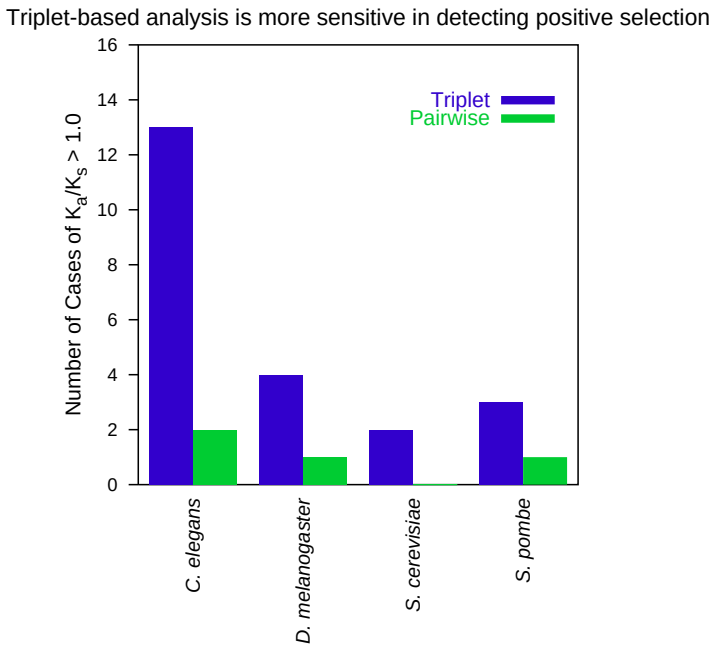


Figure 4: Number of cases where at least one duplicate in a pair has K_a/K_s >1.0 for the four organisms shown using our triplet method (black), and using the conventional pairwise estimation (grey). The total number of duplicates pairs for which we determined K_a/K_s in this analysis was 22 for *S. cerevisiae*, 14 for *S. pombe*, 44 for *D. melanogaster* and 164 for *C. elegans*.

Caenorhabditis elegans. We found 164 unsaturated triplets in the worm genome, 46 (28%) of which show asymmetric divergence. Six of the asymmetric pairs contain 7-helix transmembrane chemoreceptor domains. This nematode uses chemical signals to locate food (Delattre & Félix 2001), attract mates (Simon & Sternberg 2002) and initiate the social feeding response (de Bono *et al*. 2002), and the divergence of such pairs may increase the specificity of responses to these signals.

Like their fruit fly counterparts, two pairs of worm cytochrome P450 genes evolved asymmetrically. Cytochrome P450 is involved in detoxifying xenobiotics (Mathews & Van Holde 1996) and some worm family members have been shown to be xenobiotically inducible (Menzel *et al*. 2001). The asymmetric evolution of these genes may be related to challenges from environmental toxins.

Two further functional families with asymmetric pairs contain proteins with F-box domains and chitinases. Chitin is present in some nematodes' prey and in their own eggs (Muzzarelli & Muzzarelli 1998), suggesting the need for specialized chitinase enzymes.

Asymmetric amino acid divergence and gene expression profiles

We tested the hypothesis that asymmetric amino acid divergence is coupled to greater gene expression divergence in one of two duplicate genes. To do so, we used data from mRNA microarray experiments in yeast (Gasch *et al*. 2000) and nematodes (Kim *et al*. 2001). We found no significant correlation between degree of asymmetry in K_a and divergence in expression level. (Baker's yeast: Pearson's *r*: -0.28, **P**=0.33, Spearman's *s*: -0.17, **P**=0.56; n= 14; worm: Pearson's *r*: -0.01, **P**=0.90, Spearman's *s*: -0.04, **P**=0.69, n=119). We also calculated the statistical association between asymmetry in expression level (see Wagner 2002) and asymmetry in K_a. Once again we found no significant association (Baker's yeast: Pearson's *r*: 0.03, **P**=0.91, Spearman's *s*: 0.08, **P**=0.77; n= 14; worm: Pearson's *r*:-0.04, **P**=0.64, Spearman's *s*:-0.10, **P**=0.30, n=119).

Asymmetry and strength of selection

We examined the statistical association between asymmetry in amino acid divergence and evolutionary constraints on duplicate pairs, as indicated by K_a/K_s (see Methods). We excluded fission yeast from this analysis because of its small number of informative gene triplets. To avoid artifacts resulting from codon usage bias in baker's yeast, we excluded gene pairs where either gene had a codon bias index (Bennetzen & Hall 1982) value greater than 0.5. In baker's yeast (figure 3a) we observed a weakly significant correlation between the asymmetric amino acid divergence and selective constraint (Pearson's *r*=0.73, **P**=0.005, Spearman's *s*=0.56, **P**=0.046, n=13). The larger samples from fruit fly (figure 3b) and worm (figure 3c) both yield highly significant correlations (fruit fly: Pearson's *r*=0.64, Spearman's *s*=0.52, n=43, worm: Pearson's *r*=0.58, Spearman's *s*=0.39, n=161, **P**<0.001 for all).

Positive selection in duplicate genes

The triplet-based method of analysis permits estimation of the K_a/K_s ratio for each gene in a duplicate pair separately. As Figure 4 makes clear, such separate estimation produces many more candidate cases of $K_a/K_s > 1.0$ than does conventional pairwise analysis. None of the pairwise cases of $K_a/K_s > 1$ are statistically significant at $\mathbf{P} \leq 0.05$. In the triplet data, we found one significant case of positive selection out of 22 gene duplicates with $K_a/K_s > 1.0$: the worm gene *Y56A3A.10* (duplicate partner: *Y56A3A.14*, outgroup: *Y56A3A.15*, **P**=0.029). This gene pair also shows significant asymmetry in K_a divergence (P=0.009), consistent with the notion that one of the genes underwent directional selection. Functional information about these genes is limited, except that all three genes contain an F-box domain which is involved in ubiquitin-mediated protein degradation and spermatogenesis (Kipreos & Pagano 2000).

Discussion

Asymmetries in rates of amino acid divergence are common in our four test genomes. Sample sizes are small in some genomes (e.g., 14 gene pairs for fission yeast), but taken together, our results suggest that a genome contains at least 20% of gene duplicates that diverge asymmetrically. The largest samples come from the worm and fruit fly genomes, where between 28% and 30% of gene pairs showed asymmetric divergence. Our estimate lies in between that of Van de Peer and collaborators (Van de Peer *et al*. 2001) for DNA sequence divergence in vertebrate duplicates (50%), and that of another systematic study using various completely sequenced genomes (<5%) (Kondrashov *et al*. 2002). Differences in approach may be responsible for these discrepancies. For example, Kondrashov and collaborators required that two duplicates be closer in amino acid

sequence to each other than to the outgroup. This is a sensible assumption, but it leads to an underestimate of the number of asymmetrically diverged genes, because if asymmetric divergence occurs, one of the duplicates may have become more divergent than the outgroup gene. In addition, amino acid models (such as that used by Kondrashov *et. al.*) do not directly consider the structure of the genetic code in their estimates. Such models may thus underestimate divergence, since the codons for some amino acids are separated by multiple nonsynonymous nucleotide substitutions in the genetic code (Ota & Nei 1994). We circumvent these potential biases by using a codon model and requiring that the *synonymous* divergence of duplicates must be lower than that of the outgroup.

Caveats. The biggest caveat to our approach is that it requires gene triplets that meet several stringent criteria (see methods). It will thus yield only a moderate number of informative gene pairs. We also note that our approach may still occasionally miss asymmetrically diverged duplicates, especially if outgroup branches are long, or if the genes in question are short. In that case, the number of asymmetrically diverging genes would be higher than observed.

Asymmetry in sequence divergence and functional divergence. Is asymmetric divergence in sequence coupled to asymmetric divergence in gene function? Do rapidly evolving duplicates acquire new functions more often than slowly evolving duplicates? These obvious questions are very difficult to answer systematically. First, many asymmetrically diverging gene pairs have completely unknown function. Second, reliable indirect indicators of gene function, such as gene expression patterns, are available for many genes only in a select few organisms. Among our four organisms,

baker's yeast and the nematode contain the information necessary to address such questions.

Indirect indicators of yeast gene function include gene expression (for examples see (Gasch *et al*. 2000; Spellman *et al*. 1998), subcellular localization (Kumar *et al*. 2002), protein interactions (Ito *et al*. 2001; Uetz *et al*. 2000), and the effects of synthetic null ("knock-out") mutations on the expression of other genes (Hughes *et al*. 2000b). Such information is available for anywhere from a few hundred genes in the case of gene knock-out effects on gene expression (Hughes *et al*. 2000b), to almost all genes in the case of gene expression data. Asymmetric divergence of gene duplicates has previously been detected in baker's yeast for several of these indicators of gene function (Wagner 2002). In worm, less functional data are available, but both large microarray experiments (Kim *et al*. 2001) and whole-genome RNAi knock-down experiments (Kamath *et al*. 2003) have been performed. Is such functional asymmetry correlated with sequence asymmetry? For expression data, the answer appears to be no. Asymmetric gene expression divergence is driven by the differential evolution of regulatory regions, not coding sequences. Because the evolution of coding sequences is only weakly coupled with that of gene expression (Wagner 2000a), it is unsurprising that gene expression divergence is uncoupled to asymmetric sequence divergence. We can currently not answer whether other indicators of functional divergence are more closely coupled to asymmetric sequence divergence in baker's yeast because of our small number of asymmetric triplets.

Why asymmetric sequence divergence? Two principal forces can drive the asymmetric divergence of genes: relaxed selective constraints and directional selection.

In the first case, sequence divergence is neutral, *i.e.*, it does not involve positive selection of advantageous mutations on the more rapidly evolving gene. In the second case, divergence follows a selectionist scenario, where advantageous mutations play prominent role. The answer to the above question would contribute important evidence to the neutralist-selectionist debate (Li 1997).

For baker's yeast, fruit fly, and worm there is clear evidence for relaxed selective constraints in asymmetrically diverged genes. That is, the more asymmetrically two genes diverge, the greater is the difference in the ratio of amino acid replacement to synonymous substitutions (K_a/K_s) between them. (Fig. 3). Such differences in constraints can arise if two duplicates come to be expressed in different cell compartments or tissues, encountering different interaction partners and chemical milieus. The asymmetrically diverged baker's yeast genes *ADH1* (cytosolic) and *ADH3* (mitochondrial) (Fraenkel 1982; Pilgrim & Young 1987) and the β-tubulins in fruit fly constitute examples of this phenomenon.

We also detected several candidate gene pairs where positive selection may have taken place. Positive selection is indicated when the rate of nonsynonymous substitutions (K_a) exceeds the rate of synonymous substitutions (K_s) ($K_a/K_s>1$, Li 1997). Because the current model does not assume gene duplicates evolve symmetrically, it can look for cases of $K_a/K_s >1.0$ in *individual* genes, potentially improving sensitivity in detecting positive selection over pairwise methods. This is born out by the data in Figure 4, which shows that the triplet-based method detects many more genes with $K_a/K_s>1.0$. However, only one among them has K_a/K_s significantly greater than one. This is an (asymmetric) worm gene pair containing an F-box domain which may be involved in spermatogenesis

(Kipreos & Pagano 2000). We note that positive selection has been shown in the male reproductive genes of other organisms (Nurminsky *et al*. 1998; Wyckoff *et al*. 2000).

At first sight, the above analysis suggests that neutral relaxation of selective constraints may be largely responsible for asymmetric divergence. However, this conclusion would be premature. K_a/K_s as an indicator of positive selection averages across all nucleotides in a gene and through time since divergence, but positive selection often acts only on a small fraction of key nucleotides over a short period. To detect positive selection with K_a/K_s requires that selection have been both strong and recent. Thus, although positive selection is pervasive (Fay *et al*. 2002; Hughes *et al*. 2000a; Hughes & Hughes 1993; Smith & Eyre-Walker 2002), the ratio K_a/K_s can usually be used to demonstrate positive selection only in conjunction with phylogenetic and functional information, or with information on amino acid polymorphisms (Tsaur *et al*. 1998; Zhang *et al*. 1998). This means that many of our gene duplicates that show relaxed selective constraint may actually have experienced positive selection which cannot be detected using sequence information alone. Again, functional genomic and phylogenetic data are necessary to arrive at a firm conclusion. What will this final conclusion be? If three decades worth of molecular evolution studies are any guide, asymmetric divergence will be due to neutral divergence for some genes, and due to positive selection for others.

Chapter 3: Duplicate genes and robustness to transient gene knockouts in *Caenorhabditis elegans*

This chapter has previously appeared in substantially the same form as: Conant, G. C. and Wagner, A. (2004) "Duplicate genes and robustness to transient gene knockouts in *Caenorhabditis elegans*, *Proceedings of the Royal Society, Biological Sciences,* **271(1534)**: 89-96.

Abstract:

We examine robustness to mutations in the nematode worm *Caenorhabditis elegans* and the role of single copy and duplicate genes in it. We do so by integrating complete genome sequence and microarray gene expression data with results from a genome-scale study using RNA interference (RNAi) to temporarily eliminate the functions of more than 18,000 worm genes. 89% of single copy and 96% of duplicate genes show no detectable phenotypic effect in an RNAi knock-down experiment. We find that mutational robustness is greatest for closely related gene duplicates, large gene families, and similarly expressed genes. We discuss the different causes of mutational robustness in single copy and duplicate genes, as well as its evolutionary origin.

Introduction

Genes whose loss of function has no detectable effect number in the thousands in a typical eukaryotic genome (Kamath *et al*. 2003; Steinmetz *et al*. 2002; Winzeler *et al*. 1999). Duplicate genes comprise at least one third of eukaryotic genomes (Li *et al*. 2001; Rubin *et al*. 2000), a fact that might explain this observation, because duplicate genes often have similar function. Losing one duplicate gene can thus be tolerated because others can buffer the organism against this loss. This candidate explanation for many genes without phenotypic effects is appealing but also inadequate. A systematic analysis of the effects of knock-out mutations in the yeast *Saccharomyces cerevisiae*, a single-celled eukaryote, showed that much robustness against null mutations is caused by single-copy genes (Wagner 2000b). This analysis, based on over 250 synthetic null (gene-knockout) mutations, found that more than 40% of mutations with no phenotypic

effect occurred in single copy genes. It also showed little support for the role of gene duplications in robustness, a result due to the limited amount of gene-knockout data available at the time. A more recent study (Gu *et al*. 2003), based on more than 5700 synthetic null mutations in yeast, showed that gene duplications have an important role in mutational robustness. However, this later study also underscored the importance of single copy genes in conferring robustness. Between 41% and 77% of non-detectable mutational effects were due to single-copy genes: a number higher even than that found in the more limited study.

Whether single copy or duplicate genes are primarily responsible for mutational robustness has implications for the mechanisms providing robustness. The question itself, however, has thus far only been asked in the unicellular eukaryote yeast. Multicellular organisms might yield different answers, both because they contain more duplicate genes which form larger families (Qian *et al*. 2001; Conant & Wagner, 2002—reproduced here as an appendix; Rubin *et al*. 2000), and because developmental processes that arose with multicellular life may rely on different mechanisms to buffer the effect of null mutations. A recent genome-wide analysis that transiently eliminated the function of more than 16,000 *C. elegans* genes through RNA interference (RNAi, Fire *et al*. 1998; Kamath *et al*. 2003) allowed us to ask this question for the first time in a higher organism.

Any such analysis has caveats. For example, RNAi only temporarily deactivates genes and may not reveal all effects of a synthetic null mutation. This fact, in addition to errors in genome annotation such as the accidental inclusion of pseudo-genes, may contribute to the low proportions of genes with phenotypic effects identified in the RNAi analysis of the worm genome (~9% of the genes studied below). However, the effects

that the RNAi approach detects are representative of those found with other approaches: 64% of *C. elegans* genes with known knock-out phenotypes can also be detected with RNAi, and of those, over 92% give RNAi phenotypes similar to those observed previously (Kamath *et al*. 2003). Second, because RNAi relies on base complementarity between a (denatured) double-stranded RNA and its cognate mRNA, the method may not distinguish between closely-related gene duplicates. We alleviate this problem by using only the 13,565 genes for which an RNAi clone specific to the gene – and not affecting multiple targets – was available (Kamath *et al*. 2003) and for which an unambiguous identification in release 73 of wormpep (all protein-coding genes in the *C. elegans* genome, Stein *et al*. 2001) could be made. Third, unlike microbes, where growth rate differences can be measured with great accuracy (Steinmetz *et al*. 2002), indicators of fitness cannot be as reliably estimated for multicellular organisms. Fourth – and this is a limitation shared by all laboratory studies – phenotypic effects of mutations are usually only assessed in a small number of environments. That is, they do not necessarily reflect fitness differences in the natural environment. Despite this caveat, the resolution of such experiments is sufficient for our purpose: to distinguish the role of duplicate and single-copy genes in the buffering of mutations.

Methods

Identification of gene duplicates

We identified duplicates in the *C. elegans* genome (The *C. elegans* Sequencing Consortium 1998) using our previously described whole-genome analysis tool (Conant & Wagner 2002, reproduced here as the appendix). For this analysis, we used only

duplicate pairs separated by a nonsynonymous distance (K_a) of 1.0 or less (calculated by the maximum likelihood method of Muse and Gaut/Goldman and Yang; 1994). Use of a more liberal threshold of K_a<2.0 identified only 20 more duplicate genes, suggesting that our results are not strongly biased by this cut-off. Genes not identified as duplicates under these criteria were treated as single copy genes. We used release 73 of wormpep for this analysis (Stein *et al*. 2001), and only genes present in this release of the genome were analysed.

