DUPLICATE GENE EVOLUTION & EXPRESSION AFTER POLYPLOIDIZATION

DUPLICATE GENE EVOLUTION AND EXPRESSION AFTER POLYPLOIDIZATION

By

FRÉDÉRIC J.J. CHAIN, B.SC. (HONS.)

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2009) McMaster University
(Biology) Hamilton, Ontario

TITLE: Duplicate gene evolution and expression after polyploidization

AUTHOR: Frédéric J.J. Chain, B.Sc. Hons. (University of Guelph)

SUPERVISOR: Dr. Ben J. Evans

NUMBER OF PAGES: x, 174

ABSTRACT

Gene duplications can facilitate genetic innovation, reduce pleiotropy and catalyze reproductive incompatibilities and speciation. Therefore, the molecular and transcriptional fate of duplicate genes plays an important role in the evolutionary trajectory of entire genomes and transcriptomes. Using the polyploid African clawed frog *Xenopus*, I have investigated mechanisms that promote the retained expression of duplicate genes (paralogs) after whole genome duplication. The studies herein estimated molecular evolution and characterized expression divergence of thousands of duplicate genes and a singleton ortholog from a diploid outgroup. In this thesis, I have discussed the multiple mechanisms for the retention of duplicate genes in a polyploid genome and examined the potential effects that gene characteristics before duplication have on the odds of duplicate gene persistence. I have also explored the use of microarrays for comparative transcriptomics between duplicate genes, and between diverged genomes.

The main objectives of my thesis were to better understand the genetic mechanisms that promote the retained expression of gene duplicates. My research utilized the duplicated genome from the allopolyploid clawed frog *Xenopus*. Genome duplication in clawed frogs offers a compelling opportunity to study factors that influence the genetic fates of gene duplicates because many paralogs in these frogs are of the same age, permitting one to control for the influence of time when evaluating the impact of duplication. My work has major impacts on several biological fronts including evolutionary genomics and comparative transcriptomics, and also on technical aspects of using microarrays. I have provided among the most comprehensive studies of its kind, in terms of examining molecular and regulatory aspects of thousands of expressed duplicates of the same age, and exploring various alternative hypotheses to explain how these genes are retained.

PREFACE

The chapters of this thesis have been written as separate manuscripts, of which chapters 1, 2 and 4 have been published. For all four chapters, data collection was almost exclusively an individual effort, while analysis, manuscript preparation and editing were performed with the contributions and guidance of my supervisor Ben Evans. I orchestrated the identification and alignment of paralogs and orthologs, the analyses of molecular evolution and gene expression, and the construction of gene probemasks for the microarrays. For chapters 2 and 4, Dora Ilieva performed the microarray hybridizations and helped supervise the initial microarray analyses.

ACKNOWLEDGEMENTS

My academic persistence would be inexistent without Ben Evans. I thank my supervisor for his unyielding encouragement and guidance, immeasurable care and dedication, unparalleled enthusiasm and humour, and for not only being a facilitator but an inspiration to achieve more than I thought was possible.

My committee members Drs. Brian Golding, Richard Morton, and Jonathon Stone have given me great suggestions and have been helpful and positive influences throughout. I am also grateful to Drs. Melanie Huntley, Weilong Hao and Wilfried Haerty for contributing their time to advise me along the way. Many other students have been great study companions, among them Masters Megan Barclay, Fathima Ishna Iftikar, and Drs. Maria Abou-Chakra, Abha Ahuja, Paul Craig, John Fitzpatrick and Sheng Sun, and of course many more who deserve to be mentioned.

Every member of the Evans lab since its genesis has made lasting impressions on me, and I hope they have benefited from suffering through my antics. Eliza Wojcik and Dr. Mohammad Sumo Zubairi were the first who entertained me in the lab, while Brady Tracey, Stephanie Sun and Laura Pin have happily aided me in my frog cleaning duties. Master Dave Anderson was always willing to keep me on my toes, Adam Bewick gives me more credit than I deserve, and of course M. Iqbal Setiadi never ceases to amaze me with his relentless attempts to thwart me in all my doings, in such a good way. Thank you.

CONTENTS

LIST OF FIGURES

INTRODUCTION

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 4

PART I – INTRODUCTION

Duplicate Genes

Soon after gene duplication, one copy is expected to be silenced by acquiring deleterious mutations in the coding or regulatory regions, or both (Altschmied et al. 2002; Force et al. 1999; Li 1980; Lynch and Conery 2003; Prince and Pickett 2002; Taylor et al. 2001a; Wagner 2002). However, an unexpectedly large number of duplicate genes persist for long periods of time in many organisms (Lynch and Conery 2000; Nadeau and Sankoff 1997; Postlethwait et al. 2000). This suggests that duplicate genes play an integral role in genome evolution, possibly supplying opportunities for evolving novel genes (Ohno 1970; Zhang 2003) or reducing pleiotropy (Carroll 2005; Lynch and Force 2000). Although the theoretical plausibility of new functions arising from duplicates has been discussed (Lynch and Conery 2000; Massingham et al. 2001), the mechanisms behind duplicate gene retention, whether they counteract or exploit degenerative mutations, remain of great interest to evolutionary biologists.

Functional divergence of duplicate genes can lead to the preservation of both copies if mutations in either paralog affect protein function or gene expression regulation in such a way that losing a copy incurs a fitness disadvantage (Li 1980; Nadeau and Sankoff 1997). There are several proposed mechanisms that attempt to explain how duplicate genes are retained, but they are not necessarily mutually exclusive because different mechanisms could operate concurrently, on different parts of the genes, in both the coding and regulatory regions, and at different times after duplication (Force et al. 1999). Furthermore, subsequent neutral mutations (Zhang et al. 1998), different evolutionary rates between lineages (Lynch and Conery 2000; Nembaware et al. 2002), genomic locations of each gene copy (Lercher et al. 2004), and the incorrect identification of duplicates that are actually pseudogenes or allelic variants are all variables that can distort the molecular signal and influence the assignment of mechanisms. For example, positive selection on a gene may be restricted to specific functional sites or may only act during a short period of time so that it may not be detected by molecular analyses (Zhang et al. 1998). Nonetheless, if a particular mechanism operates for an extended period of time or on a large portion of the paralog(s), the molecular changes incurred should make it detectable.