RNAi interference (knock-down) data

Data on gene knock-down effects were obtained from the RNA interference (Fire *et al*. 1998) experiments of Kamath and collaborators (Kamath *et al*. 2003). Because interfering RNAs may not distinguish between closely related gene duplicates, we excluded clones annotated as affecting multiple targets (Kamath *et al*. 2003).

We grouped phenotypic knock-down effects into three categories: no phenotype, viable but detectable phenotype, and lethal phenotypes and assigned numerical scores to the categories in order of increasing defect: 0 for no phenotype, 1 for moderate (viable) phenotype, and 2 for lethal phenotype.

Effect of gene family size and evolutionary distance on knock-down phenotype

We first asked whether the distribution of genes into the three phenotypic categories was affected by the number of paralogs a gene has. We grouped genes into 5 classes (genes with 0, 1, 2-3, 4-7, or 8 eight or more paralogs, see figure 5). We then asked whether the proportion of genes with the three phenotypes differed (i) between

single copy genes and genes that have (at least one) duplicate, and (ii) between the gene families of various sizes shown in figure 5.

To address question (i), we calculated the expected number of genes with each of the three knock-down effects among the genes with at least one duplicate, using the phenotypic proportions seen in the single copy genes. By comparing these 3 expected values to the observed number of genes of each phenotype among duplicated genes, we were able to use a χ^2 goodness-of –fit test with 2 degrees of freedom to ascertain statistical significance.

To address question (ii), we used the same approach, limiting our comparisons to adjacent duplication classes. For example, we asked whether the phenotype distribution is the same for genes with one duplicate as for genes with 2-3 duplicates.

To ask whether phenotypic effects were correlated with the evolutionary distance between duplicates, we compared the proportion of genes in the three phenotypic categories to both the amino acid distance (the fraction K_a of substitutions per non-synonymous site) between closest duplicates and the synonymous distance (the fraction K_s of substitutions per synonymous site) between closest duplicates. We calculated the Pearson product-moment correlation *r* between the distance (K_s or K_a) and the proportion of genes in each of the three phenotypic categories (figures 6 and 8). In the case of K_s, we included only duplicate pairs where K_s <2.0 and in which both genes showed an effective number of codons (ENC, Wright 1990) greater than 43. This choice of ENC cut-off excludes approximately 10% of genes in the *C. elegans* genome with the lowest values of ENC. Although failing to exclude any genes with high codon bias yields a correlation between knock-down effect and K_s (presumably due to the association between

expression level and knock-down effect seen in figure 9), varying the ENC cut-off so as to exclude between 4% and 30% of genes yields the same result (no significant association) as reported in Results (data not shown). To test the statistical significance of *r*, we randomly reshuffled the phenotypic effects with respect to the distances 1000 times and recalculated *r* for each reshuffled data set.

To examine whether duplicate genes show similar phenotypic effects, we counted the number of duplicate gene pairs within a given window of K_a where one member showed no knock-down phenotype and the other showed either a lethal or a moderate phenotype. We tested for significance using the same randomisation test.

Association between knock-down effect and gene expression

To identify a statistical relationship between knock-down effect and gene expression, we used a large microarray expression dataset comprising 553 experiments and most *C. elegans* genes (Kim *et al*. 2001). The data consists of logarithmically ($\log_2$) transformed expression changes relative to a reference condition that depends on the particular experiment (Kim *et al*. 2001). We identified pairs of duplicate genes (see above) for which RNAi data were present and which were separated by a pairwise K_s of 0.2 or more. For each duplicate pair, we assembled all microarray experiments for which data were available for both genes and calculated the Pearson correlation coefficient (*r*) between the two genes' expression changes. We then calculated the correlation between this expression similarity of the pairs and their average knock-down effect (calculated using the numerical scheme above). A randomisation analysis was used for significance testing. We also repeated this analysis substituting the proportion of gene pairs where exactly one gene had a lethal effect (see results) for the average knock-down effect.

Using the duplicate pairs identified for the expression analysis above, we next calculated the statistical association between the pairwise correlations in duplicate expression from the experiments by Kim and collaborators (see above) and the pairwise K_a between the duplicates (figure 10).

To assess whether highly expressed genes show strong knock-down effects, we used results of an experiment (Hill *et al*. 2000) that had determined the expression levels of 18,791 *C. elegans* open-reading frames at 8 time points during the worm's lifecycle. Using Affymetrix gene chips, these authors estimated the concentration of transcripts (in parts per million) at each time point. We considered only the 2624 genes where RNAi knock-down data were available, where a transcript was detected by all hybridization replicates, and where that transcript showed an expression level above 20 parts per million (ppm). We compared the $\log_{10}$ transform of each gene's highest concentration across the eight timepoints to the RNAi knock-down effect. We again evaluated significance using a randomization test as outlined above.

Our final analysis compared the level of gene expression (using the same concentration values as above) to amino acid distance (K_a). This analysis allowed us to judge whether effects of sequence and expression similarity on knock-down might be differing measures of the same underlying phenomena. Only genes which appeared in the knock-down data generated by Kamath and coauthors were used for this analysis. We again used the base-ten logarithm of the maximum concentration, comparing it to K_a and determining significance with a permutation test.

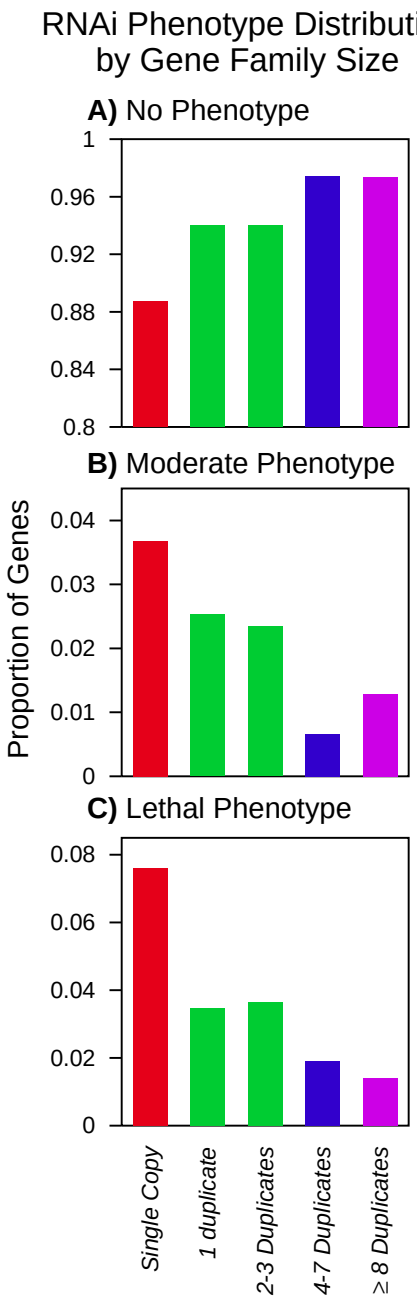


Figure 5: Proportions of genes with **A)** no RNAi knock-down phenotype, **B)** a detectable and viable phenotype, and **C)** a lethal phenotype, categorized by gene family size, that is, the number of paralogs per gene. Absolute numbers of genes in each family size category are 8861 (single copy genes), 316 (genes with 1 duplicate), 1624 (genes with 2-3 duplicates), 1209 (genes with 4-7 duplicates), and 1555 (genes with 8 or more duplicates).

Results

Many weak phenotypic effects are due to single-copy genes. We grouped phenotypic effects of RNAi interference (gene knock-down) into three categories: no phenotype (12,387 genes), viable but detectable (moderate) phenotype (395 genes), and lethal phenotypes (783 genes). The detectable category groups phenotypes with slow or arrested post-embryonic growth and post-embryonic phenotypes without such growth defects (Kamath *et al*. 2003) together.

Among the 13565 genes analyzed, 8861 are single copy, of which 88.8% (7872) show no detectable phenotype in an RNAi knock-down experiment. The proportion of genes with no RNAi phenotypes that occur in gene families of size two or greater is somewhat larger: 96.0% (4515 of 4704) of such genes have no knock-down phenotype. For the other two classes of phenotypes, the relationship is reversed: more lethal phenotypes are due to single copy genes

(7.5% or 668 genes) than to duplicated genes (2.4% or 115 genes), as are more moderate phenotypes (single copy genes: 3.6% or 321 genes; duplicate genes: 1.6% or 74 genes). The large numbers of genes involved makes even these small differences statistically highly significant (χ^2=243; df=2; **P**<10^{-10}). We now examine in greater detail the relationship between gene family size and RNAi phenotype.

Gene family size is correlated with RNAi phenotype. Figure 5 demonstrates a correlation between the size of a gene family and the frequency of the different RNAi knock-down effects. Specifically, the larger a gene family, the more likely that its members have no RNAi phenotype (Figure 5A), and the less likely that they have either a detectable (Figure 5B) or a lethal phenotype (Figure 5C). Absolute differences in proportions are again small: 88.8% of single-copy genes but 94.0% of genes with one duplicate have no detectable RNAi phenotype. We asked whether any two adjacent size categories in the panels of Figure 5 contain equal proportions of genes (see methods). Because we are making four comparisons in this analysis, we used a Bonferroni correction (Sokal & Rohlf 1995), performing individual tests at a significance level of 0.0125 to yield a family error rate of 0.05. Adjacent categories of gene family sizes with the same coloring in figure 5 indicate cases where we cannot reject the hypothesis of equal proportions across the three phenotypes. Only the categories with one duplicate and with 2-3 duplicates show such equal proportions: all others contain different proportions of genes (**P**≤0.0125). In sum, there is strong evidence that the phenotypic effect detected in knock-down experiments changes with increasing gene family size.

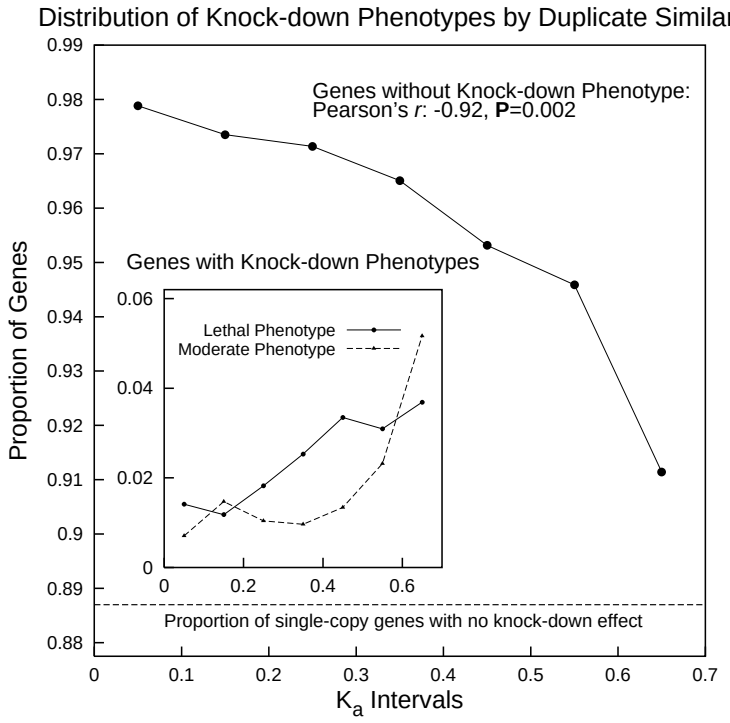


Figure 6: Relationship of nonsynonymous distance to nearest gene duplicate (K_a) and proportion of genes with no RNAi knock-down phenotype. 4639 total gene pairs were analyzed. Inset shows proportions of viable and lethal knock-down phenotypes.

The more similar two duplicates are, the less severe is their knock-down effect. We next examined the proportion of genes with a given phenotypic effect as a function of similarity between duplicates, using the amino acid distance K_a (number of non-synonymous substitutions per non-synonymous site, Li 1997) to measure similarity. The proportion of genes with no phenotypic effect decreases with amino acid distance to the nearest paralog (Fig. 6; Pearson's r =-0.92; n=4639, **P**=0.002, significance calculated using a permutation test on the binned data, see methods). Likewise, the proportion of genes with moderate and lethal effects increases with increasing amino acid distance (Fig. 6; inset; r=0.77; **P**=0.04 and r=0.95; **P**=0.001, respectively, n=4639 for both). We also asked whether two duplicates generally have similar knock-down effects and found that amino acid distance (K_a) and the proportion of duplicate pairs with different knock-down effects have a strong positive correlation (Pearson's r=0.96; n=3314; **P**<0.001). That is, the more distant two duplicates are, the more likely it is that

one of them has a more severe knock-down effect than the other (Figure 7). Previous genome-scale analyses in various organisms showed that many duplicate genes have asymmetric sequence or functional divergence, as indicated by protein interactions, sequence divergence, and gene expression patterns (Wagner 2002). For example, for some 30% of worm duplicate genes, one duplicate diverges faster than the other on the amino acid level (Conant & Wagner 2003, chapter 2 of this manuscript). Asymmetric divergence, which may increase with amino acid distance and divergence time, could explain why distantly related duplicates often show different mutational effects.

As noted above, we removed from our analysis all genes with possible cross-reactivity according to Kamath and collaborators (2003). In addition, we assessed whether there were any remaining cross-reactivity biases in the above two analyses by repeating these analyses excluding gene pairs with $K_a<0.1$. Doing so changed neither the association of knock-down effect and amino acid sequence similarity nor the association of asymmetry of knock-down effect and sequence similarity (data not shown).

We also assessed whether time since duplication affects knock-down phenotypes by comparing knock-down effect to K_s (number of substitutions per synonymous site, Li 1997). K_s is a better indicator of divergence time than K_a because it is subject to fewer evolutionary constraints and thus may change at an approximately constant (neutral) rate (Li 1997). Interpretation of K_s values is confounded by codon usage bias, a feature of very highly expressed genes which can lead to slower rates of synonymous evolution in such genes (Bernardi & Bernardi 1986; Comeron & Aguade 1998). In *C. elegans*, a measure of codon usage bias is the effective number of codons (ENC, Wright 1990). It shows a significant correlation (Pearson's r = -0.57, Spearman's s=-0.45, n=3160,

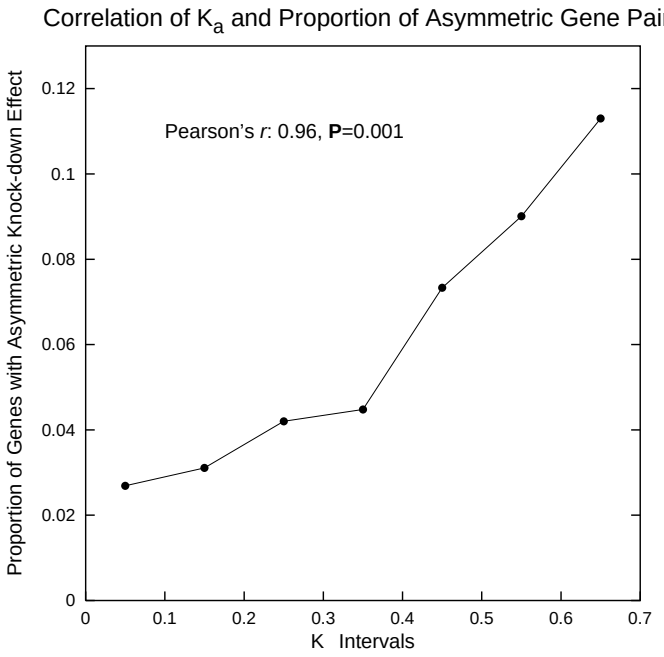


Figure 7: Relationship of nonsynonymous distance to nearest duplicate (K_a, x-axis) and proportion of genes with asymmetric knock-down effects (y-axis; see text for details).

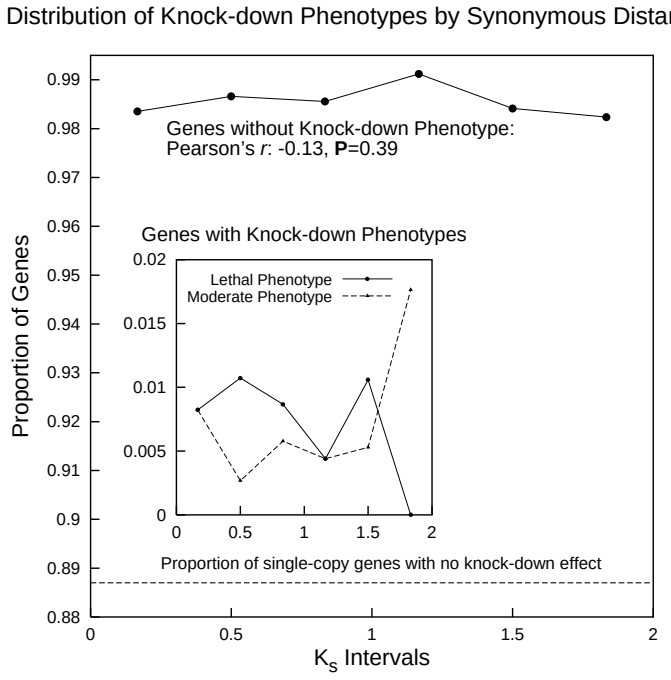


Figure 8: Relationship of synonymous distance to nearest gene duplicate (K_s) and proportion of genes with no RNAi knock-down phenotype. Inset shows proportions of viable and lethal knock-down phenotypes.