Following duplication, transcripts from paralogous genes will either remain similar or diverge. Paralogs that differ in coding sequence may be functionally divergent; one duplicate can carry out a novel function (*neofunctionalization),* or both duplicates can undergo changes in complementary regions so that multiple ancestral functions have now been subdivided between each duplicate (*subfunctionalization).* Regulatory changes may ensue so that they are further divergent in terms of expression. Consistent with neofunctionalization is the divergence of the eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN), genes in primates which originated from a duplication; ECP has since developed a novel anti-pathogen toxicity that is not present in EDN, nor was this activity present before the duplication event (Zhang et al. 1998). Consistent with subfunctionalization, which can occur if different functional domains on each

paralog were enhanced (subfunction co-option) or degraded (subfunction partitioning of coding sequences), are the *mitf* genes. These microphthalmia-associated transcription factor (*MITF*) paralogs in zebrafish complement each other to perform the function of their singleton ortholog in birds and mammals (Altschmied et al. 2002; Lister et al. 2001). These duplicated genes differ in that alternative exons and regulatory elements have shown degeneration (Altschmied et al. 2002).

Although single amino acid substitutions can radically alter protein function (Gibson and Spring 1998; Golding and Dean 1998), paralogs that share high coding sequence identity likely perform similar functions. Under these scenarios, paralogs display a form of *redundancy* where a duplicate is probably retained for regulatory reasons; the extra copy may assist in gene dosage or dosage balance, or may be expressed in a different location or at a different time (*regulatory neofunctionalization or subfunctionalization*) (Adams and Wendel 2005; Aury et al. 2006; Force et al. 1999; Tirosh and Barkai 2007). Consistent with redundancy are the SHATTERPROOF and SEPALLATA genes in *Arabidopsis thaliana*; they lack any sign of loss-of-function after the knockout of one of the paralogs, and protein divergence indicates that both gene pairs are under functional constraint, suggesting that they are both maintained for gene dosage purposes (Moore et al. 2005). Signs of regulatory subfunctionalization have been found in zebrafish; the *engrailed* paralogs are expressed in different tissues (Force et al. 1999; Prince and Pickett 2002) while the *hoxb1* paralogs are expressed at different developmental stages in different tissues (McClintock et al. 2002; Prince and Pickett 2002). For both sets of genes, a single gene in mice is expressed in both areas, suggesting that these duplicates stemming from whole genome duplication in fish have partitioned two regulatory-related functions between them.

To summarize the different mechanisms of retention, duplicate genes can maintain functional redundancy by contributing to gene dosage or maintaining protein stoichiometry. Other duplicates may diverge in expression in a spatial, temporal or quantitative fashion; these are all forms of regulatory subfunctionalization. Some may be beneficial to an organism by altering protein function and conferring a selective advantage through neofunctionalization, or by dividing up protein functions and refining them, releasing pleiotropy through functional subfunctionalization (Carroll 2005). These mechanisms can furthermore act concurrently or sequentially on the same or different parts of a gene; subneofunctionalization predicts that subfunctionalization initially acts to retain paralogs in a short transition period, until neofunctionalization plays a more prominent long-term role in preserving duplicate genes (He and Zhang 2005b). The relative importance of these different retention mechanisms remains an area of debate.

Polyploid African clawed frogs (*Xenopus* and *Silurana*)

Xenopus belongs to the Pipidae subfamily Xenopodinae. Within *Xenopus*, there exist at least 12 tetraploid species (2n=36), 5 octoploids (2n=72) and 2 dodecaploids (2n=108), some having arisen through independent polyploidization events in different lineages (Figure 1). *Silurana tropicalis* is the closest known diploid relative (2n=20) and is the only known surviving diploid clawed frog; most other lower ploidy ancestors are believed to be extinct (Evans et al. 2005). Allopolyploidy, the fusion of two separate genomes through backcrossing of hybrids who lay unreduced eggs, is believed to have given rise to these polyploid frogs, and their genomes are primarily disomic; bivalents form between a chromosome and its homolog, as opposed to polysomy where multivalents form and recombination between paralogous loci is more prevalent (Evans et al. 2005; Kobel 1996; Osborn et al. 2003; Tymowska 1991). This allopolyploid origin, the first of which occurred prior to the diversification of the majority of the *Xenopus* species, is evident when considering interspecies gene relationships; each paralog is more similar to its homolog in another species (ortholog), than to its duplicate (paralog).

Xenopus laevis, which is the most widely studied amphibian, has thousands of genes and EST (expressed sequence tag) libraries that are publicly available. Its diploid relative, *Silurana tropicalis* has had its entire genome sequenced and has various EST libraries available. In addition, we collected new data using 454 pyrosequencing from a related species, *X. borealis* and new expression data from *X. laevis*, *X. borealis*, and their hybrids using Affymetrix microarrays designed for *X. laevis*. Together, these databases provide information from polyploid animals and a closely-related living diploid relative, enabling one to investigate if duplicate genes evolve at different rates and are expressed differently relative to themselves and to a singleton ortholog, and what type of genes persist as duplicates.