P<0.0001 for both) to a gene's maximum expression level during *C. elegans* development, as measured by oligonucleotide microarrays (Hill *et al*. 2000). We thus eliminated genes with a high codon usage bias (low ENC) before analysis. The remaining genes showed no significant association between K_s and the propensity to have either no, a viable, or a lethal phenotypic defect (*n*=1791, no phenotype: *r* =-0.13, **P**=0.39; viable phenotypes: *r*=0.53, **P**=0.10; lethal phenotypes: *r*=-0.59, **P**=0.14; see figure 8). To be certain that this lack of association is not an

artefact of our permutation test, we have also applied a χ^2-goodness-of-fit test to these data, testing the null hypothesis that the different ranges of K_s all show the same proportions of null, moderate and lethal phenotypes. This test is conservative in the sense that it can reject the null hypothesis even if there is no linear trend in the data. However, the χ^2 test reinforces our conclusions of no association (χ^2=6.7; df=17; P=0.99).

Expression level and knock-down effect. Similarity in amino acid sequence is only one indicator of functional similarity among gene duplicates. Studies of individual gene duplicates have shown that functional divergence sometimes occurs through diverging expression patterns rather than diverging sequences (Hanks *et al*. 1995; Li & Noll 1994; Wang *et al*. 1996). This raises the question whether expression divergence among gene duplicates, which is generally rapid (Gu *et al*. 2002; Wagner 2000a), is also associated with phenotypic effect. To address this question, we compared similarity in expression levels (see Methods) between duplicate genes to the average RNAi knock-down effect. To avoid artefacts from cross-reactivity in microarray experiments, we excluded duplicate pairs where K_s<0.2. There is a significant correlation between similarity of expression pattern and the average knock-down effect: (*r*=-0.80, n=3028, **P**=0.04, Figure 9A). We observe a similar association if we replace the average knock-down effect with the proportion of gene duplicates where one gene shows a lethal knock-down effect while the other does not (*r*=-0.88, n=3028, **P**=0.02). Excluding genes with high codon usage bias (low ENC) does not change this pattern either (low ENC; *r*=-0.81, n=2535, **P**=0.05).

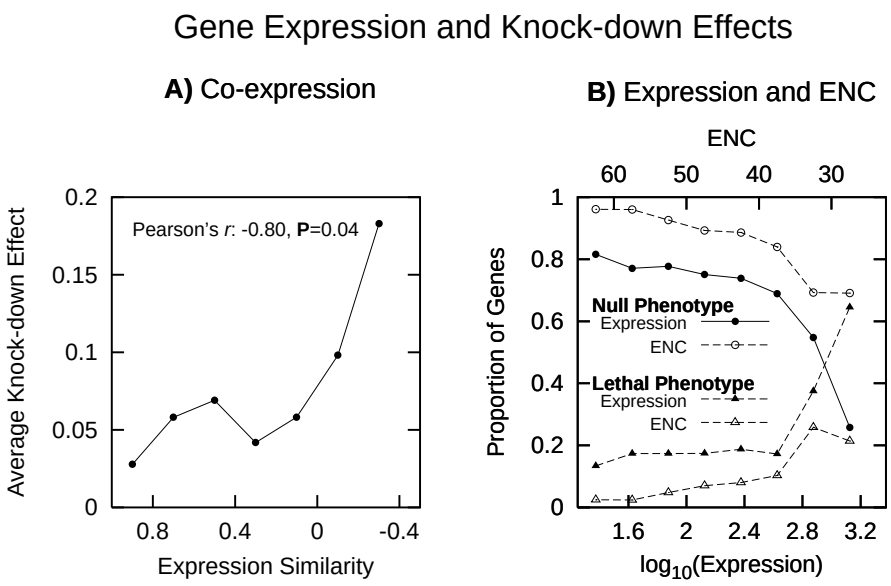


Figure 9: A) Association of duplicate expression similarity and average knock-down effect. The x-axis shows the Pearson's *r* for the correlation of the expression levels of the two duplicates, while the y-axis shows the average RNAi knock-down effect (see methods). **B)** Distribution of knock-down effects by expression level. The x-axes show two measures of gene expression: the $\log_{10}$ of the parts-per million counts for each gene (the relative expression level, taken at the maximal expression timepoint—see methods) or the effective number of codons (ENC—see methods). The y-axis shows the proportion of genes with no knock-down phenotype and with a lethal knock-down phenotype.

It is possible that sequence similarity and expression similarity co-vary, and hence that the association between knock-down effect and the two therefore reflects the same underlying phenomenon. However, the magnitude of the correlation in expression similarity and amino acid sequence distance is weak (Pearson's *r*:-0.29, Spearman's *s*:-0.27, n=3032, **P**<0.001 for both, figure 10, see methods). Moreover, considering only restricted ranges of K_a in the above analysis should eliminate the observed correlation in figure 9A if it is truly a result of the covariance of K_a and expression. We determined the association between gene expression level and knock-down effect separately for gene duplicates within five ranges of K_a (0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, and 0.5-0.6). Despite small sample size (several bins had fewer than 30 elements), four of the five bins showed a negative association, just as the complete data.

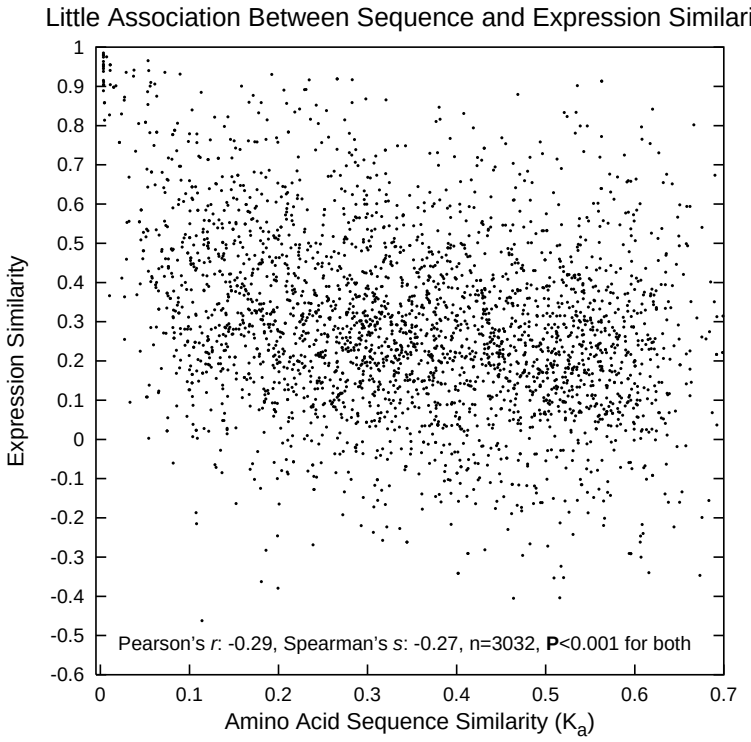


Figure 10: No strong association between the sequence similarity of a duplicate gene pair and the similarity of their expression patterns (see methods)

We also find a statistically significant relationship between maximal expression level and knock-down effect, consistent with the results of others in yeast (Gu *et al*. 2003; Pál *et al*. 2003). Using the highest expression level of each gene measured during eight time points in the worm's life cycle (Hill *et al*. 2000), we find that highly expressed genes are more likely to show a lethal effect (Pearson's r = 0.77; n=2624; **P**=0.02) and less likely to show no effect from knock-down (r= -0.83; n=2624; **P**=0.005; Figure 9B). A similar statistical association holds if codon usage bias (*low* ENC) is used as an indicator of high expression. (No effect knock-downs: Pearson's r=0.93, **P**=0.001, lethal knock-downs: Pearson's r=-0.89, **P**=0.005; n=13529 for both; Figure 9B). This result is unsurprising, given the negative correlation of ENC and microarray gene expression levels seen above.

Finally, it has been noted in yeast (Pál *et al*. 2001) that highly expressed genes are under stronger evolutionary constraints and thus evolve more slowly. Data from duplicate genes in *C. elegans* are consistent with this finding: Amino acid distance, K_a, and

expression level (data as described for figure 9B) show a significant negative correlation (Pearson's r = -0.80, **P**=0.01; n=1552).

Discussion

Although the absolute number (7872) of single-copy genes with no knock-down effect is higher than the number of duplicate genes with no knock-down effect, proportionally more duplicate genes have no knock-down effect than do single copy genes. Kamath and collaborators noted a similar pattern using a different method of identifying duplicates (Kamath *et al*. 2003). In addition, mutational robustness is greatest for closely related and similarly expressed gene duplicates, as well as for duplicates in large gene families. These findings show the important role of both single copy genes and duplicate genes in robustness against mutation. Weak knock-down phenotypes for duplicate genes can be explained by gene redundancy and overlapping gene functions. Much less clear is how single copy genes can be eliminated without detectable effect, even though this phenomenon is now established in two organisms. One possibility is that for many single-copy genes the worm genome harbors at least one other gene with a convergent function, yet no sequence similarity. Consistent with this possibility is the observation that sequence similarity search algorithms miss many genes with dissimilar sequence but convergent tertiary structure (Hubbard *et al*. 1998). Whether such convergent evolution could explain most cases of single-copy genes with no phenotypic effect is unknown. However, the massive scale – more than 7,000 genes – at which such convergence would have to occur makes this seem unlikely. A second possibility is that much mutational robustness is due to interactions of unrelated genes in genetic networks. Mechanistically, this kind of buffering is best understood in metabolic networks. Such networks can

compensate loss-of-function mutations in many (non-redundant) genes by rerouting the flux of metabolites through alternative pathways (Edwards & Palsson 2000).

Is gene redundancy more important in the multicellular worm than in the unicellular yeast? In yeast 39.5% of single copy genes versus 64.3% of duplicates gene account for synthetic null mutations with weak or no effect on growth (Gu *et al*. 2003), a ratio of 1:1.63. The proportions we find in the worm indicate a ratio of 1:1.08, less strongly skewed towards gene duplicates. Conversely, in yeast 29.0% of single copy genes versus 12.4% of duplicate gene account for synthetic null mutations with lethal effects, a ratio of 1:0.43. In the worm, the corresponding percentages are 7.5% and 2.4%, yielding a ratio of 1:0.32. From this perspective, gene duplication in the worm is less important than in yeast for causing weak phenotypic effects. However, gene duplication is slightly more important in the worm in preventing lethal phenotypic effects.

A complementary analysis follows that of Gu and collaborators, who estimated lower and upper bounds on the proportion of weak gene knockout effects that can be attributed to duplicate genes (Gu *et al*. 2003). Their lower bound derives from the assumption that the difference in proportions of mutations with no effect between single copy genes and duplicate genes is due to gene duplication. For our worm data, 89% of single copy genes and 96% of duplicate genes had no knock-down effect. This difference of 7% indicates that at least 323 duplicate genes show no knock-down effect because they are duplicates. The lower bound in the worm is thus approximately 3% (323/12387), compared to 23% in yeast. The main caveat to this lower bound is that RNAi detects fewer phenotypic effects than does gene knockout in yeast, biasing the estimate. To obtain an upper bound on the contribution by gene duplicates, Gu and

collaborators assumed that all weak knock-out effects in duplicate genes are due to redundancy among duplicates. In the worm, this implies that all 4515 duplicate genes with no phenotypic effect showed this phenomenon because of functional redundancy, and hence that roughly 36% of robustness is due to buffering from duplication. In sum, the available yeast data suggests that the contribution of duplicate genes to weak phenotypic effects ranges between 23% and 59%, whereas the corresponding range for the worm is 3%-36%. An important caveat to this comparison is that synthetic-null mutations in yeast and RNAi represent fundamentally different approaches to generating phenotypic effects. Moreover, the patterns of duplication in these two organisms have resulted in different functional distributions of duplicate genes (Conant & Wagner 2002).

Despite uncertainties in estimating the relative contribution of gene duplicates to the buffering of null mutations, it is clear that much gene redundancy exists in eukaryotes. Why is this so? At least three possibilities exist. First, gene redundancy may be an accidental by-product of gene duplications, serving no adaptive role. If so, redundancy is just a transient state after gene duplication. Because multiple lines of evidence indicate that sequence and functional divergence after gene duplication is rapid (Gu *et al*. 2002; Lynch & Conery 2000; Wagner 2002), redundancy should then only be observed in recent gene duplicates. This prediction is contradicted by at least two lines of evidence. First, many genomes contain ancient gene duplicates with very similar functions. Examples include the yeast TPK gene family (catalytic subunits of cyclic AMP-dependent protein kinase, Toda *et al*. 1987) and the yeast CLN gene family (cyclins required for the G1-S transition in the cell cycle, Nasmyth 1993). Although synthetic null mutations in member genes of both (well-characterized) families show only

subtle fitness defects (Benton *et al*. 1993; Smith *et al*. 1996), even the youngest duplicate pair within each family is ancient (>100 million years old, Wagner 2001). A second line of evidence is our figure 6, which shows that even highly diverged duplicate genes are more likely to show no phenotypic effect in RNAi than single copy genes. The age of duplicates cannot be reliably estimated from amino acid divergence. However, for a third of the duplicates shown, synonymous sites on DNA have completely diverged (results not shown), demonstrating that these duplicates are ancient. Mutational robustness through gene redundancy is not just a transient phenomenon.

The second possibility is that redundancy is maintained whenever it is advantageous for an organism to produce copious amounts of gene product (Seoighe & Wolfe 1999). Clearly, for duplicate genes to fulfill such a role, they must maintain a high degree of functional similarity. Consistent with this notion is our observation that highly expressed genes are more likely to have a duplicate with high sequence similarity than other genes. This pattern has been previously described for duplicate genes in yeast although it may have other causes (Pál *et al*. 2001). The major difficulty with this argument is that if most redundant gene duplicates are maintained because the genes must be highly expressed then gene duplication can not be responsible for many weak gene knock-out effects, because eliminating one of two duplicates would then have deleterious effects. Indeed, our results show that the loss of highly expressed genes in the worm tends to result in severe phenotypic effects.

The last remaining possibility regards an adaptive role for redundant gene functions. Population genetic modeling (Cooke *et al*. 1997; Nowak *et al*. 1997; Wagner 1999; Wagner 2000c) has shown that gene redundancy can be maintained by natural

selection of genotypes robust against mutations. Such robustness is maintained indirectly, as organisms with redundant genes do not have higher fitness but rather accumulate in populations because they are less susceptible to deleterious mutations. The problem is that the selection pressure is very weak, of the order of the genic mutation rate μ (Wagner 1999; Wagner 2000c). Redundancy can thus only be indefinitely maintained if mutation rates are very high or populations are very large (effective size $N_e>1/\mu$, Hartl & Clark 1997) . However, even in small populations, this evolutionary mechanism can substantially delay the functional divergence of duplicates and the concomitant loss of redundancy (Wagner 2000c). In addition, multifunctional gene duplicates with many pleiotropic interactions can also diverge very slowly in function, even in small populations (Wagner 2000c). *C. elegans*, whose populations consist largely of self-fertilizing hermaphrodites, is likely to have small effective population size. Nevertheless, it shows considerable redundancy in ancient gene duplicates (Figure 6), consistent with a slowing of duplicated gene divergence due to an adaptive role of redundancy.

In sum, the worm genome contains thousands of single-copy genes with absent phenotypic effects. This phenomenon is most likely due to complex interactions in genetic networks that are still incompletely understood. Whether such robustness is an evolved or an intrinsic feature of genetic networks is an open question for future research. Conversely, gene duplications also contribute to numerous cases of genes with absent phenotypic affects. Many of these duplicates are ancient, raising the possibility that the functional divergence of genes may be slowed by selection for mutational robustness.

Chapter 4: Convergent evolution of gene circuits

A version of this chapter has previously appeared as: Conant, G. C. and Wagner, A. (2003) “Convergent evolution of gene circuits, *Nature Genetics*, **34(3)**: 264-266.

Abstract:

Convergent evolution occurs on all levels of biological organization and is a potent indicator of optimal design. We here show that convergent evolution also occurs in genetic networks, where it had previously not been demonstrated. Specifically, we show that multiple types of transcriptional regulation circuitry in *Escherichia coli* and the yeast *Saccharomyces cerevisiae* have evolved independently and not by duplication of one or a few ancestral circuits.

Introduction

Repeated appearances of a complex structure in the history of life are *a priori* sufficiently unlikely as to require an explanation (Futuyma 1998). Cases where a structure reoccurs in independent evolutionary lineages, as do the wings of birds, bats, and pterodactyls, are examples of convergent evolution: the reappearance of a structure due to selective pressures that favor similar solutions to similar problems.

If, on the other hand, such a structure reoccurs in the same organism, there is a second possible explanation: duplication. Families of gene duplicates, such as the human globin family, are the most well-known of such repeated structures. In this case, gene duplication is a more likely source of commonality than is convergent evolution, especially since the similarity between family members extends to parts of the genes not involved in protein function, such as silent sites (Li 1997). Below, we will consider another repeated structure: genetic circuits in the transcriptional regulatory network of two micro-organisms. In the case of genetic circuits, although certain circuits are very

abundant, it is not clear *a priori* whether duplication or convergence is responsible for this abundance.