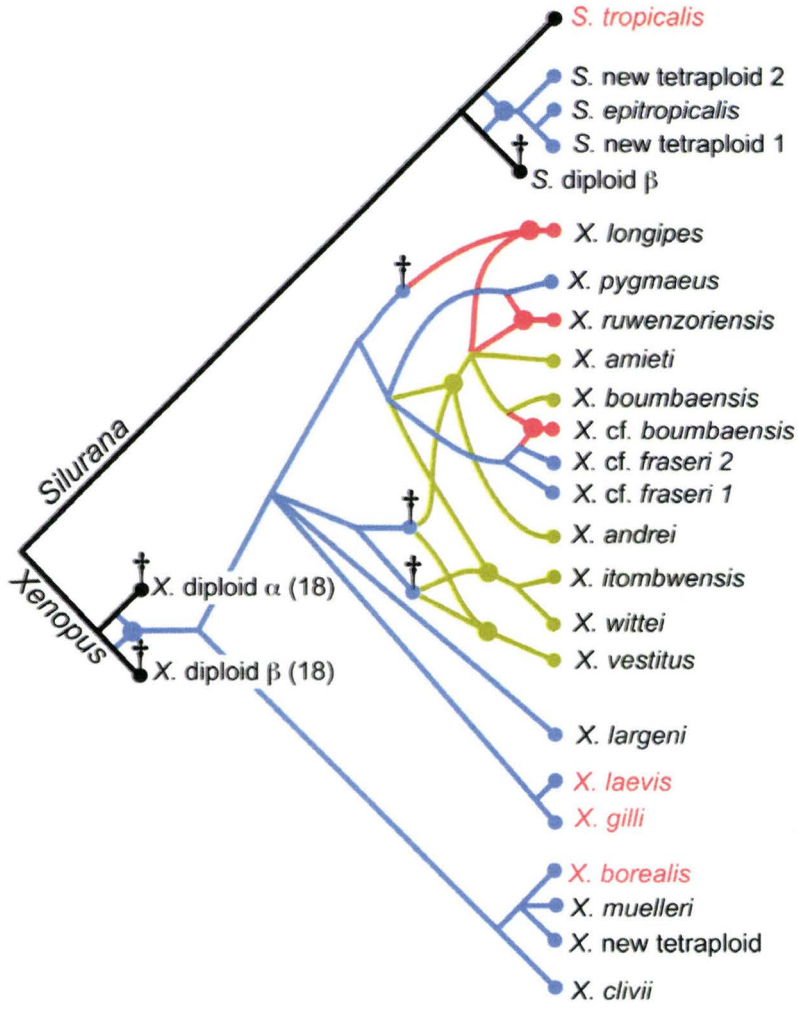


Figure 1. Phylogeny of clawed frogs; adapted from Evans (2007). Branch colors reflect lineage ploidy levels with black being diploid, blue tetraploid, green octoploid, and red dodecaploid. Some species names used in this thesis are highlighted in red. The original genomes that fused together in the ancestral tetraploid are labeled as **α** and **β**.

CHAPTER 1

Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*

Chain, F.J.J. and B.J. Evans (2006) *PLoS Genetics* **2:** e56.

PREFACE

Several hypotheses have been proposed to explain the widespread persistence of duplicate genes. In an attempt to quantify the frequency of different duplicate gene retention mechanisms that act at the protein level, we analyzed aspects of the molecular evolution of paralogs in *Xenopus laevis* and a singleton ortholog in the diploid relative *Silurana tropicalis*.

ABSTRACT

Gene duplication provides a window of opportunity for biological variants to persist under the protection of a co-expressed copy with similar or redundant function. Duplication enables innovation (neofunctionalization), subfunction degeneration (subfunctionalization), and genetic buffering (redundancy), and the genetic survival of each paralog is triggered by mechanisms that add, compromise, or do not alter protein function. We tested the applicability of three types of mechanisms for promoting the retained expression of duplicated genes in 290 expressed paralogs of the tetraploid clawed frog, *Xenopus laevis*. Tests were based on explicit expectations concerning the *ka/ks* ratio, and the number and location of nonsynonymous substitutions after duplication. Functional constraints on the majority of paralogs are not significantly different from a singleton ortholog. However, we find strong support that some of them have an asymmetric rate of nonsynonymous substitution – 6% match predictions of the neofunctionalization hypothesis in that (1) each paralog accumulated nonsynonymous substitutions at a significantly different rate and (2) the one that evolves faster has a higher *ka/ks* ratio than the other paralog and than a singleton ortholog. Fewer paralogs (3%) exhibit a complementary pattern of substitution at the protein level that is predicted by enhancement or degradation of different functional domains, and the remaining 13% have a higher average *ka/ks* ratio in both paralogs that is consistent with mechanisms that select for regulation and/or altered functional constraints. We estimate that these paralogs have been retained since they originated by genome duplication between 21 and 41 million years ago. Multiple mechanisms operate to promote the retained expression of duplicates in the same genome, in genes in the same functional class, over the same period of time following duplication, and sometimes in the same pair of paralogs. None of these paralogs are superfluous; degradation or enhancement of different protein subfunctions and neofunctionalization are plausible hypotheses for

the retained expression of some of them. Evolution of most *X. laevis* paralogs, however, is consistent with retained expression via mechanisms that do not radically alter functional constraints, such as selection to preserve post-duplication stoichiometry or temporal, quantitative, or spatial subfunctionalization.

SYNOPSIS

Gene duplication plays a fundamental role in biological innovation but it is not clear how both copies of a duplicated gene manage to circumvent degradation by mutation if neither is unique. This study explores genetic mechanisms that could make each copy of a duplicate gene different, and therefore distinguishable and potentially preserved by natural selection. It is based on DNA sequences of the protein-coding region of 290 expressed duplicated genes in a frog, *Xenopus laevis*, that underwent complete duplication of its entire genome. Results provide evidence for multiple mechanisms acting within the same genome, within the same functional classes of genes, within the same period of time following duplication, and even on the same set of duplicated genes. Each copy of a duplicate gene may be subject to distinct evolutionary constraints and this could be associated with degradation or enhancement of function. Functional constraints of most of these duplicates, however, are not substantially different from a single copy gene; paralog persistence in the first dozens of millions of years after duplication may more frequently be explained by mechanisms acting on their expression patterns rather than on their protein function.

INTRODUCTION

By providing a redundant genetic template, gene duplication could relax purifying selection on one or both gene copies and facilitate functional divergence. Duplication catalyzes reproductive incompatibilities and speciation (Lynch and Conery 2000; Lynch and Force 2000; Taylor et al. 2001b), facilitates exon shuffling (Bailey et al. 2002) and microfunctionalization (Hancock 2005), buffers genetic pathways against null mutations (Gu et al. 2003), decreases pleiotropy (Carroll 2005), increases the diversity of gene expression (Gu et al. 2004; Li et al. 2005), and increases specialization of genes and genetic pathways. Duplicated genes exchange information through recombination, gene conversion, and epigenetic processes (Wendel 2000). However, unless natural selection favors the retained expression of both paralogs, mutations are generally expected to silence one gene copy soon after duplication (Haldane 1933; Kimura 1983). Duplication by polyploidization, for example, is accompanied by extensive and rapid genome restructuring and gene silencing; gene silencing is achieved in a variety of ways including mutations in the protein coding sequence or regulatory elements, and changes in methylation, histones, and chromatin structure (Liu and Wendel 2003; Osborn et al. 2003; Soltis and Soltis 1999). In order to retain expression of both copies, evolutionary mechanisms must therefore counteract or exploit mutation-induced degeneration.

Thus, the questions of how both paralogs retain expression, and how molecular evolution changes after duplication has captured the interest of evolutionary biologists.

Central to our understanding of the fate of gene duplicates are the questions of whether paralogs evolve differently from singletons, whether they evolve differently from each other, and whether their retained expression is more frequently triggered by mechanisms that add, compromise, or do not alter protein function (Figure 1.1; Force et al. 1999; Hughes and Hughes 1993; Kondrashov et al. 2002; Lynch and Conery 2000; Nembaware et al. 2002; Robinson-Rechavi and Laudet 2001). Molecular evolutionary analyses can be used to test the applicability of alternative explanations for the retained expression of duplicate genes that predict a unique molecular signature in the protein-coding portion in terms of the rates and locations of nonsynonymous substitutions, and the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (hereafter referred to as the *ka/ks* ratio).

Of course, these proposed mechanisms are not mutually exclusive because they could operate concurrently, on different parts of the genes, inside and/or outside of the coding region, and at different times after duplication. Moreover, if these mechanisms involve positive selection on one or both paralogs, their genetic signature will be difficult to detect in old duplicates if positive selection occurred soon after duplication, on only a portion of amino acid sites, or if it was followed by a long period of purifying selection. Other obstacles to dissecting out these mechanisms include variation in the rate of evolution over time, between lineages (Lynch and Conery 2000; Nembaware et al. 2002), functional classes of genes (Robinson-Rechavi and Laudet 2001), and genomic locations of each gene copy (Lercher et al. 2004), variation in the rate of gene duplication (Long and Thornton 2001), saturation of synonymous substitutions (Hughes et al. 1990), and mistaken identification of expressed duplicates that are actually pseudogenes or allelic variants. Nonetheless, if a particular mechanism operates for an extended period of time or on a large portion of the paralog(s), or if it involves a change in protein function or expression, it should be detectable by comparison to closely related orthologous singletons.

Clawed frogs (genera *Xenopus* and *Silurana*) offer a useful model system for exploring evolution of gene duplicates. Multiple species in this clade have undergone genome duplication via allopolyploidization and these polyploid genomes are primarily disomic in that each chromosome has only one homolog, as opposed to being polysomic, where multivalents form and recombination between paralogous loci is more prevalent (Evans et al. 2005; Kobel 1996; Osborn et al. 2003; Tymowska 1991). Extant tetraploids originated once in *Xenopus* and once in *Silurana* (Evans et al. 2005) and as a result, duplicate genes originating from tetraploidization in *Xenopus* are the same age. Detailed studies have been performed on hundreds of expressed duplicate genes (Table 1.1) and synonymous substitutions are generally not saturated (Taylor et al. 2001a).

A landmark study by Hughes and Hughes (1993) used this system to explore molecular evolution of 17 pairs of expressed gene duplicates in the tetraploid *Xenopus laevis*. They found evidence for an elevated *ka/ks* ratio after duplication, but still below the neutral expectation, and no evidence for a significantly different rate of non-synonymous substitution relative to single copy orthologs in mammalian outgroups. Their results are not consistent with neofunctionalization (Ohno 1970), wherein expression of duplicates is retained because one gene copy acquires novel function while the other carries out an ancestral function. Since this research was published, new mechanisms for duplicate gene retention have been proposed (Figure 1.1) and genomic sequences of the closely related diploid *Silurana tropicalis* have become available. In order to further evaluate these proposals, we have reanalyzed genes examined by Hughes and Hughes (1993) and also deployed new data, for a total of 290 gene duplicates expressed in the tetraploid *X. laevis* (Table 1.1).

RESULTS

The *ka/ks* Ratio, Expressed Paralogs in *X. laevis*, and Hypothesis Testing

Rates and types (nonsynonymous or synonymous) of substitution in the coding region are influenced by factors that are not directly linked to protein function, such as GC content, RNA secondary structure, and methylation (D'Onofrio et al. 1991; Fryxell and Moon 2004; Katz and Burge 2003), and also by factors that are related to protein activity but not unique to a particular function, such as level of expression (Drummond et al. 2005; Duret and Mouchiroud 2000; Pál et al. 2001; Subramanian and Kumar 2004). However, because nonsynonymous changes by definition affect the amino acid sequence of a protein, this class of substitution is more strongly affected by natural selection than synonymous substitutions. Evaluation of the *ka/ks* ratio therefore provides information on functional constraints on proteins, under the assumption that the effective population size does not change (Fay and Wu 2001; Fay and Wu 2003; Nei and Kumar 2000; Zhang et al. 2003).

Unfortunately this assumption is rarely met. If the *ka/ks* ratio of low frequency polymorphisms is different from the *ka/ks* ratio of fixed differences, demographic changes will alter the *ka/ks* ratio of fixed differences by changing the fixation probability of polymorphisms. This is not a problem when comparing the *ka/ks* ratio (or the rate of nonsynonymous substitution) between paralogs in the same species because they share the same demographic. However, unique demographic fluctuations could affect the *ka/ks* ratio of homologous genes in separate diploid and tetraploid species, even if selective constraints on proteins in these species were equal. For example, if mildly deleterious amino acid substitutions segregate at a low frequency, a reduction in population size of one species would increase the *ka/ks* ratio of fixed differences (Fay and Wu 2001). In comparing the *ka/ks* ratio of homologous genes in a diploid and a tetraploid species, we thus make the assumption that the effect of the unique demographic histories of each species is small compared to the effect of the unique selective constraints in these different types of genomes. In this study, we also do not have polymorphism information

with which to distinguish fixed and segregating differences, and we therefore make a second assumption that the observed differences between paralogs are fixed.

We identified 290 paralogs expressed in *X. laevis* by searching the literature and molecular databases for sequences expressed at the RNA and/or protein level (Table 1.1). Tetraploidization of this species probably occurred via allopolyploidization (Figure 1.2). Both paralogs were used to identify an *S. tropicalis* ortholog (JGI, assembly 3.0). Phylogenetic and phenetic methods were used to confirm that these sequences were paralogous rather than allelic and that they originated from tetraploidization of *X. laevis* as opposed to a separate gene duplication event. By comparing each pair of paralogs to closely related orthologs from *S. tropicalis*, we minimize the confounding effects of functional differences in the comparison. Because genes in a polyploid are simultaneously duplicated, we have standardized across all duplicates the impact of variation in the genome-wide rate of evolution over time after duplication.