Convergent evolution is a strong indicator of optimal design. It occurs on all levels of biological organization, from complex organ systems down to the structure of genes and proteins. For instance, eyes of similar basic design may have evolved multiple times independently; the wings of birds and bats have similar architectures, and both fish and whales have similarly streamlined body shapes, although the latter are secondarily descended from land mammals (Futuyma 1998). On the smallest scale, lysozymes have been repeatedly and independently recruited for digestion food in foregut fermenting herbivores including bovids, colubine monkeys such as languars, a bird, and the fruit fly *Drosophila* (Kornegay *et al*. 1994; Regel *et al*. 1998; Stewart *et al*. 1987). Antifreeze glycoproteins in two groups of fish that live in extremely cold environments at opposite ends of the globe, antarctic notothenioids and northern cods, have independently evolved similar amino acid sequences (Chen *et al*. 1997; Fletcher *et al*. 2001). We here demonstrate a case of convergent evolution on the level of genetic networks. Our examples of convergent evolution are particularly interesting because they occur repeatedly in the same organism and cannot be attributed to duplication.

Our work builds on recent studies that have identified small and abundant genetic circuit motifs in transcriptional regulation networks of the yeast *Saccharomyces cerevisiae* (Lee *et al*. 2002; Milo *et al*. 2002) and the bacterium *Escherichia coli* (Milo *et al*. 2002; Shen-Orr *et al*. 2002). These authors represent the transcriptional network as a directed graph (see figure 11). In this graph, nodes are genes and directed edges link transcription factors with the genes they regulate. Circuits, therefore, are clusters of (in

this case) three or more nodes and any edges that connect those nodes. Milo and coauthors (2002) performed a statistical analysis of networks from both yeast and *E. coli*, looking at all possible combinations of three and four nodes . There a limited number of possible three and four node sub-graphs, so these authors where able to count the number of occurrences of each type of subgraph. For instance, there are 13 possible three-node subgraphs, and of those 13, only 1, the feed-forward loop (figure 11), occurred in either yeast or *E. coli* more often than one would expect by chance.

We will consider circuit motifs identified both by Milo and collaborators (2002) as well as by Lee and coauthors (2002). They include regulatory chains, feed-forward circuits, and a "bi-fan" circuit (see figure 12b). The identification of such motifs invites questions about their evolutionary origin, for which there are two possibilities. First, these circuits may have come about through the duplication – and subsequent functional diversification – of one or a few ancestral circuits. Given the high frequency at which single genes and large genome fragments undergo duplication (Bailey *et al*. 2002; Lynch & Conery 2000; Wolfe & Shields 1997), this is a plausible scenario. However, it is equally possible that most of these circuits arose independently by recruitment of unrelated genes. If such convergent circuit evolution is prevalent, then evolved circuit motifs must have favorable functional properties, and natural selection will have played an important role in their creation. If, however, most circuits share a common ancestry, historical accidents may be primarily responsible for their abundance.

Methods

Circuit types. In *E. coli*, we examined feed-forward loops and "bi-fans" (see fig. 12b). To identify these motifs, a purpose-built computer program was used to locate

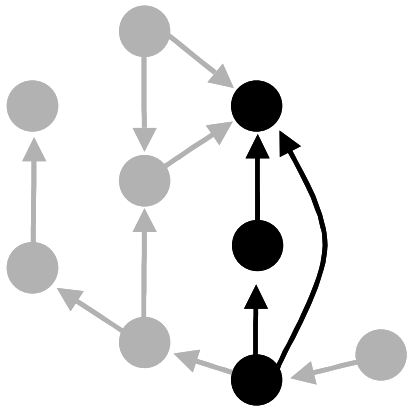


Figure 11: One type of subgraph, a feed-forward circuit (shown in black), embedded in a larger network graph. Circles are nodes (genes), while arrows (edges) indicate regulatory interactions

three and four node motifs within the full transcriptional regulatory network (released by Shen-Orr and collaborators 2002). Before analysis, we removed any regulatory elements representing an operon rather than a single gene product from the *E. coli* data. A third kind of *E. coli* circuit, the dense-overlapping regulon described by Shen-Orr and collaborators (2002), does not have uniform topology, and is thus not suitable for our approach.

In *Saccharomyces cerevisiae* we considered six classes of circuits, five of which were described by Lee *et al.* (2002). These are autoregulation loops (one regulator influencing its own transcription), multi-component loops (a chain of regulators forming a closed regulatory loop), single input motifs (a regulator with multiple target genes), feed-forward loops (but see below), multi-input motifs, and regulatory chains (see fig. 12b for the last three types of circuits).

Although the feed-forward loop was identified by Lee and collaborators, the definition of the motif used differed slightly from that of Shen-Orr and collaborators (2002). In particular, Lee and coauthors occasionally included as an intermediate regulator a gene that actually regulated the master regulator of the feed-forward loop. For the sake of consistency and to allow us to analyze motifs at varying levels of

microarray stringency, we used the feed-forward loop definition of Shen-Orr and collaborators rather than that of Lee and collaborators. Thus, in yeast, we identified feed-forward loops and bi-fans (a circuit not considered by Lee and collaborators: see description of *E. coli* circuits) using Lee and coauthors' data but locating the circuits with the computer program used above in the *E. coli* analysis.

Of the six circuit types available in yeast, autoregulation and single input motifs were inappropriate for our analysis as they contain only a single regulatory gene. The multi-component loop was also unsuitable because there are only 3 circuits of this type, containing only three distinct genes. Analysis of multi-input motifs and regulatory chains is complicated because these circuits contain variable numbers of regulatory genes. We analyzed each different size of these circuit types separately (a total of 16 analyses).

We use two indicators of common gene circuit ancestry. To conceptualize these indicators, consider a genome with n regulatory circuits of identical topology, each comprising k genes. We call any pair of circuits related by common descent if all k gene pairs in the circuit pair are gene duplicates. Our first indicator, an index A of common circuit ancestry, is best understood in the context of a graph whose n nodes represent the n circuits (fig. 12a). Two nodes are connected in this graph if they are identical by descent. There are two extreme possibilities. First, none of the n circuits might share common ancestry. In this case, the graph consists of n disconnected vertices (A=0). Second, all circuits may share common ancestry, in which case the graph is fully connected (A≈1). Between the extremes lies a spectrum of possibilities, where the graph has C connected components ($1<C<n$), each of which corresponds to one family of circuits that derive from a single common ancestor. We define the index of common

ancestry A simply as A=1-(C/n). The greater A, the greater the fraction of circuits sharing a common ancestor. Our second indicator of common ancestry is simply the size of the largest family of circuits with common ancestry (F_{max}). In terms of the circuit graph, F_{max} is the size of the graph's largest component.

In analyzing all these circuits, we considered only the regulatory genes in the circuits and not their downstream targets (Including target genes would result in even fewer circuits showing common ancestry). We identified duplicated genes using gapped BLAST (Altschul *et al*. 1997) at a threshold value of $E \leq 10^{-5}$. Varying E from 10^{-3} to 10^{-11} did not change the conclusions reached. There are more rigorous methods to identify duplicate genes. For example, one can call two genes duplicates only if their alignable regions exceed a certain length, if there is a high ratio of mismatches to gaps, and if a minimum percentage of nucleotides match (e.g., Conant & Wagner 2002, appendix). Any such criterion serves to eliminate false positive duplicates. However, we did not pursue a more stringent approach because false positive duplicates disfavor our hypothesis of independent circuit origin. That is, by using a liberal assay for identifying gene duplicates, the number of duplicate circuits may appear larger than it is. That our hypothesis holds up in spite of this methodical bias against it speaks in its favor.

When evaluating instances of gene circuit duplication it is important to distinguish circuit duplication from simple gene duplication. We thus evaluated the probability that two circuits appear to be duplicates of each other merely because they both happen to contain duplicated gene pairs. For any circuit which showed potential for common circuit origin (*i.e.* A>0), we created a distribution of 1000 randomized circuit graphs, each formed by substituting randomly chosen genes for each gene in the original

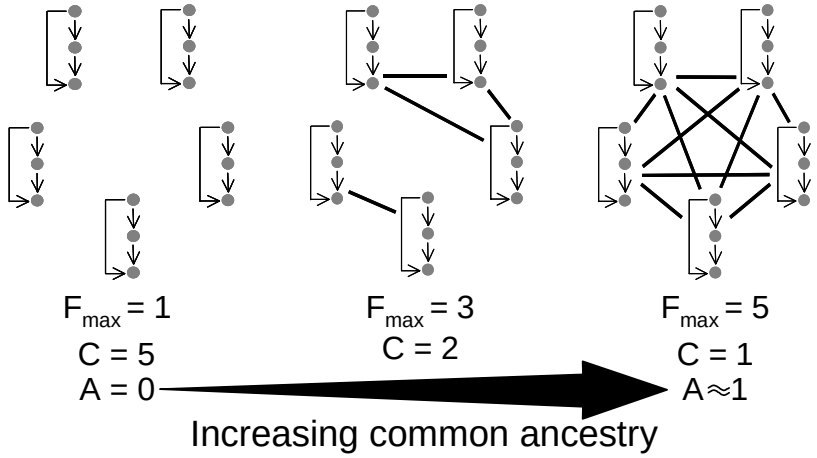
graph. These new random graphs contain the same number of circuits and genes as the original graph, except that two circuits (nodes) are connected if the randomly chosen genes in them are paralogs of each other (selected at $E \leq E_{crit}$). Naively, one might think that the random genes should be drawn from the genome as a whole. However, because only 112 regulator genes could potentially occur in these circuits under the experimental design of Lee and collaborators (Lee *et al*. 2002), the most appropriate pool of genes to draw from is these regulatory genes. Calculating A and F_{max} for each resampled graph yields a distribution of their values. We then asked whether the observed values of A and F_{max} lie in the tails of this distribution.

Finally, we assessed whether genes which are duplicates of each other are more likely to occur in the same type of circuit. There are 112 regulators in yeast, but not every one occurs in each circuit type. We define P_{motif} as the probability that a randomly-chosen regulator will occur in a particular circuit type (for example a bifan, see table 1). We then calculated $\mathbf{P}_{motif|duplicate}$, the probability of a regulator occurring in that circuit type, given that at least one of its duplicates does. If many gene circuits originated through gene duplication, members of one gene family will be more likely to co-occur in a circuit type than genes at large ($\mathbf{P}_{motif} < \mathbf{P}_{motif|duplicate}$). To evaluate the statistical significance of observed values for $\mathbf{P}_{motif|duplicate}$, we tested the hypothesis $\mathbf{P}_{motif} = \mathbf{P}_{motif|duplicate}$ with an (exact) one-sided binomial test (table 1) for all circuit types where $\mathbf{P}_{motif} < \mathbf{P}_{motif|duplicate}$.

Results

Neither of two abundant circuit types in *E. coli* show evidence of common ancestry, that is, A=0 and F_{max}=1 for both types of circuits (Fig. 12b). To identify

A)



B)

	Circuit Type	Number of Circuits	Number of Families (C)	Index of common ancestry (A)	Largest Circuit Family (F_{max})
Yeast	Feed-forward	48	44 (46.8 ± 1.9; **P** = 0.08)	0.082 (0.023 ± 0.035; **P** = 0.08)	5 (1.9 ± 1.4; **P** = 0.05)
	Bi-fan	542	435 (469.0 ± 37.7; **P** = 0.18)	0.197 (0.135 ± 0.070; **P** = 0.18)	49 (41.0 ± 31.1; **P** = 0.33)
	MIM-2	176	168 (164.5 ± 8.8; **P** = 0.60)	0.045 (0.065 ± 0.050; **P** = 0.60)	5 (7.4 ± 6.2; **P** = 0.59)
	Reg. Chain (3)	33	33	0	1
E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

Figure 12: Circuit duplication is rare in yeast and *E. coli*. **A)** Two indicators of common ancestry for gene circuits. Each of *n* (=5) circuits of a given type (a feed-forward loop for illustration) is represented as a node in a 'circuit graph'. Nodes are connected if they are derived from a common ancestor, that is, if all *k* pairs of genes in the two circuits are duplicate genes. A=0 if no circuits share a common ancestor (the graph has *n* isolated vertices), A≈1 if all circuits share one common ancestor (the graph is fully connected). The number C of connected components indicates the number of common ancestors (two in the middle panel) from which the n circuits derive. F_{max} is the size of the largest family of circuits with a single common ancestor (the graph's largest component). **B)** Little common ancestry in six circuit types. We considered two circuits as related by common ancestry if each pair of genes at corresponding positions in the circuit had significant sequence similarity (see text). Each row of the table shows values of C, A, and F_{max} for a given circuit type, followed in parentheses by their average values ± standard deviations and P-values (see methods).

duplicated genes and circuits, we used gapped BLAST (Altschul *et al*. 1997) at a

deliberately liberal threshold value of $E \leq 10^{-5}$ (Varying E from 10^{-3} to 10^{-11} did not change the conclusions reached; results not shown). Among the 18 abundant yeast circuit types we studied (see methods), only three, the feed-forward loops, multi-input modules of size 2, and bi-fans showed evidence of common ancestry ($A>0$ and $F_{max}>1$; Fig. 12b). This, however, might be due to chance alone, simply because genomes contain many duplicate genes. We thus applied the above permutation test to determine the significance of the observed values.

None of the observed values of A were significantly different from values expected by chance alone (see methods). For example, yeast contains 542 bi-fan motifs that show an index of common ancestry A=0.197. The mean of A for 1000 randomly resampled circuit graphs was A=0.135 (standard deviation 0.070). The probability of observing the actual value of A=0.197 by chance alone is **P**=0.18: not small enough to reject the null hypothesis. A similar pattern holds for the maximum circuit family size F_{max}=49 for this circuit. In comparison to the randomly resampled circuit graphs with an average F_{max} of 41 (s.d. 31.1), there is a probability of **P**=0.33 of obtaining F_{max}=49 by chance alone. We only observed a marginally significant value of F_{max} =5 for feed-forward loops (**P**=0.05). The corresponding circuit graph has a single large component of 5 circuits, and 43 single circuit components. That is, 43 of the 48 instances of this circuit type have no common ancestor. The remaining five circuits contain only 5 different genes (ABF1, MBP1, MOT3, SWI4, and SWI6), that is, they share multiple genes. The repeated presence of the highly duplicated SWI4 gene in these circuits is responsible for the single large component.

Table 1: Gene families are not over-represented in circuit types

Organism	Circuit Type	P_{motif} [a]	$\mathbf{P_{motif|duplicate}}$ [b]	P [c]
S. cerevisiae	Bi-fan	**0.82**	**0.80**	**NA**
	Feed-forward	**0.38**	**0.42**	**0.21**
	Multi-input motif	**0.77**	**0.76**	**NA**
	Regulator Chains	**0.64**	**0.67**	**0.30**
E. coli	**Bi-fan**	**0.50**	**0.67**	**0.11**
	Feed-forward	**0.82**	**0.67**	**NA**

[a]: Probability of a randomly-chosen regulatory gene occurring in a given circuit type.
[b]: Probability of a regulatory gene occurring in a circuit type given that one of its duplicates occurs in that circuit type (see Methods)
[c]: P-value for one-sided exact binomial test of the null-hypothesis $P_{motif} = P_{motif\ |\ duplicate}$. 'NA' indicates that a test has not been carried because $P_{motif} > P_{motif\ |\ duplicate}$. The number of transcriptional regulators was n=112 and n=22 for the yeast and *E. coli* analyses, respectively.

The study that first described the structure of the yeast transcriptional regulation network relies on genome-scale chromatin precipitation experiments that use a statistical error model to distinguish true from spurious regulatory interactions (Lee *et al*. 2002). In our analysis thus far, we have used the error threshold $\mathbf{P_e}=10^{-3}$ employed by the authors. In order to gain additional confidence in our conclusions, we repeated our analysis with error thresholds ranging from a very liberal $\mathbf{P_e}=10^{-2}$ to a conservative $\mathbf{P_e}=10^{-5}$ for the yeast bi-fan and feed-forward circuits. For the bi-fans, no evidence of common circuit ancestry emerged with varying error threshold. That is, both A and F_{max} were no different than expected by chance alone (results not shown). For feed-forward loops, we observed a marginally significant value of A=0.11 (**P**=0.03) and F_{max} =3 (**P**=0.03) at $\mathbf{P_e}\leq10^{-4}$. However, lowering $\mathbf{P_e}$ further to $\mathbf{P_e}=10^{-5}$ yielded both A=0 and F_{max}=1.

In a complementary analysis, we asked whether members of one gene family preferentially occur in one type of gene circuit. This is the pattern expected if many gene circuits originated through gene duplication. Specifically, we asked whether the likelihood that a gene occurs in a circuit of a given type increases if one of its duplicates occurs in a circuit of this type. Table 1 shows that none of the circuit types we examined

showed this biased pattern of occurrence. In addition, we examined whether the genes within any one circuit type preferentially arose through gene duplication but found no such pattern (results not shown). Gene duplicates seem to be distributed randomly across circuit types, with no discernable regularities.

Discussion: Convergent Circuit Evolution

Our analysis of 2 circuit types in *E. coli* did not find any hints of common ancestry. The same was the case for the yeast regulatory chains (of any length), or the yeast multi-input motifs with more than two regulators. Of the remaining three yeast circuit types studied, two show common ancestry indistinguishable from that expected by chance alone. Only feed-forward loops show marginally significant values of either A or F_{max}, but this finding is not robust to changes in the error threshold $\mathbf{P_e}$. In addition, the vast majority of feed-forward loops have clearly independent evolutionary origins. Among 48 circuits of this type, 43 are composed of unrelated genes (Figure 12b). We also note that the probability of falsely identifying a pair of circuits as duplicates of each other scales as p^n, where p is the probability of randomly choosing a pair of genes that are duplicates and n is the circuit size. In other words, the larger a circuit is, the smaller is the probability of falsely identifying it as a duplicate of another circuit. The larger circuits are exactly the circuits where we see not even non-significant evidence of duplication, that is, where A=0 and F_{max}=0.

Multiple indicators of gene function, including gene expression, promoter structure, and protein interactions, indicate that most duplicate genes diverge rapidly in function (Dermitzakis & Clark 2002; Gasch *et al*. 2000; Gu *et al*. 2002; Wagner 2000a; Wagner 2001). Our finding that small gene circuits do not share common ancestry, and that

duplicate regulatory genes are randomly distributed across gene circuit types also supports this point, because it implies that duplicate transcriptional regulators can readily evolve new regulatory interactions. The short DNA binding sites of transcriptional regulators may account for much of this plasticity. In microbes like yeast and *E. coli*, with huge population sizes, new binding sites for transcriptional regulators can arise by chance alone on very short time scales (Stone & Wray 2001). Results of laboratory evolution experiments in yeast, which can permanently alter the expression of thousands of genes within several hundred generations, testify to this evolutionary plasticity (Ferea *et al*. 1999). Transcriptional regulation circuits are thus ideal to observe instances of convergent evolution, because natural selection has much raw material – variation in regulatory interactions – to shape such circuits. The resulting abundance of convergently evolved circuits also speaks to the longstanding neutralist-selectionist debate (Li 1997; McDonald & Kreitman 1991; Zhang *et al*. 1998) about the role of selection in shaping genes and their products. It suggests that natural selection may be very important in enriching genomes for well-designed circuits.

The finding that the same type of gene circuit has evolved again and again makes a strong case that a circuit type has optimal design features. For example, the design of a feed-forward loop may serve to activate the regulated ('downstream') genes only if the upstream-most regulator is persistently activated. Conversely, it can rapidly deactivate downstream genes once this regulator is shut off (Shen-Orr *et al*. 2002). Our results also suggest that convergent evolution, which may be rare on the level of protein sequences (Doolittle 1994), may play an important role on the higher level of biological organization of gene circuits. Stephen Jay Gould famously asked what would be

conserved if life's tape – its evolutionary history – was rewound and run a second time (1989). Transcriptional regulation circuits would, at least in abundance, come out just about the same.

Chapter 5: Conclusions and Discussion

The three studies described above individually offer insights into distinct problems in molecular evolution which are discussed in their respective chapters. However, taken together, the three studies also provide valuable perspective on some larger questions. In particular, I have explored the relationship of duplication (gene and otherwise) and the appearance of novel features in organisms (functional divergence). The most naïve expectation would be that functional divergence and duplication would be tightly coupled, so that surviving duplicates would almost invariably show some evidence for diversification in function. (This argument does not, however, preclude diversification occurring without duplication). My results, however, indicate that duplication and diversification need not always go hand in hand. For instance, even in chapter 2 where I find that many gene duplicates show asymmetric sequence divergence, it is important to note that the majority of duplicates (~70%) did not show such asymmetry, meaning that I have no sequence-based indications of functional divergence for more than two-thirds of these gene pairs. Of course, both parts of this statement need to be interpreted cautiously: asymmetric divergence alone is insufficient to demonstrate functional divergence. Asymmetry is nonetheless suggestive of functional divergence, particularly because asymmetric divergence of synonymous sites is much rarer (unpublished data). On the other hand, the rate of asymmetric sequence divergence likely understates the overall rate of divergence since evolution occurs not only in coding sequences, but also in expression patterns. As the results above and those of other researchers suggest (Fig. 10; Wagner 2000a; Gu *et al*. 2002), evolution of expression patterns and sequences is generally not highly correlated, so duplicates without

asymmetric evolution in coding sequence may show asymmetries in expression (in fact Wagner 2002 showed such asymmetries to be common).

Assuming that not all duplicates have evolved distinct functions, what is the explanation for the preservation of the remaining duplicate pairs? As discussed in the introduction, selective forces are needed to protect duplicate pairs from the degenerative effects of genetic drift. In addition to the three possibilities already discussed, one other potential means of preservation is suggested by work on RNA interference (see chapter 3). Researchers have found that at least one pair of duplicate genes (*Stellate* and *Suppressor of Stellate*) is preserved in fruit flies so that one member of the paralog pair can down regulate the other by an RNAi-like mechanism (Aravin *et al*. 2001). The importance of this mechanism in maintaining duplicates is not yet clear, and it differs from the three already discussed in that the duplicate need not be preserved in its entirety as long as sufficient sequence similarity remains for the RNAi pathway. Of course, my work in chapter 3 offers further evidence that at least some duplicate genes may be retained to provide mutational robustness. Both of these examples are cases where duplicates are maintained to confer advantages other than those provided by functional divergence.

In chapter 4, I also found cases where novel structures were created through the action of natural selection without intervening duplication. In particular, transcriptional regulatory circuits have repeatedly and independently evolved because of their beneficial attributes. This suggests that transcriptional regulatory networks, at least, are flexible enough to produce new functions without the need to "backup" existing structures through duplication. Researchers are increasingly aware of the importance of various

types of networks in biology (Alon 2003; Bray 2003), be they of metabolites (Fell & Wagner 2000; Jeong *et al*. 2000), interacting proteins (Fraser *et al*. 2002; Hahn *et al*. 2004; Jeong *et al*. 2001) or transcription factors (Lee *et al*. 2002; Milo *et al*. 2002; Shen-Orr *et al*. 2002). One important characteristic of such networks is that changing the way network member genes or proteins interact is generally less difficult than creating new members (for example adding a protein interaction or transcription factor binding site rather than evolving a new protein). As a result, many evolutionary novelties may owe their origins to such changes of interaction, changes that do not involve duplication and may be difficult to detect using traditional sequence analysis.

The diverse fates of duplicate genes and the various ways in which novelty appears also reminds us of the undirected nature of evolution. This property of evolution is most clear in the unexpected connections between structures of different function in organisms. The panda's thumb and birds' flight feathers are connected by descent with other structures (wrist bones and insulating material) to which they bear no obvious functional affinity. There are similar examples at the molecular level, where existing enzymes have been co-opted for new roles. Thus, the ε-crystallin of birds and reptiles is actually transcribed from the same gene locus as is lactate dehydrogenase B_4 (Hendriks *et al*. 1988; Wistow *et al*. 1987), while the lactalbumin protein in mammalian milk is derived by gene duplication from the bacteriolytic enzyme lysozyme (Graur & Li 2000).

The differing fates of gene duplicates are a less obvious but equally relevant example of evolution's lack of direction. Duplicates do not have a predestined function at the time of their creation. Rather, they are raw material which may be co-opted to meet a current need of the organism. If a duplicate by chance undergoes a mutation that

allows it to perform some new function, it may enter a path to diversification. If, on the other hand, the organism is in an environment where redundancy is beneficial, the duplicate may maintain its current function. Duplicates may even be maintained to increase or control gene dosage. The fate of a duplicate gene pair is at least in part a contingent effect of history.

As the above discussion of duplication will suggest, undirected does not mean random. The fate of a duplicate depends on both random factors (which genes actually are duplicated and what mutations they undergo) as well as selective forces (features of the organism which dictate the need for high gene dosage or the usefulness of a new enzymatic function). It is equally important to remember both of these types of factors when considering the evolution of the transcriptional regulatory network. As my analysis shows, the evolution of this network is by no means random: useful circuits have evolved repeatedly in the network. At the same time, the addition of new circuits to the network must be constrained by network structure. As I note, a given transcription factor is generally not over-represented in any one circuit type, suggesting that transcription factors are not forced by their structure into only one part of a circuit (a master regulator of a feed-forward loop, say). Thus, the network structure will owe its origins both to its evolved features and well as to more random effects such as, again, which of its members have recently undergone duplication. Understanding the evolution of novelty within complex existing structures such as regulatory networks will be a major challenge in the future study of evolution, requiring tools that detect both the relics of past selection as well as the signals of historical continuity.

Appendix: GenomeHistory: a software tool and its application to fully sequenced genomes

This appendix has previously appeared in substantially the same form as: Conant, G. C. and Wagner, A. (2002) "GenomeHistory: a software tool and its application to fully sequenced genomes", *Nucleic Acids Research*, **30**: 3378-3386.

Abstract

We present a publicly available software tool (http://www.unm.edu/~compbio/software/GenomeHistory) that identifies all pairs of duplicate genes in a genome and then determines the degree of synonymous and non-synonymous divergence between each duplicate pair. Using this tool, we analyze the relations between (i) gene function and the propensity of a gene to duplicate, and (ii) the number of genes in a gene family and the family's rate of sequence evolution. We do so for the complete genomes of four eukaryotes (fission and budding yeast, fruit fly, nematode) and one prokaryote (*Escherichia coli*). For some classes of genes we observe a strong relationship between gene function and a gene's propensity to undergo duplication. Most notably, ribosomal genes and transcription factors appear less likely to undergo gene duplication than other genes. In both fission and budding yeast, we see a strong positive correlation between the selective constraint on a gene and the size of the gene family of which this gene is a member. In contrast, a weakly negative such correlation is seen in multicellular eukaryotes.

Introduction

That gene duplication is a major force in genome evolution was first pointed out forcefully in Ohno's pioneering book (Ohno 1970). Since then, considerable progress has been made in determining how gene duplicates evolve and what role they play in organismal evolution (Ferris & Whitt 1979; Force *et al*. 1999; Iwabe *et al*. 1996; Li 1980; Lundin 1999; Nadeau & Sankoff 1997; Nei & Roychoudhury 1973). The availability of complete genome sequences has not only made it clear that genomes are replete with

duplicate genes, but has also spawned new and varied avenues of research. These include studies of the fate of gene duplicates produced in a genome duplication (Wolfe & Shields 1997) and of the production and distribution of pseudogenes (Harrison *et al*. 2001; Seoighe & Wolfe 1998). Further research has focused on estimates of the rate at which gene duplications occur (Lynch & Conery 2000) and on the distribution of gene family sizes in genomes (Gerstein 1997; Huynen & van Nimwegen 1998; Qian *et al*. 2001), which was found to obey a power-law.

Through this report and through an accompanying web site (http://www.unm.edu/~compbio/software/GenomeHistory), we make public a flexible and portable tool that allows one to extract the number of non-synonymous nucleotide substitutions per nucleotide site (K_a) and the number of synonymous nucleotide substitutions per nucleotide site (K_s) for all gene duplicates in a genome from information on coding regions contained in FASTA files. With suitable precautions, K_s can be used to estimate the time that elapsed since a gene duplication. The ratio K_a/K_s is an enormously useful quantity in gauging the selective constraint a given sequence pair is subject to (Li 1997). We have named our tool GENOMEHISTORY. It relies on existing algorithms, but uses user-configurable parameters to automate the analysis of large datasets with minimal user input.

Below, we use GENOMEHISTORY to examine patterns of gene duplication in five fully sequenced genomes. Several genome sequencing consortia have begun this task in their original reports published with the genome sequences (Rubin *et al*. 2000). Extending this and other work (Kondrashov *et al*. 2002; Lynch & Conery 2000), we here address three questions: (i) Do genes of different functions differ in their propensity to

undergo duplication? (ii) Do selective constraints differ among duplicate genes with different functions? (iii) Does the selective pressure acting on a gene depend on the number of its duplicates?

Materials and Methods

Sequence Analysis: GENOMEHISTORY pre-screens a genome for similar amino acid sequences using gapped BLASTP (Altschul *et al*. 1997), then carries out a global alignment of promising candidates using CLUSTAL (Thompson *et al*. 1994), and subsequently estimates K_a and K_s, the number of non-synonymous and synonymous mutations per non-synonymous and synonymous site on DNA, respectively (Li 1997). We analyzed five genomes with GENOMEHISTORY: those of the yeasts *Saccharomyces cerevisiae* (Goffeau *et al*. 1996) and *Schizosaccharomyces pombe* (Wood *et al*. 2002), the fruit fly *Drosophila melanogaster* (Adams *et al*. 2000), the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998) and the bacterium *Escherichia coli* (Blattner *et al*. 1997). For each genome, we obtained the complete set of protein sequences and corresponding nucleotide sequences from sources listed in the above references. We considered protein pairs for further analysis if their similarity was greater than indicated by the following BLAST E-value thresholds. $E<10^{-8}$ (yeasts), $E<10^{-10}$ (*Drosophila*), $E<10^{-10}$ (*C. elegans*) and $E< 10^{-7}$ (*E. coli*). The differences in E-value thresholds reflect a correction accounting for varying numbers of pairwise comparisons due to different genome sizes. After globally aligning candidate duplicates, we retained all gene pairs with more than 40% amino acid similarity over the entire alignment. In addition, we required at least 100 aligned amino acid residues for *S. cerevisiae*, *S. pombe*, *Drosophila*, and *C. elegans*, and 70 aligned residues for *E. coli*.

For each of the retained gene pairs, we calculated K_a and K_s. This calculation is performed in GENOMEHISTORY by maximum likelihood estimation using our own implementation of the codon-based models of sequence evolution proposed by Muse and Gaut as well as Goldman and Yang (1994; 1994). The computation is often referred to as Yang and Nielsen's method (2000). Our routine produces results very similar to Yang and Nielsen's implementation of the model in the PAML package. The likelihood maximization is performed using two different computational methods: Powell's routine (Press *et al*. 1992) to find the ratio K_a/K_s as well as the transition/transversion ratio, and Yang's method (2000) to find branch lengths. The latter uses a modified Newton's method (Press *et al*. 1992).

To increase the proportion of true duplicates in our analysis, we report results only for gene pairs where $K_a<0.75$. In our analysis of evolutionary rates, we further restrict ourselves to duplicates with $K_s<3$ (in addition to $K_a<0.75$) and $K_a/K_s<1$. In addition, we excluded all pairs with $K_a< 10^{-4}$ or $K_s< 10^{-4}$. (Such pairs had either no non-synonymous or no synonymous substitutions).

Because of their potentially unusual pattern of sequence evolution, we also wished to highlight and exclude transposon-related genes from our analysis. In *E. coli*, this is easily done because such genes carry a distinct annotation. In *S. cerevisiae* we screened for transposon-related genes (Fig. 15) by using BLASTP to identify all genes similar at $E< 10^{-17}$ to reverse transcriptase (GenBank protein sequence ID AAA91746.1) or the GAG/POL family (based on similarity to gene YFL002W-B). For *S. pombe*, we used Genbank gene descriptions to filter transposon-related genes. In *C. elegans*, we used similarity to the sequence with GenBank sequence ID NP_502686.1 as the criterion.

(In this case, we excluded only genes with BLASTP E< 10^{-77}, because lowering this threshold led to inclusion of genes with other annotations.). Available *Drosophila* genes are already filtered for transposons—only one annotation indicated transposase activity, and there were no large (>20 member) gene families related to transposable elements, as in other organisms. We used the list of *Drosophila* transposons from http://flybase.bio.indiana.edu/transposons/lk/melanogaster-transposon.html as a final filter, which removed only a single gene pair.

Annotations: For genome-scale analyses, manual assignment of genes to functional categories based on their annotations is possible in principal (Chervitz *et al*. 1998), but prohibitive in cost. We thus took to an automated approach. To study the distribution of gene duplicates in different functional categories, we obtained annotations for the yeasts, fruit fly, and nematode genomes from the Gene Ontology (GO) database (The GeneOntology Consortium 2000) (*http://www.geneontology.org/*). The GO database is divided into three high-level annotation groups: Cellular Component, Biological Process, and Molecular Function. We selected 10 functional categories from different levels of the GO hierarchy, mainly from the "Biological Process" annotation group (Ribosomal proteins and transcription factors were identified from the Molecular Function group, and the cytoskeletal genes from the Cellular Component group). We therefore find it helpful to view these annotations as primarily "pathway-based", as opposed to the more biochemical "Molecular Function" annotations.

To prevent single genes from falling into multiple categories, we used an exclusion scheme, whereby genes assigned to specific categories (such as transcription factors and ribosomal proteins) were excluded from more general categories (such as metabolism).

Although requiring genes to fall only into a single pathway does not always match the more complex realities of gene function, we impose this requirement for two reasons. Firstly, we chose annotations at a high enough level that most genes would be seen as fitting best into a single category. For instance, although some actin genes can be placed into the cell cycle category due to their role in cytokinesis, they fit better into the cytoskeletal category. Secondly, allowing genes to occur in more than one category can result in the artefact of observing that different functional classes of genes show different propensity to undergo duplication, when these differences are due to a single underlying cause.. For instance, genes encoding transcription factors are less likely to have multiple duplicates than other genes. Including transcription factors in the "cell cycle" category would then falsely indicate that all genes important for the cell cycle also have a reduced propensity to duplicate.