We assigned mechanisms for duplicate gene retention to each of these paralogs based on three analyses that test specific predictions about the *ka/ks* ratio and the rate and location of nonsynonymous substitutions in their coding region after duplication (Figure 1.3). Analysis 1 tests whether the average *ka/ks* ratio in both paralogs increased after duplication and is consistent with diversifying selection, positive selection on a subset of sites, activity-reducing mutations, or relaxed purifying selection after duplication (which is probably a consequence rather than a cause of retained expression). Analysis 2 tests whether the nonsynonymous substitution rate differed between paralogs and is consistent with neofunctionalization. Because variation in evolutionary rate due to genomic location could influence rates of nonsynonymous substitution, for the second test we imposed the criterion that the *ka/ks* ratio of the paralog with the significantly higher rate of nonsynonymous substitution be higher than the *ka/ks* ratio of the other paralog and also higher than the *ka/ks* ratio of the singleton lineage, but we did not stipulate that the higher ratio be significantly higher. Analysis 3 tests whether the pattern of substitution in each paralog was complementary in that substitutions occurred in different parts of each paralog. This pattern of nonsynonymous substitutions is consistent with either complementary degeneration or enhancement of different protein functional domains. For each gene, we applied the sequential Bonferroni correction for these three tests (Rice 1989).

Diversifying Selection and/or Relaxed Purifying Selection in Both Paralogs

We compared the likelihood of a model with a higher *ka/ks* ratio after duplication (Model B in Figure 1.3) to a model with no change in the *ka/ks* ratio (Model A in Figure 1.3). Thirty-eight out of 290 of these paralogs (13%) have a significantly higher average *ka/ks* ratio than the diploid lineage (even though this ratio does not exceed neutral expectations), but based on other tests they have a similar rate of nonsynonymous substitution between paralogs and do not have a complementary pattern of nonsynonymous substitution. This difference is significant table-wide (Fisher's test; $P << 0.0001$).

Interestingly, the diploid lineages of the alpha and beta globin genes acquired nonsynonymous substitutions much faster than their paralogous lineages and also much faster than other genes (Tables 1.2 and 1.3). The *ka/ks* ratios over all sites of the diploid alpha and beta globin are near neutral expectations (0.799 and 1.068, respectively; Table 1.2).

Neofunctionalization

Under the protein neofunctionalization hypothesis, one paralog carries out the ancestral (pre-duplication) function and the other paralog acquires a useful novel function due to amino acid changes during a period of relaxed purifying selection. A prediction of neofunctionalization is that one paralog acquires nonsynonymous substitutions at a different rate than the other paralog and also faster than a homologous singleton. We tested a neofunctionalization model that has a different rate of nonsynonymous substitution on each paralog (Model C in Figure 1.3). This was compared to a null model with an equal rate of nonsynonymous substitution in each paralog (Model B in Figure 1.3). With the criterion that the faster paralog also have the highest *ka*/ks ratio, an individually significant difference in nonsynonymous substitution rates was achieved for 40 genes (Table 1.3) and this difference is significant table-wide (Fisher's test; P = 0.0004). This significant difference between nonsynonymous but not synonymous substitutions was also confirmed with an alternative statistical framework (see below). After correcting for multiple tests, 18 out of 290 of these paralogs (6%) are individually consistent with the neofunctionalization model and also do not have a complementary pattern of substitution.

An extreme scenario of neofunctionalization would involve one paralog remaining unchanged after duplication and the other paralog acquiring many substitutions. Interestingly, *X. laevis* paralogs of liver-type arginase have this genetic signature (Figure 1.4A). One paralog (X69820) incurred an in-frame deletion of one amino acid, an inframe insertion of one amino acid, a new stop codon that terminates the protein seven codons upstream from the other paralog, and 25 amino acid substitutions. The other paralog (BC043635) is identical to the maximum likelihood reconstruction of the ancestral sequence, although a maximum parsimony reconstruction of this ancestral sequence suggests that two synonymous substitutions occurred in this paralog. This paralog (BC043635) is similar in size to *S. tropicalis* and to outgroups such as humans and mice, indicating that the indels occurred in the other paralog (X69820). Of course, this pattern of substitution could also occur if the rapidly evolving paralog were a pseudogene. However in the case of liver-type arginase, polyclonal antibodies generated from protein translated from cDNA of the rapidly evolving paralog recognize two differently sized proteins in tadpole livers but only one size in adult liver (Xu et al. 1993). Although this would suggest expression of both paralogs at the protein level, cross hybridization to other genes or splice variants is a concern, and further studies are needed to confirm expression and translation of both of these paralogs.

Complementary Substitutions

Another way that the retained expression of paralogs could be promoted is if different functional domains on each paralog were enhanced or degraded (Force et al. 1999; Hughes 1994). These mechanisms predict a complementary pattern of substitution on each paralog, and this pattern is not expected if neofunctionalization or regulatory changes drive their retained expression. We tested this possibility with the paralog heterogeneity test (Dermitzakis and Clark 2001) and the runs test for dichotomous variables (Sokal and Rohlf 2003), after excluding genes with two or fewer substitutions in one or both paralogs. Using a more conservative null distribution than Dermitzakis and Clark (Dermitzakis and Clark 2001), the paralog heterogeneity test identified more clustered nonsynonymous substitutions than expected by chance, depending on the number of domains assumed (235 genes tested, 13 or 18 genes were significant under the assumption of two or three domains at $P < 0.05$, Table 1.4). The runs test identified more genes with runs of nonsynonymous mutations on the same paralog than expected by chance (235 tests, 19 significant at $P < 0.05$, Table 1.4). Some duplicates were identified by both tests, and a few of these genes appear to have complementary substitutions in positions that correspond to distinct functional domains (See below). We used the lowest p-value from both methods for Bonferroni correction across these analyses (Figure 1.3).

As a qualitative test for Type I error, we also performed these tests on synonymous substitutions because we would not expect this class of mutations to be more heterogeneous in duplicates than in singletons. When synonymous substitutions were analyzed, both tests identified more significantly complementary mutations than expected by chance. The paralog test identified 24 or 22 out of 286 genes tested under the assumption of two or three domains, and the runs test identified 22 genes ($P < 0.05$; Table 1.4). One explanation for this observation is that synonymous substitutions of some paralogs are complementary. Synonymous substitutions can, for example, be heterogeneous (Pond and Muse 2005). Another explanation is that, although these tests help target some candidates for retention by subfunction co-option or subfunction partition in the coding region, both may suffer from Type I error. In any case, tests for complementary nonsynonymous and synonymous substitutions are both significant table-wide ($P = 0.001$ and $P < 0.0001$, respectively)

According to these tests, eight out of 235 of these paralogs (~3%) exhibit a significant complementary pattern of nonsynonymous substitution. One of them, *fibroblast growth factor receptor* (FGFR), also had a significantly different rate of nonsynonymous substitution. This could be explained by differently sized functional domains being co-opted or degraded, or by a different number of domains being altered in each paralog. In FGFR, it is a combination of these possibilities (Figure 1.4B). This gene has three immunoglobulin domains that are roughly 70 amino acids long and one tyrosine kinase domain that is roughly 300 amino acids long. Four out of five substitutions in the first immunoglobulin domain are in one paralog (M55163) whereas the second immunoglobulin domain has five out of

seven unique mutations in the other paralog (U24491) plus one in the same position in both paralogs. Six out of six substitutions in an approximately 115 amino acid long region between the third immunoglobulin domain and the tyrosine kinase domain are in one paralog (U24491). The tyrosine kinase domain has a similar number of mutations in both paralogs (four or five), but their distribution differs in that each paralog has most of it substitutions in either the beginning or in the end of this domain.

Tests over Multiple Loci: Codon Bias, Evolutionary Rates, and Functional Categories

Codon bias affects the *ka/ks* ratio due to selection on synonymous sites and this bias could change after gene duplication, especially if it is linked to expression levels. However, we did not find a significant partial correlation between codon bias and the number of extra copies (zero or one) of the gene over all loci when the effect of the number of synonymous substitutions is held constant ($r = -0.0004$, $t_s = 0.0068$, $df = 287$, $P = 0.4973$) or over just the loci with a significantly higher average *ka/ks* ratio ($r = -0.0470$, $t_s = 0.3908$, $df = 68$, $P = 0.3481$). This indicates that the elevated *ka/ks* ratio after duplication in some paralogs cannot be attributed to increased selection on synonymous sites after duplication.

To further explore the null hypothesis of equal evolutionary rates in each paralog, we developed a method to use the equal mean Skellam distribution framework proposed by Lynch and Katju (2004) over multiple loci. The null hypothesis of this test is that the number of nonsynonymous substitutions on each paralog follows the same Poisson distribution (i.e. the paralogs have equal rates). We used permutations to derive a probability distribution for the difference in the number of substitutions observed between all paralogs, and performed simulations to evaluate whether the observed distribution was significantly different from the expected equal mean Skellam distribution. To minimize the impact of variation in evolutionary rate due to genomic location, we restricted our analysis to genes in which synonymous substitutions met Poisson expectations (i.e., that the mean number of substitutions equal the variance in the number of substitutions); 260 out of the 290 genes met this criterion (89%). As a conservative measure, we also excluded one gene (met *mesencephalon-olfactory transcription factor 1*) from this analysis because we suspect a sequencing error increased the number of nonsynonymous substitutions of one paralog (AF041138), causing a run of eight amino acid differences that could be eliminated by shifting the nucleotide alignment out of frame by one base pair.

This analysis confirms the results of the likelihood test for unequal rates of nonsynonymous substitution (Analysis 2). The set of genes with an individually significant difference in nonsynonymous rates according to Analysis 2 also have a significant departure from the equal mean Skellam distribution null hypothesis for nonsynonymous substitutions (36 genes were analyzed; $\lambda_{ML} = 543$, $P < 0.001$, Figure 1.5A) even though, as expected, synonymous substitutions of these genes were not significantly different ($\lambda_{ML} = 1039$, $P = 0.857$, Figure 1.5B). The other

genes did not have a significant departure from the equal means Skellam distribution null hypothesis for nonsynonymous (224 genes were analyzed; λ_{ML} = 2,737, P = 0.505, Figure 1.5C) or synonymous substitutions (λ_{ML} = 5,760, P = 0.124, Figure 1.5D). Thus, even after excluding loci with synonymous substitutions that do not meet Poisson expectations and also a locus with a potential sequencing error, these results strongly reject the null hypothesis of equal evolutionary rates in about 14% of these genes. The estimated percentage of genes consistent with neofunctionalization (6%) is lower because it is calculated in the context of multiple tests on each gene.

We also explored whether expression of paralogs of certain functional categories tends to be retained by a particular type of mechanism. Second-level gene ontology annotations from the three main categories (Biological Process, Molecular Function, and Cellular Component) were assigned to *X. laevis* paralogs based on the annotations (when available) of the most homologous hits that were obtained with the Gene Ontology Consortium Browser and BLAST tool (http://www.godatabase.org). After correcting for multiple tests, we did not find a significant overrepresentation of retention mechanisms in any of the functional classes based on a hypergeometric distribution performed with GeneMerge (Castillo-Davis and Hartl 2003). In contrast, we find that expression of paralogs within functional classes are consistent with a diversity of mechanisms.

Selective Constraints of Most Paralogs Are Not Significantly Different from an Orthologous Singleton.

In gene duplication by polyploidy, as opposed to by doubling of a single gene or a fragment of the genome, selection to maintain protein stoichiometry could play a prominent role in preserving both copies of a duplicate gene because entire genetic networks are duplicated. In a polyploid genome, spatial, quantitative, or temporal subfunctionalization of expression could also promote retained expression of duplicate genes. Under these hypotheses, functional constraints (and pleiotropic interactions) of both paralogs are similar and nonsynonymous substitutions would not be in complementary locations in each paralog because each one performs an identical function to their singleton ancestral gene (though perhaps within a marginalized expression domain).

In all paralogs the average *ka/ks* ratio over all sites is less than one, indicating that the impact of purifying selection after duplication is pervasive (Table 1.2). However, in 226 out of 290 genes (78%), the average *ka/ks* ratio was not significantly higher after duplication, neither paralog had a significantly higher rate of nonsynonymous substitution and higher *ka/ks* ratio than the orthologous diploid lineage, and there was not a significantly complementary pattern of nonsynonymous substitution in each paralog. The degree to which this estimate is inflated by Type II error is expected to vary from gene to gene depending on the power of each test, the amount of data, unique parameter values of the data (transition/transversion ratios, base frequencies, branch lengths), and the degree to which the data depart from the null hypothesis.