Instead of allowing genes to occur in multiple categories, we have used the "Molecular Function" annotations in the GeneOntology database to ask whether genes with multiple molecular functions differ from those with a single molecular function in their propensity to duplicate. Using the 34 top-level "Molecular Function" annotations, we divided the genes of the four eukaryotes into two categories: those with a single top-level function annotation and those with more than one such annotation. While this approach has many obvious imperfections, it serves as an automatable first approximation to address the above question. We then divided all duplicate genes into those with a single duplicate and those with more than one duplicate. For each of these two groups we determined the proportion of genes that had only one functional annotation. Although the multiply-duplicated *Drosophila* and *C. elegans* genes were more likely to have single annotations

than expected by chance (P<0.01), the difference was small (*Drosophila*: fraction of genes with single functional annotations: all genes/ multiply-duplicated genes=0.71/0.75; *C. elegans*: fraction of genes with single functional annotations: all genes/multiply-duplicated genes=0.66/0.70). No such difference was observed for multiply duplicated genes in the other genomes, or for any singly duplicated genes (results not shown). This suggests that our strategy of restricting each gene to be in only one functional pathway did not substantially bias our results.

For *S. pombe*, transcription factors and ribosomal proteins were not specifically annotated in GO. We therefore used the GenBank gene description tables (*http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/euk.html*) to identify these genes. Genes were not specifically annotated as cytoskeletal elements in either set of annotations used for this organism and this annotation category is thus not included in our analysis of *S. pombe*.

The K12 strain of *E. coli* is not included in the GO database. We thus obtained annotations from the University of Wisconsin website (http://www.genome.wisc.edu/) and slightly modified the 23 categories used by the sequencing center to yield 19 functional categories (Fig. 13e).

Availability, implementation, and validation of GenomeHistory

GENOMEHISTORY is available from our website, (http://www.unm.edu/~compbio/software/GenomeHistory) and includes HTML documentation (also available on-line at http://www.unm.edu/~compbio/software/GenomeHistory/GenomeHistory.html). The tool

was developed under RedHat Linux 7.1 (kernel version 2.4 and compiler 2.96). Although we have no reason to expect difficulties on other UNIX platforms, we cannot guarantee that our code will work on untested platforms. However, we expect portability to other operating systems, as long as they support perl and stand-alone BLAST. To facilitate modification of the tool by those wishing to overcome platform incompatibilities, we also make public the source code of the routines estimating divergence.

We have compared data obtained with GENOMEHISTORY to data from published work and found the results qualitatively identical. For example, we calculated a 'survivorship' curve of youngest duplicates in *S. cerevisiae* and compared the results to those of Lynch and Conery (2000). The rate of duplication loss was statistically identical (exponential decay coefficient d=7.5 for Lynch and Conery vs. 7.23 for GENOMEHISTORY).

To analyze the *S. cerevisiae* genome's approximately 6000 genes at a BLASTP E-value of $1x10^{-6}$, a dual 800Mhz Pentium system (Redhat Linux 7.1) needs approximately 17.5 hours. BLAST is able to use multiple processors, so this time would be somewhat longer on an equivalent single-processor machine. Which step in the analysis is most time-consuming depends on the BLAST threshold selected: if this threshold is very stringent ($E<10^{-15}$), the maximum likelihood estimations in step 3 dominate, but for more permissive thresholds, the pairwise sequence alignments by CLUSTALW (step 2) dominate.

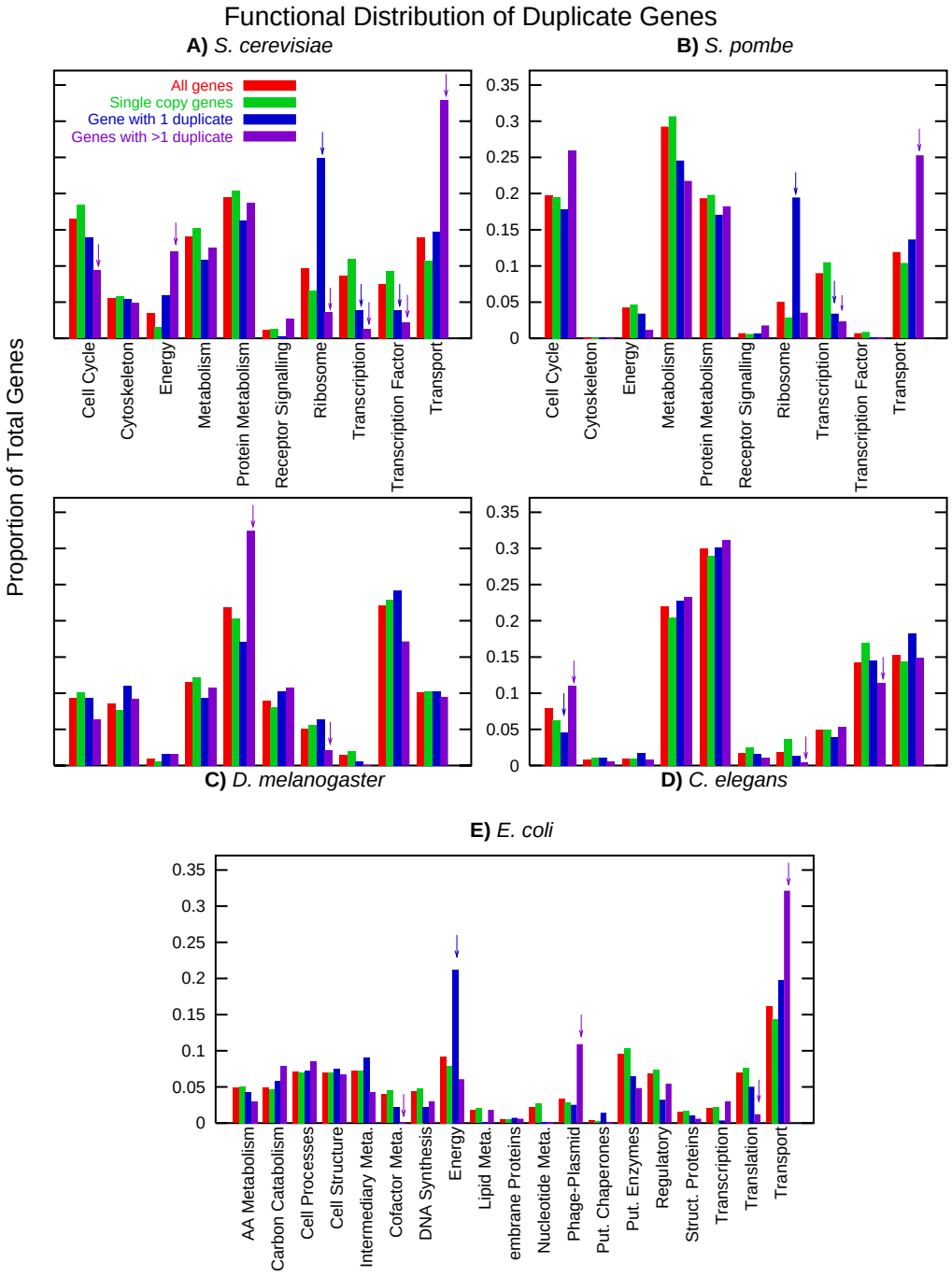


Figure 13: Distribution of genes among functional categories for five organisms. Genes were divided into three groups: single-copy genes, genes with one duplicate and genes with more than one (multiple) duplicates. Proportions significantly different from the overall distribution at a Bonferroni significance level of 0.05 are marked with arrows. **A:** *S. cerevisiae* (2077 total genes).; **B:** *S. pombe* (2298 total genes); **C**: *Drosophila* (2181 total genes); **D:** *C. elegans* (3417 total genes); **E:** *E. coli* (2609 total genes)

The input to GENOMEHISTORY consists of two files in FASTA format, one containing all protein sequences to be analyzed, and the other the nucleotide sequences

corresponding to these proteins. GENOMEHISTORY produces an output file (in tab-delimited text format) that contains K_s and K_a estimates for each sequence pair that meets the analysis criteria. GENOMEHISTORY also generates an error file logging unexpected results that inevitably occur when comparing millions of gene pairs.

To allow testing of a GENOMEHISTORY installation, the GENOMEHISTORY website includes a small test dataset containing the first few dozen genes of the *S. cerevisiae* nuclear genome, as well as sample output from our installation.

Results

What does GenomeHistory do? Comparing all gene pairs in a genome requires considerable computational effort. To eliminate obviously unrelated genes rapidly and to restrict computationally costly divergence estimates only to similar genes, our tool analyses genomes in three distinct stages. **(1)** Identification of potentially interesting gene pairs using the BLAST sequence similarity search algorithm (Altschul *et al*. 1990); **(2)** Alignment of the pairs identified in **(1)** using an exact alignment program (CLUSTAL-W, Thompson *et al*. 1994); **(3)** Calculation of the K_s and K_a values for those aligned sequences whose pairwise similarity is above a user-specified threshold.

For the first step, BLAST analysis, GENOMEHISTORY uses the Washington University implementation of gapped BLASTP (available from http://blast.wustl.edu/) for an initial comparison of protein sequences provided in a FASTA file. BLASTP compares sequences very quickly, allowing us to eliminate highly dissimilar gene pairs rapidly. This reduces the number of further comparisons to a manageable value. Through the BLASTP E-value (Altschul *et al*. 1990; Altschul *et al*. 1997) we allow the user to tune the similarity threshold below which gene pairs are eliminated. We suggest a relatively

liberal threshold choice, such as $E>1x10^{-7}$, deferring the stringent removal of sequence pairs to step two.

In this second step, any two protein sequences deemed promising by the BLAST analysis are aligned using CLUSTAL-W (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/) and the default BLOSUM 62 matrix. Since only pairs of sequences are compared, each alignment will be computationally exact. Using these alignments, sequence pairs go through an additional step of screening before K_s and K_a are estimated. They must have (i) sequence identity in a minimal, user-specified number of residues, (ii) a minimal user-specified length for each sequence, and (iii) a minimal user-specified number of residues aligned at non-gap positions. This final criterion is required because it is possible to align even two long sequences such that each sequence has very few residues aligned with non-gap residues in the other sequence.

In the third step, GENOMEHISTORY calculates a nucleotide alignment corresponding to the obtained protein alignment for the sequence pairs left after steps 1 and 2. The required DNA sequence information is obtained from a sequence file containing nucleotide sequences for all analyzed genes in FASTA format. This alignment is then used to calculate K_s and K_a via a computationally costly but unbiased maximum likelihood algorithm.

Distribution of duplicates by function. The most basic questions about the distribution of gene duplicates with respect to gene functions are these: Are genes with one duplicate over-represented or underrepresented in any of the major functional annotation categories? Does the same hold for genes with multiple duplicates? The simplest and crudest way to address these questions is via χ^2 goodness-of-fit tests to

evaluate the null-hypothesis that the proportion of genes with single (multiple) duplicates in different functional categories is identical to the overall number of genes in these categories. Except for genes with single duplicates in *Drosophila* (**P**=0.040) and in *C. elegans* (**P**=0.133), this null-hypothesis must be rejected at **P**<0.01 for singly and multiply-duplicated genes in all genomes studied. Genes in different functional categories are thus not equally likely to undergo duplication. We now analyze the observed patterns of deviation in detail.

To determine which functional categories had an over- or under abundance of duplicates, we applied a two-tailed binomial ("exact") test. To perform this test, we first calculated the number n_i and fraction p_i of all annotated genes that fell into each functional category *i*. For each *i*, we then tested the null-hypothesis that the observed number of (singly or multiply) duplicated genes in functional category *i* follows a binomial distribution with the same parameter p_i. For the yeasts, fruit fly, and nematode, the analysis involved making 10 hypothesis tests (1 per category). *E. coli* has 19 functional categories making 19 such tests necessary. We used a Bonferroni correction (Sokal & Rohlf 1995) to ensure an overall type I error rate (false rejection of the null hypothesis) of 5%. Proportions significantly different from the overall distribution are marked with arrows on figure 13.

In the yeasts, fruit fly and nematode, the most conspicuous patterns regard ribosomal protein genes. (The *E. coli* genome is not annotated in a directly comparable way). Ribosomal genes with multiple duplicates are underrepresented (**P**<0.0028) in all but the *S. pombe* genome (**P**=0.24). We speculate that this general pattern is due to the high expression level of these genes, and the resulting strong deleterious effects of

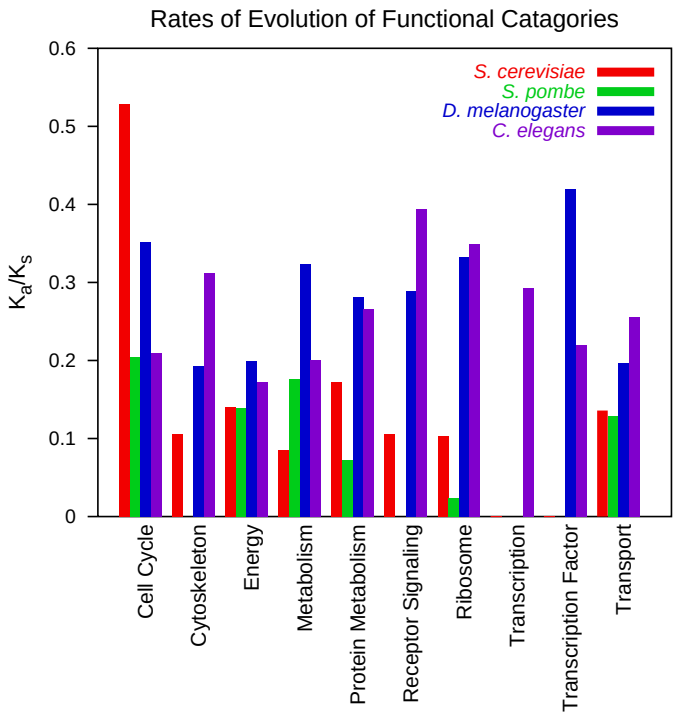


Figure 14: Average K_a/K_s for genes in different functional categories for *S. cerevisiae*, *S. pombe*, *Drosophila*, and *C. elegans*. Blanks indicate cases where no duplicates met the selection criteria ($K_s<3$, $K_a<0.75$, $K_a/K_s<1$)

changes in gene dosage. In contrast to this pattern, ribosomal protein genes with one duplicate are over-represented in both yeasts. For *S. cerevisiae*, this observation, which has also been reported by Planta and Mager (1998), is probably due to an ancient genome duplication that occurred approximately 100 Myrs ago (Wolfe & Shields 1997). Gene dosage effects may have prevented the elimination of these duplicates from the budding yeast genome. Because the common ancestor of budding and fission yeast probably lived before the *S. cerevisiae* genome duplication (Wood *et al*. 2002), it is unlikely that overrepresentation of ribosomal duplicates in fission yeast reflects the same genome duplication. However, it is tempting to speculate that fission yeast has undergone its own genome duplication.