Age of *Xenopus* Paralogs

If tetraploidization occurred by allopolyploidization, paralogs of *X. laevis* co-evolved in the same genome for a period of time that is shorter than the duration of their divergence (Figure 1.2). Using a relaxed molecular clock calibrated with geological and fossil data, we estimated the divergence time of *Xenopus* paralogs based on portions of the RAG1 and the cytokine receptor 4 genes. To avoid the possibility that an accelerated rate of nonsynonymous substitution after duplication could affect our estimates, we included only synonymous substitutions at fixed amino acid positions and four-fold degenerate sites. This analysis indicates that divergence of *Silurana* and *Xenopus* occurred 53 million years ago (mya) with a 95% confidence interval (CI) of 40 – 80 mya. The age of the most recent common ancestor of the α and β paralogs (Node 1 in Figure 1.2), which corresponds to the diversification of the diploid ancestors of *Xenopus* tetraploids, is estimated to be 41 mya (CI 29 –66 mya). Diversification of *Xenopus* tetraploids (Node 3 in Figure 1.2) is estimated to be about 21 mya (95% c.i. 13 – 38 mya). We did not directly estimate the timing of allopolyploidization (Node 2 in Figure 1.2) because no extant descendant of the most recent diploid ancestor of *X. laevis* is known (Evans et al. 2005). Thus we have narrowed down the age of *Xenopus* genome duplication to between 21 and 41 mya, but with broad confidence limits for these upper and lower boundaries.

The estimated time of divergence of *Silurana* and *Xenopus* (~53 mya) and the estimated time of tetraploid divergence (between 21 mya) are less than corresponding estimates based on mitochondrial DNA (~64 mya and 42 mya, respectively) (42 mya; Evans et al. 2004). However, all of them are about twice as old as estimates based on immunological distances of antiserum to albumin (about 30 and 10 mya, respectively; Bisbee et al. 1977). We suspect that these immunological distances could underestimate divergence between sister tetraploid species because intraspecific divergence between expressed paralogs is similar to or greater than interspecific divergence between paralogs (Figure 1.2). Divergence between expressed paralogs in a tetraploid could also reduce immunological distances between a tetraploid species and a diploid species as compared to two similarly diverged diploid species. Another estimate of 110 million years for the divergence of *Silurana* and *Xenopus* (Knochel et al. 1986) is clearly an overestimate because it is based on globin proteins with an atypically rapid rate of evolution in diploid clawed frogs (Table 1.2).

DISCUSSION

Genome duplication provides an approximation of the assumption of initial redundancy made by some models for retained expression of gene duplicates (Nowak et al. 1997; Wagner 1999; Walsh 1995) because intact regulatory elements are duplicated with the coding region. However, many duplicates in diploid genomes are gene fragments or have incomplete regulatory elements (Katju and Lynch 2003), and extensive and rapid genome restructuring can also fragment

protein-coding and regulatory regions in polyploids (Blanc et al. 2000; Soltis and Soltis 1999). Initial population genetic dynamics of duplicates in polyploid genomes differ from those of duplicates in diploid genomes. In a diploid genome duplicates must become fixed, whereas in a polyploid genome duplicates must stay fixed. Selective pressures to maintain expression stoichiometry also differ in each system; duplication by polyploidy does not change stoichiometry, but singleton duplication does (Lynch and Conery 2000). Nonetheless, a recent comparison of expressed duplicates derived from whole genome duplication to paralogs from smaller scale duplication found that while the functional attributes differ between these types of expressed duplicates, molecular evolutionary changes are analogous (Davis and Petrov 2005).

Our results suggest that most of these paralogs do not have significantly different selective constraints from a diploid ortholog. The extent to which this is applies to duplicate genes in diploid species depends on how many of these *X. laevis* paralogs are expressed due to attributes unique to polyploids (such as selection to maintain the stoichiometry of expression in a duplicated genome) versus other mechanisms common to both types of genomes (such as quantitative, spatial, and regulatory subfunctionalization). Retained expression of duplicates in either type of genome might be favored, for example, if overexpression is advantageous (Kondrashov et al. 2002).

Increased *ka/ks* Ratio after Duplication

Other studies have reported a higher *ka/ks* ratio following duplication and the magnitude that this ratio increases differs among groups (Hughes 1994; Hughes and Hughes 1993; Jordan et al. 2004; Kondrashov et al. 2002; Li 1985; Lynch and Conery 2000; Nembaware et al. 2002), but see (Robinson-Rechavi and Laudet 2001). Conservative sites are more apt to change after duplication (Seoighe and Wolfe 1998) and a burst of nonsynonymous substitutions following duplication is suggested by comparison of young to old duplicates (Nembaware et al. 2002). This change is often attributed to relaxed purifying selection following duplication but could also be explained if some aspects of ancestral function disappear in both copies after duplication. The ability to self-dimerize, for example, is lost when a duplicated homodimer becomes a heterodimer. An increased tolerance of activity-reducing mutations in both paralogs could also occur such that the function of both is needed to recover the activity of the singleton ancestor (Stoltzfus 1999).

Interestingly, age discrepancies between the duplicates and the singletons could affect comparison of the *ka/ks* ratio (Nembaware et al. 2002). An unexpected positive correlation between *ka/ks* and *ks* was reported in comparisons between distantly related orthologs of some mammals, but a negative correlation exists between closely related mammalian comparisons (Wyckoff et al. 2005). In the closely related sequences, a negative correlation is expected as a result of stochastic sampling of synonymous substitutions at low mutation rates (Wyckoff et al. 2005). Consistent with this expectation, linear regression of data from clawed frogs

indicates a weak negative correlation between *ka/ks* and *ks* ($r^2 = 0.055$, unpublished data); this relationship is more obvious when data are binned (Figure 1.6).

The duration of divergence of *X. laevis* paralogs is twice their age, or about 82 million years. We estimate that the total divergence time of the diploid lineage (between node 1 and *S. tropicalis* in Figure 1.3) is about 75 million years. Because the ages of these lineages are similar, we expect that the effect of stochastic sampling of synonymous substitutions would also be similar (Wyckoff et al. 2005). However, some paralogs have a significantly higher *ka/ks* ratio, and many of them have a slightly higher *ka/ks* ratio after duplication even though the difference is not significant (Figure 1.7). Thus, these data provide strong evidence that some duplicates (~13%) evolve differently, averaged over both paralogs, than singletons, even though a significant change was not observed in the majority of these loci.