Energy metabolism (in *S. cerevisiae* and *E. coli*) and transport genes (in both yeasts and *E. coli)* show markedly higher proportions of duplicates, which may reflect an historical imprint of the chemically diverse environments these microbes encountered in their evolutionary history. In *S. cerevisiae*, the presence of a large gene family of 17 annotated hexose transporters partly accounts for the expansion of transport-related

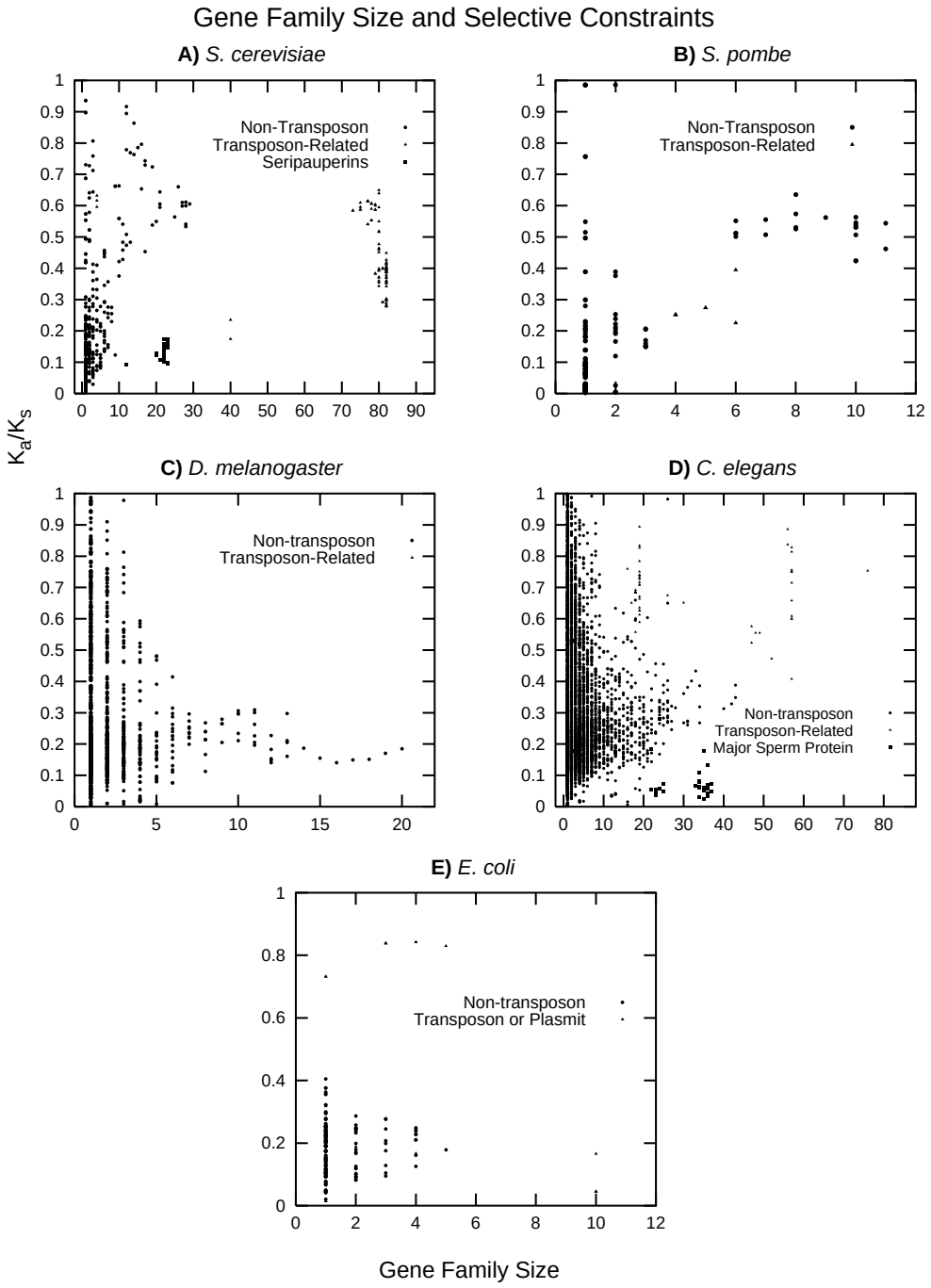


Figure 15: Statistical association between the number of members of a gene family and selective constraints on sequence evolution, as indicated by the ratio K_a/K_s averaged over all family members. **A:** *S. cerevisiae*—Seripauperin genes are highlighted based on their sequence similarity (BLASTP E< 10^{-17}) to ORF YJL223C*;* **B:** *S. pombe;* **C:** *Drosophila;* **D:** *C. elegans*—major sperm family proteins highlighted based on similarity to gene MSP-36 (C04G2.4) (BLASTP E < 10^{-6})*;* **E:** *E. coli;*

genes. Yeast grown in a glucose-limited laboratory environment can undergo multiple duplications of hexose transporters in as few as 450 generations (Brown *et al*. 1998). This

raises the question whether the observed duplicates in the yeast genome are due to the long history of cultivating yeast in the laboratory under similar conditions. This seems unlikely, however, because only 6 of these transporters seem to have been duplicated within the last 10Myrs (K_s<0.11, Wagner 2001).

Several patterns of duplication are specific to only one of the taxa we analyzed. The largest deviation from expected frequencies of duplicates in *Drosophila* is the over-abundance of protein metabolism genes with many duplicates. 28 of the 64 genes in this group - sufficient to explain the deviation - have kinase activity. The presence of many duplicated protein kinases in *Drosophila* and other metazoans has been previously described by other authors (Chervitz *et al*. 1998; Rubin *et al*. 2000; Suga *et al*. 1999).

C. elegans shows an overabundance of proteins with multiple duplicates annotated as cell-cycle proteins. This appears to be the result of numerous duplicates of histone genes (Roberts *et al*. 1987). For instance, there are more than 20 gene duplicates with strong similarity to histone H3 in *C. elegans*, but only two in *Drosophila*, three in *S. cerevisiae*, and five in *S. pombe*.

Do genes with different functions show different evolutionary constraints? To address this question, we determined the average ratios of K_a/K_s for all duplicates in an annotation class, and assessed significant differences via a one-sample t-test. In neither *E. coli*, *C. elegans*, nor *Drosophila* did any functional categories evolve at rates significantly different from the average. In *S. cerevisiae*, the metabolism genes showed significantly slower evolution (**P**=0.003), while in *S. pombe* the ribosomal protein genes evolved significantly more slowly (**P**=0.0006). This paucity of significant results is unsurprising when one considers the high levels of variance in K_a/K_s within categories.

Most variation in K_a/K_s occurs within categories, not among them. Interestingly, the average K_a/K_s ratio in *Drosophila* and *C. elegans* duplicates is higher in almost all categories than in the yeasts (Figure 14).

Do evolutionary constraints correlate with gene family size? Figure 15 shows K_a/K_s (averaged over members of a gene family) plotted against the number of duplicates a gene has. Both yeasts show a positive correlation between K_a/K_s and the number of duplicates (*S. cerevisiae*: Pearson's *r*=0.397; Spearman's *s*=0.508, **P**<0.0001; *S. pombe*: Pearson's *r*=0.533, Spearman's *s*=0.511, **P**<0.0001 for both). In *S. cerevisiae*, removing the seripauperins, a poorly characterized but very large gene family (Viswanathan *et al*. 1994) further increases the magnitude of these associations (Pearson's *r*: 0.591; Spearman's *s*=0.561).

Perhaps surprisingly, both *C. elegans* and *Drosophila* show a negative correlation between the number of duplicates and the K_a/K_s ratio (*C. elegans*: Pearson's *r*: -0.122, Spearman's *s*: -0.073, **P**<0.0001; *Drosophila*: Pearson's *r*: -0.116, **P**<0.0001 and Spearman's *s*: -0.061, **P**=0.017). Both associations are weak in magnitude but significant because of the sheer number of observations. Finally, *E. coli* shows no significant association between K_a/K_s and the number of gene duplicates.

Discussion

Caution is necessary in applying any automated software tool to analyze the evolutionary history of genomes. The reason is that choice of analysis parameters by an investigator can critically influence results. We had to make such choices not only in the assignment of genes to categories, but also in setting similarity thresholds for including gene pairs. For instance, we deliberately chose a conservative approach, admitting only

highly similar gene pairs to our analysis. This may explain why some statistical patterns detected in other analyses, e.g., the expansion of certain regulatory gene families in fruit fly and nematode (Chervitz *et al*. 1998; Rubin *et al*. 2000), have not been detected here. Their expansion occurred so long ago that individual gene family members may have become too dissimilar to be detected in a conservative assay. On the other hand, the advantage of our conservative approach is that detected patterns are less likely to be spurious.

A number of evolutionary patterns found here may be easily explained. They include the overrepresentation of duplicates in transport and metabolic genes in the microbial genomes, as well as a general under-representation of ribosomal protein genes with multiple duplicates. Dosage effects may make it difficult to maintain duplicate ribosomal proteins in a genome, unless, as in budding yeast, a whole-genome duplication has duplicated all of the proteins at once. Some of the patterns we see have been observed independently by others, which adds to our confidence in them. They include the amplification of duplicates related to hexose transport in yeast (Brown *et al*. 1998), as well as amplification of the histone gene family in *C. elegans* (Roberts *et al*. 1987), and the kinase gene family in *Drosophila* (Chervitz *et al*. 1998; Rubin *et al*. 2000; Suga *et al*. 1999). Such patterns suggest that the rate of gene duplication is by no means homogenous across the genome. Rather, this rate is affected by both biochemistry and cell biology (as illustrated by how dosage effects of highly expressed genes influence duplication probability) as well as by conditions specific to particular organisms and their environments (for instance in the case of the yeast hexose transporters).

Our analysis also considered selective constraints specific to gene families, as indicated by the ratio of K_a/K_s. While very few significant differences occur among functional categories, we observed higher K_a/K_s ratios (weaker constraints) in the two multicellular eukaryotes relative to the microbial eukaryotes. This trend might reflect a previously reported stronger relaxation of K_a/K_s shortly after duplication in higher organisms (Lynch & Conery 2000).

Striking taxon-specific differences exist in the association between selective constraint (K_a/K_s) and gene family size. *E. coli* shows no such association, the microbial eukaryotes show a highly positive association, and the higher eukaryotes show a weakly negative (but highly significant) association. The most straightforward explanation of the correlation seen in the yeasts is that large gene families "buffer" the effect of mutations in one of their members and thus allow a higher amino acid substitution rate. That this pattern is not observed in the many-celled eukaryotes is in line with population genetic arguments showing that only very large populations (as are likely to occur in yeasts) can sustain such buffering through redundancy (Wagner 2000c). In addition, the manifold greater possibilities for tissue-specific expression of duplicates in the multicellular organisms may prevent duplicates in large families from experiencing relaxed constraints.

Complementary data further supports a relation between gene family size and buffering for budding yeast. Among 540 genes with one or more duplicates that meet our criteria ($K_s<3$, $K_a<0.75$, $K_a/K_s<1$), only 18 are known to be essential in yeast (as indicated by the lethality of a synthetic-null mutation). Moreover, none of these 18 genes have more than 5 duplicates (Winzeler *et al*. 1999). (Previous analysis had found four

essential genes with duplicates Winzeler *et al*. 1999). We also observe, anecdotally, that no budding yeast gene with more than 9 duplicates has a functional annotation in the gene ontology database (The Gene Ontology Consortium 2000). This indicates the well-known difficulty of identifying gene functions in large gene families by genetic means. However, while such evidence may insinuate a simple explanation for an observed statistical pattern, caution is appropriate. First, perhaps as many as half of all yeast gene deletions with no phenotypic effect affect single copy genes, showing that redundancy through gene duplication is not all there is to buffering of mutational effects (Wagner 2000b). Also, highly similar duplicates did not generally show weaker effects in synthetic-null mutations in early analyses (Wagner 2000b), although that picture has been somewhat revised by larger datasets (Gu *et al*. 2003, see chapter 3). And finally and most importantly, the lack of an association between gene family size and evolutionary constraint in *E.coli* is squarely at odds with the above interpretation.

The negative correlation between gene family size and K_a/K_s in the two multicellular eukaryotes is more difficult to understand. We suspect that the *Drosophila* correlation is largely a result of the very small number of large gene families, which simply do not show the variation in K_a/K_s that the small gene families do. Figure 15C indicates this, with the very high and low K_a/K_s values all being located among small gene families. The correlation in *C. elegans* is stronger, and we suspect that there are one or more large gene families with specific functions that are driving the relationship. In particular, removing the major sperm protein family (Klass *et al*. 1984) (which functions both in sperm motility and in oocyte signaling Miller *et al*. 2001; Roberts & Stewart

2000) reduces Pearson's *r* from –0.122 to –0.093 and Spearman's *s*: from –0.073 to –0.058, although the significance in each case is unchanged at **P**<0.0001 (see figure 15D).

Unfortunately, difficult to explain patterns are still the norm rather than exception in analyzing genome evolution. Other such patterns include an under-representation of duplicated transcription factor genes (Fig. 13), a large difference in numbers of histone genes between nematode and fruit fly, and the disproportionately large major sperm protein family of the nematode. However, such unexplained patterns make clear that genome sequencing projects have accomplished something very important. They have opened new frontiers of inquiry.

References

Adams, M. D. (and 194 others) 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195.

Alon, U. 2003 Biological networks: The tinkerer as an engineer. *Science* **301**, 1866-1867.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped Blast and Psi-Blast : A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.

Aravin, A. A., Naumova , N. M., Tulin, A. V., Vagin, V. V., Rozovsky , Y. M. & Gvozdev , V. A. 2001 Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Current Biology* **11**, 1017-1027.

Bailey, J. A., Yavor, A. M., Viggiano, L., Misceo, D., Horvath, J. E., Archidiacono, N., Schwartz, S., Rocchi, M. & Eichler, E. E. 2002 Human-specific duplication and mosaic transcripts: The recent paralogous structure of chromosome 22. *American Journal of Human Genetics* **70**, 83-100.

Bakker, R. T. 1986 *The Dinosaur Heresies*. New York: Zebra Books.

Bennetzen, J. L. & Hall, B. D. 1982 Codon selection in yeast. *Journal of Biological Chemistry* **257**, 3026-3031.

Benton, B. K., Tinkelenberg, A. H., Jean, D., Plump, S. D. & Cross, F. R. 1993 Genetic analysis of Cln/Cdc28 regulation of cell morphogenesis budding yeast. *EMBO Journal* **12**, 5267-5275.

Bernardi, G. & Bernardi, G. 1986 Compositional constraints and genome evolution. *Journal of Molecular Evolution* **24**, 1-11.

Bisbee, C. A., Baker, M. A. & Wilson, A. C. 1977 Albumin phylogeny for clawed frogs (*Xenopus*). *Science* **195**, 785-787.

Blattner, F. R.(and 16 others) 1997 The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462.

Bray, D. 2003 Molecular networks: The top-down view. *Science* **301**, 1864-1865.

Bronowski, J. 1973 The ladder of creation. In *The Ascent of Man*, pp. 291-319. Boston: Little, Brown and Company.

Brown, C. J., Todd, K. M. & Rosenzweig, R. F. 1998 Multiple duplications of yeast hexose-transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution* **15**, 931-942.

Brown, C. S., Goodwin, P. C. & Sorger, P. K. 2001 Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences, U.S.A.* **98**, 8944-8949.

The *C. elegans* Sequencing Consortium. 1998 Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012-2018.

Chen, L., DeVries, A. L. & Cheng, C.-H. C. 1997 Convergent evolution of anti-freeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proceedings of the National Academy of Sciences, U. S. A.* **94**, 3817-3822.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. 1998 SGD: Saccharomyces Genome Database. *Nucleic Acids Research* **26**, 73-80.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J. M. & Botstein, D. 1998 Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282**, 2022-2028.

Chiu, C.-H., Nonaka, D., Xue, L., Amemiya, C. T. & Wagner, G. P. 2000 Evolution of *Hoxa-11* in lineages phylogenetically positioned along the fin-limb transition. *Molecular Phylogenetics and Evolution* **17**, 305-316.

Comeron, J. M. & Aguade, M. 1998 An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution* **47**, 268-274.

Conant, G. C. & Wagner, A. 2002 GenomeHistory: A software tool and its application to fully sequenced genomes. *Nucleic Acids Research* **30**, 3378-3386.

Conant, G. C. & Wagner, A. 2003 Asymmetric sequence divergence of duplicate genes. *Genome Research* **13**, 2052-2058.

Cooke, J., Nowak, M. A., Boerlijst, M. & Maynard-Smith, J. 1997 Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends in Genetics* **13**, 360-364.

Daffre, S., Kylsten, P., Samakovlis, C. & Hultmark, D. 1994 The lysozyme locus in *Drosophila melanogaster*: An expanded gene family adapted for expression in the digestive tract. *Molecular and General Genetics* **242**, 152-162.

Darwin, C. 1859 *The Origin of Species by Means of Natural Selection*. London: John Murry.

de Bono, M., Tobin, D. M., Davis, M. W., Avery, L. & Bargmann, C. I. 2002 Social feeding in *Caenorhabditis elegans* is induced by neurons that detect aversive stimuli. *Nature* **419**, 899-903.

Delattre, M. & Félix, M.-A. 2001 Microevolutionary studies in nematodes: A beginning. *BioEssays* **23**, 807-819.

Dermitzakis, E. T. & Clark, A. G. 2001 Differential selection after duplication in mammalian developmental genes. *Molecular Biology and Evolution* **18**, 557-562.

Dermitzakis, E. T. & Clark, A. G. 2002 Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Molecular Biology and Evolution* **19**, 1114-1121.

Doolittle, R. F. 1994 Convergent evolution : The need to be explicit. *Trends in Biochemical Sciences* **19**, 15-18.

Edwards, J. S. & Palsson, B. O. 2000 The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences, U.S.A.* **97**, 5528-5533.

Fackenthal, J. D., Hutchens, J. A., Turner, F. R. & Raff, E. C. 1995 Structural analysis of mutations in the *Drosophila* β2-tubulin isoform reveals regions in the β-tubulin molecule required for general and for tissue-specific microtubule functions. *Genetics* **139**, 267-286.

Fackenthal, J. D., Turner, F. R. & Raff, E. C. 1993 Tissue-specific microtubule functions in *Drosophila* spermatogenesis require the β2-tubulin isotype-specific carboxy terminus. *Developmental Biology* **158**, 213-227.

Fay, J. C., Wycoff, G. J. & Wu, C.-I. 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024-1026.

Fell, D. A. & Wagner, A. 2000 The small world of metabolism. *Nature Biotechnology* **18**, 1121-1122.

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376.

Ferea, T. L., Botstein, D., Brown, P. O. & Rosenzweig, R. F. 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 9721-9726.

Ferris, S. D. & Whitt, G. S. 1977 Loss of duplicate gene expression after polyploidisation. *Nature* **265**.

Ferris, S. D. & Whitt, G. S. 1979 Evolution of the differential regulation of duplicate genes after polyploidization. *Journal of Molecular Evolution* **12**, 267-317.

Fire, A., X., S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. 1998 Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-811.

Fisher, R. A. 1930 *The Genetical Theory of Natural Selection*. Oxford: Clarendon.

Fletcher, G. L., Hew, C. L. & Davies, P. L. 2001 Antifreeze proteins of teleost fishes. *Annual Review of Physiology* **63**, 359-390.

The FlyBase Consortium. 2002 The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research* **30**, 106-108.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. & Postlethwait, J. 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-1545.

Fraenkel, D. G. 1982 Carbohydrate Metabolism. In *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (ed. J. N. Strathern, E. W. Jones & J. R. Broach), pp. 1-37. Cold Spring Harbor: Cold Spring Harbor Laboratory.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. 2002 Evolutionary rate in the protein interaction network. *Science* **296**, 750-752.

Fromental-Ramain, C., Warot, X., Messadecq, N., LeMeur, M., Dollé, P. & Chambon, P. 1996 *Hoxa-13* and *Hoxd-13* play a crucial role in the patterning of the limb autopod. *Development* **122**, 2997-3011.

Futuyma, D. J. 1998 *Evolutionary Biology: 3rd Edition*. Sunderland, MA: Sinauer Associates, Inc.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. 2000 Genomic expression programs in the response of yeast cells to environmental change. *Molecular Biology of the Cell* **11**, 4241-4257.

The Gene Ontology Consortium. 2000 Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29.

Gerstein, M. 1997 A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *Journal of Molecular Biology* **274**, 562-576.

Giaever, G. (and 72 others) 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-391.

Gibson, T. J. & Spring, J. 1998 Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in Genetics* **14**, 46-49.

Goffeau, A. (and 15 others) 1996 Life with 6000 genes. *Science* **274**, 546-567.

Goldman, N. 1993 Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182-198.

Goldman, N. & Yang, Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725-736.

Gould, S. J. 1980 The Panda's Thumb. In *The Panda's Thumb*. New York: W. W. Norton.

Gould, S. J. 1989 *Wonderful Life: The Burgess Shale and the Nature of History*. New York: W. W. Norton.

Graur, D. & Li, W. H. 2000 *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.

Gu, Z., Nicolae, D., Lu, H. H.-S. & Li, W.-S. 2002 Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18**, 609-613.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W.-H. 2003 Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-66.

Haack, H. & Gruss, P. 1993 The establishment of murin Hox-1 expression domains during patterning of the limb. *Developmental Biology* **157**, 410-422.

Hahn, M. W., Conant, G. C. & Wagner, A. 2004 Molecular evolution in large genetic networks: Connectivity does not equal constraint. *Journal of Molecular Evolution* **58,** 203-211.

Haldane, J. B. S. 1933 The part played by recurrent mutation in evolution. *American Naturalist* **67**, 5-9.

Hanks, M., Wurst, W., Ansoncartwright, L., Auerbach, A. B. & Joyner, A. L. 1995 Rescue of the En-1 mutant phenotype by replacement of En-1 with En-2. *Science* **269**, 679-682.

Harris, J. I., Sanger, F. & Naughton, M. A. 1956 Species differences in insulin. *Archives of Biochemistry and Biophysics* **65**, 427-438.

Harrison, P. M., Echolds, N. & Gerstein, M. B. 2001 Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Research* **29**, 818-830.

Hartl, D. L. & Clark, A. G. 1997 *Principles of Population Genetics, 3rd Edition*. Sunderland MA: Sinauer associates.

Hedstrom, L., Perona, J. J. & Rutter, W. J. 1994 Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry* **33**, 8757-8763.

Hendriks, W., Mulders, J. W. M., Bibby, M. A., Slingsby, C., Bloemendal, H. & de Jong, W. W. 1988 Duck lens ε-crystallin and lactate dehydrogenase B_4 are identical: A single copy gene product with two distinct functions. *Proceedings of the National Academy of Sciences, U.S.A.* **85**, 7114-7118.

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. & Brown, E. L. 2000 Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809-812.

Hillis, D. M., Moritz, C. & Mable, B. K. 1996 *Molecular Systematics: Second Edition*. Sunderland, MA: Sinauer Associates.

Holland, P. W. H. 1999 Gene duplication: Past, present and future. *Seminars in Cell and Developmental Biology* **10**, 541-547.

Hoyle, H. D. & Raff, E. C. 1990 Two *Drosophila* Beta-tubulin isoforms are not functionally equivalent. *Journal of Cell Biology* **111**, 1009-1026.

Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. 1998 SCOP, structural classification of proteins database: Applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallographica Section D-Biological Crystallography* **54**, 1147-1154.

Hughes, A. L., Green, J. A., Garbayo, J. M. & Roberts, R. M. 2000a Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proceedings of the National Academy of Sciences, U.S.A.* **97**, 3319-3323.

Hughes, M. K. & Hughes, A. L. 1993 Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Molecular Biology and Evolution* **10**, 1360-1369.

Hughes, T. R. (and 21 others) 2000b Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126.

Huynen, M. A. & van Nimwegen, E. 1998 The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution* **15**, 583-589.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences, U.S.A.* **98**, 4569-4574.

Iwabe, N., Kuma, K. & Miyata, T. 1996 Evolution of gene families and relationship with organismal evolution: Rapid divergence of tissue-specific genes in the early evolution of chordates. *Molecular Biology and Evolution* **13**, 483-493.

Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651-654.

Ji, Q., Norell, M. A., Gao, K.-Q., Ji, S.-A. & Ren, D. 2001 The distribution of integumentary structures in a feathered dinosaur. *Nature* **410**, 1084-1088.

Kamath, R. S.(and 12 others) 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231-237.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. & Davidson, G. S. 2001 A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087-2092.

Kimbel, M., Incardona, J. P. & Raff, E. C. 1989 A variant β-tubulin isoform of *Drosophila melanogaster* (β3) is expressed primarily in tissues of mesodermal origin in embryos and pupae, and is utilized in populations of transient microtubules. *Developmental Biology* **131**, 415-429.

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

Kipreos, E. T. & Pagano, M. 2000 The F-box protein family. *Genome Biology* **1**, reviews3002.1 - 3002.7.

Klass, M. R., Kinsley, S. & Lopez, L. C. 1984 Isolation and characterization of a sperm-specific gene family in the nematode *Caenorhabditis elegans*. *Molecular and Cellular Biology* **4**, 529-537.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. 2002 Selection in the evolution of gene duplicates. *Genome Biology* **3**, 0008.1-0008.9.

Kornegay, J. R., Schilling, J. W. & Wilson, A. C. 1994 Molecular adaptation of a leaf-eating bird: Stomach lysozyme of the Hoatzin. *Molecular Biology and Evolution* **11**, 921-928.

Krakauer, D. C. & Nowak, M. A. 1999 Evolutionary preservation of redundant duplicated genes. *Seminars in Cell and Developmental Biology* **10**, 555-559.

Kumar, A. (and 14 others) 2002 Subcellular localization of the yeast proteome. *Genes and Development* **16**, 707-719.

Kylsten, P., Kimbrell, D. A., Daffre, S., Samakovlis, C. & Hultmark, D. 1992 The lysozyme locus in *Drosophila melanogaster*: Different genes are expressed in midgut and salivary glands. *Molecular and General Genetics* **232**, 335-343.

Langkjær, R. B., Cliften, P. F., Johnston, M. & Piskur, J. 2003 Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. *Nature* **421**, 848-852.

Lee, T. I.(and 20 others) 2002 Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799-804.

Lewis, P. O. 2001 Phylogenetic systematics turns over a new leaf. *Trends in Ecology and Evolution* **16**, 30-37.

Li, W.-H. 1980 Rate of gene silencing at duplicate loci: A theoretical study and interpretation of data from tetraploid fish. *Genetics* **95**, 237-258.

Li, W.-H. 1997 *Molecular Evolution*. Sunderland, MA.: Sinauer Associates.

Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. 2001 Evolutionary analyses of the human genome. *Nature* **409**, 847-849.

Li, X. L. & Noll, M. 1994 Evolution of distinct developmental functions of 3 *Drosophila* genes by acquisition of different cis-regulatory regions. *Nature* **367**, 83-87.

Li, Y.-J. & Tsoi, S. C.-M. 2002 Phylogenetic analysis of vertebrate lactate dehydrogenase (LDH) multigene families. *Journal of Molecular Evolution* **54**, 614-624.

Liò, P. & Goldman, N. 1998 Models of molecular evolution and phylogeny. *Genome Research* **8**, 1233-1244.

Lundin, L. 1999 Gene duplications in early metazoan evolution. *Cell and Developmental Biology* **10**, 523-530.

Lynch, M. & Conery, J. S. 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.

Mathews, C. K. & Van Holde, K. E. 1996 *Biochemistry*. Menlo Park: The Benjamin/Cummings Publishing Company Inc.

McDonald, J. H. & Kreitman, M. 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652-654.

Menzel, R., Bogaert, T. & Achazi, R. 2001 A systematic gene expression screen of *Caenorhabditis elegans* cytochrome P450 genes reveals CYP35 as strongly xenobiotic inducible. *Archives of Biochemistry and Biophysics* **395**, 158-168.

Miller, M. A., Nguyen, V. Q., Lee, M.-H., Kosinski, M., Schedl, T., Caprioli, R. M. & Greenstein, D. 2001 A sperm cytoskeletal protein that signals oocyte meiotic maturation and ovulation. *Science* **291**, 2144-2147.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii., D. & Alon, U. 2002 Network motifs: Simple building blocks of complex networks. *Science* **298**, 824-827.

Muse, S. V. & Gaut, B. S. 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715-724.

Muse, S. V. & Weir, B. S. 1992 Testing for equality of evolutionary rates. *Genetics* **132**, 269-276.

Muzzarelli, R. A. A. & Muzzarelli, C. 1998 Native and modified chitins in the biosphere. In *Nitrogen-containing macromolecules in the bio- and geosphere*, vol. 707 (ed. P. F. van Bergen), pp. 148-162. Washington, DC: American Chemical Society.

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

Nasmyth, K. 1993 Control of the yeast cell cycle by the Cdc28 protein kinase. *Current Opinion in Cell Biology* **5**, 166-179.

Nei, M. & Roychoudhury, A. K. 1973 Probability of fixation of nonfunctional genes at duplicate loci. *American Naturalist* **107**, 362-372.

Nelson, C. E., Morgan, B. A., Burke, A. C., Laufer, E., DiMambro, E., C., M., Gonzales, E., Tessarollo, L., Parada, L. F. & Tabin, C. 1996 Analysis of *Hox* gene expression in the chick limb bud. *Development* **122**, 1449-1466.

Nowak, M. A., Boerlijst, M. C., Cooke, J. & Maynard-Smith, J. 1997 Evolution of genetic redundancy. *Nature* **388**, 167-171.

Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. & Hartl, D. L. 1998 Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396**, 572-575.

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

Ota, T. & Nei, M. 1994 Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *Journal of Molecular Evolution* **38**, 642-643.

Pál, C., Papp, B. & Hurst, L. D. 2001 Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-931.

Pál, C., Papp, B. & Hurst, L. D. 2003 Rate of evolution and gene dispensability. *Nature* **421**, 496497.

Patel, P. H. & Loeb, L. A. 2000 DNA polymerase active site is highly mutable: evolutionary consequences. *Proceedings of the National Academy of Sciences, U.S.A.* **97**, 5095-5100.

Pilgrim, D. & Young, E. T. 1987 Primary structure requirements for correct sorting of the yeast mitochondrial protein ADH III to the yeast mitochondrial matrix space. *Molecular and Cellular Biology* **7**, 294-304.

Planta, R. J. & Mager, W. H. 1998 The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast* **14**, 471-477.

Press, W. H., Teukolsky, S. A., Vetterling, W. A. & Flannery, B. P. 1992 *Numerical Recipes in C*. New York: Cambridge University Press.

Qian, J., Luscombe, N. M. & Gerstein, M. 2001 Protein family and fold occurrence in genomes: Power-law behavior and evolutionary model. *Journal of Molecular Biology* **313**, 673-681.

Regel, R., Matioli, S. R. & Terra, W. R. 1998 Molecular adaptation of *Drosophila melanogaster* lysozymes to a digestive function. *Insect Biochemistry and Molecular Biology* **28**, 309-319.

Roberts, S. B., Sanicola, M., Emmons, S. W. & Childs, G. 1987 Molecular characterization of the histone gene family of *Caenorhabditis elegans*. *Journal of Molecular Biology* **196**, 27-38.

Roberts, T. M. & Stewart, M. 2000 Acting like actin: The dynamics of the nematode major sperm protein (MSP) cytoskeleton indicate a push-pull mechanism for amoeboid cell motility. *Journal of Cell Biology* **149**, 7-12.

Rubin, G. M. (and 40 others) 2000 Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215.

Ruse, M. 2003 Is evolution a secular religion? *Science* **299**, 1523-1524.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. & Herzel, H. 2000 Normalization strategies for cDNA microarrays. *Nucleic Acids Research* **28**, E47i-E47v.

Seoighe, C. & Wolfe, K. H. 1998 Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 4447-4452.

Seoighe, C. & Wolfe, K. H. 1999 Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology* **2**, 548-554.

Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. 2002 Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31**, 64-68.

Simon, J. M. & Sternberg, P. W. 2002 Evidence of a mate-finding cue in the hermaphrodite nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences, U.S.A.* **99**, 1598-1603.

Smith, N. G. C. & Eyre-Walker, A. 2002 Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022-1024.

Smith, V., Chou, K. N., Lashkari, D., Botstein, D. & Brown, P. O. 1996 Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069-2074.

Sokal, R. R. & Rohlf, F. J. 1995 *Biometry: 3rd Edition*. New York: W. H. Freeman and Company.

Sordino, P., van der Hoeven, F. & Duboule, D. 1995 *Hox* gene expression in teleost fins and the origin of vertebrate digits. *Nature* **375**, 678-681.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. 1998 Comprehensive identification of cell-cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J. & Spieth, J. 2001 WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research* **29**, 82-86.

Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T., Chu, A. M., Giaever, G., Prokisch, H., Oefner, P. J. & Davis, R. W. 2002 Systematic screen for human disease genes in yeast. *Nature Genetics* **31**, 400-404.

Stewart, C.-B., Schilling, J. W. & Wilson, A. C. 1987 Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401-404.

Stone, J. R. & Wray, G. A. 2001 Rapid evolution of *cis*-regulatory sequences via local point mutations. *Molecular Biology and Evolution* **18**, 1764-1770.

Suga, H., Koyanagi, M., Hoshiyama, D., Ono, K., Iwabe, N., Kuma, K. & T., M. 1999 Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and Protein Tyrosine Kinases from sponge and hydra. *Journal of Molecular Evolution* **48**, 646-653.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994 CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

Toda, T., Cameron, S., Sass, P., Zoller, M. & Wigler, M. 1987 Three different genes in *S. cerevisiae* encode the catalytic subunits of the cAMP-dependent protein kinase. *Cell* **50**, 277-287.

Tsaur, S. C., Ting, C. T. & Wu, C. I. 1998 Positive selection driving the evolution of a gene of male reproduction, Acp26aa , of *Drosophila* : II. Divergence versus polymorphism. *Molecular Biology and Evolution* **15**, 1040-1046.

Uetz, P.(and 19 others) 2000 A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.

Van de Peer, Y., Taylor, J. S., Braasch, I. & Meyer, A. 2001 The ghost of selection past: Rates of evolution and functional divergence of anciently duplicated genes. *Journal of Molecular Evolution* **53**, 436-446.

Venter, J. C. (and 273 others) 2001 The sequence of the human genome. *Science* **291**, 1304-1351.

Viswanathan, M., Muthukumar, G., Cong, Y. S. & Lenard, J. 1994 Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. *Gene* **148**, 149-153.

Wagner, A. 1999 Redundant gene functions and natural selection. *Journal of Evolutionary Biology* **12**, 1-16.

Wagner, A. 2000a Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proceedings of the National Academy of Sciences, U.S.A.* **97**, 6579-6584.

Wagner, A. 2000b Robustness against mutations in genetic networks of yeast. *Nature Genetics* **24**, 355-361.

Wagner, A. 2000c The role of population size, pleiotropy, and fitness effects of mutations in the evolution of overlapping gene function. *Genetics* **154**, 1389-1401.

Wagner, A. 2001 The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution* **18**, 1283-1292.

Wagner, A. 2002 Asymmetric functional divergence of duplicate genes in yeast. *Molecular Biology and Evolution* **19**, 1760-1768.

Wang, Y. K., Schnegelsberg, P. N. J., Dausman, J. & Jaenisch, R. 1996 Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin. *Nature* **379**, 823-825.

Watson, J. D. & Crick, F. H. C. 1953 Molecular structure of nucleic acids. *Nature* **171**, 737.

Winzeler, E. A. (and 50 others) 1999 Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901-906.

Wistow, G. J., Mulders, J. W. M. & de Jong, W. W. 1987 The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature* **326**, 622-624.

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

Wood, V.(and 132 others) 2002 The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871-880.

Wright, F. 1990 The 'effective number of codons' used in a gene. *Gene* **87**, 23-29.

Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97-159.

Wyckoff, G. J., Wang, W. & Wu, C.-I. 2000 Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304-309.

Xu, X., Zhou, Z.-h. & Prum, R. O. 2001 Branched integumental structures in *Sinornithosaurus* and the origin of feathers. *Nature* **410**, 200-204.

Yang, A. & Nielsen, R. 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32-43.

Yang, Z. 2000 Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution* **51**, 423-432.

Zhang, J. Z., Rosenberg, H. F. & Nei, M. 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 3708-3713.