Asymmetric Evolutionary Rates

The neofunctionalization hypothesis for the retained expression of duplicate genes has been criticized because expression of duplicate genes is retained more frequently and for a longer time than expected if neofunctionalization is the principal mechanism for retention (Ahn and Tanksley 1993; Amores et al. 1998; Ferris and Whitt 1977; Nadeau and Sankoff 1997; Seoighe and Wolfe 1998). This hypothesis also lacks a known mechanism for sequestering beneficial mutations to only one of the two duplicate genes (Lynch and Katju 2004). However, after correcting for multiple tests on each gene, 6% of these paralogs have an asymmetric rate of nonsynonymous substitution, and a joint analysis of 14% of these paralogs also supports significant asymmetry, an observation that is consistent with neofunctionalization. One explanation for a different number of nonsynonymous substitutions in each paralog is that each diploid ancestor of *X. laevis* had a substantially different effective population size and that this introduced unequal levels of polymorphism in alternative paralogs of the allopolyploid ancestor of *X. laevis*. But this scenario is not supported by the data: paralogs with significantly different rates of nonsynonymous substitution do not have significantly different rates of synonymous substitution (Table 1.3, Figures 1.5A and B).

Other studies have found conflicting results with respect to whether paralogs have a different (Blanc and Wolfe 2004; Conant and Wagner 2003; Dermitzakis and Clark 2001; Van de Peer et al. 2001; Zhang et al. 2003) or have a generally similar (Kondrashov et al. 2002; Robinson-Rechavi and Laudet 2001) rate of nonsynonymous substitution. One way that asymmetry in nonsynonymous substitutions could be realized is via positive selection on one paralog. Accounts of positive selection have been found in many individual duplicated genes (Duda and Palumbi 1999; Hughes 1999; Van de Peer et al. 2001; Zhang et al. 1998; Zhang et al. 2002) but based on a branch-specific test over all sites, this study found only two out of 580 individual paralogs with a *ka/ks* ratio over one, and in both cases (*c-jun* and *deleted in colorectal cancer tumor suppressor*), the ratio was very close to one (Table 1.3). That this ratio is generally below neutral expectations suggests that neither copy is superfluous; selection is maintaining expression of both, either to

preserve advantageous unique functions or to preserve redundant functions. Because the *ka/ks* ratio over all sites is a conservative estimate of the frequency of positive selection we cannot rule out a role for site-specific positive selection in generating an asymmetric rate of nonsynonymous substitution in some paralogs. Comparison of the number of retained genes in species with different population sizes, for example, supports a role for positive selection in duplicate gene retention although it is not clear whether this is due to changes in amino acid sequences or regulation (Shiu et al. 2006).

Rates of nonsynonymous substitution are also correlated with levels of expression (Duret and Mouchiroud 2000; Pál et al. 2001; Subramanian and Kumar 2004) and one way that mutations could be sequestered to only one of the two paralogs is if regulation diverged prior to the accumulation of different numbers of substitutions. In *Saccharomyces cerevisiae*, for example, highly expressed paralogs evolve more slowly than paralogs with low expression levels (Drummond et al. 2005). Asymmetric rates could also be realized through enhancement or degradation of differently sized functional domains in each paralog. It will be interesting therefore, to combine these results with information on expression to further evaluate the role of neofunctionalization versus other mechanisms in promoting the retained expression of these paralogs.

Low Incidence of Complementary Replacement Substitutions.

Only 3% of these paralogs were identified with complementary patterns of substitution and this could be due to multiple factors. If retention is promoted by a small number of complementary mutations, mutations in the same (or nearby) positions, or splice variants, then the tests that we used would lack power. Additionally, amino acid substitutions in the diploid ancestors prior to allopolyploidization or near-neutral substitutions in either paralog after allopolyploidization could obscure an otherwise complementary pattern of substitution that occurred after allopolyploidization. That the estimated frequency of complementary substitutions is much lower in *X. laevis* than in paralogs shared by humans and mice (Dermitzakis and Clark 2001) suggests that subfunction specialization or degeneration in the coding region is more prevalent in much older expressed duplicates. Substantive changes in functional domains of each paralog may occur more commonly in older duplicates, for example after regulatory changes have occurred (Rastogi and Liberles 2005).

Retention of Genes with Overlapping or Redundant Functions

That we did not detect a significant change in the *ka/ks* ratio after duplication in most genes suggests that changes in functional constraints following duplication are small in many of them. The degree to which retained expression of most of these paralogs (78%), whose evolution was not significantly different from a singleton ortholog, is attributable to mechanisms that do not necessitate changed functional constraints depends on the level of Type II error of the tests that we deployed. Some paralogs have a lower *ka/ks* ratio than the diploid lineage and, under

assumptions and caveats discussed earlier, functional constraints of these loci either did not change or even became more extreme after duplication.

Many scenarios exist in which these paralogs could be retained without substantial changes in functional constraints, and our results indicate that this class of mechanisms is pervasive in the early stages of duplicate gene evolution. Selection to maintain stoichiometry, selection for over expression, and quantitative, temporal, or spatial subfunctionalization (Figure 1.1) preserve paralogs with identical function. Functional differences can be achieved by a small number of amino acid substitutions (Gibson and Spring 1998; Golding and Dean 1998) and a small number of activity-reducing substitutions could sufficiently impair function of both paralogs to the extent that both are required (Stoltzfus 1999).

Multiple Mechanisms

We have used a simple paradigm to associate duplicate genes to nonoverlapping categories of retention mechanism, although in reality there is reason to believe that a combination of factors may operate on a single duplicate copy. A functional study of *Saccharomyces cerevisiae* indicates that multiple mechanisms promote the retention of duplicate genes and that these mechanisms sometimes collaborate to promote retention of the same paralogs (Kuepfer et al. 2005). Evidence from *X. laevis* also supports this notion. Duplicated copies of the *estrogen receptor* α, ERα1 (AY310906) and ERα2 (AY310905), for example, exhibit signs of a combination of types of subfunctionalization in *X. laevis*: ERα2 is missing the N-terminal domain and splice variants of each paralog are expressed in different tissues (Wu et al. 2003). A combination of mechanisms is also suggested by *X. laevis embryonic fibroblast growth factor* (FGF4), which is a secreted protein with mesoderm-inducing activity: one *X. laevis* FGF4 paralog (X62594) has five out of six amino acid mutations in a hydrophobic signaling domain, part of which gets cleaved after expression, whereas the other paralog (X62593) has seven out of eight substitutions and a four amino acid deletion in a different domain that elicits fibroblast growth factor activities (Figure 1.4C; Isaacs et al. 1992). These paralogs also have divergent timing and stoichiometry of expression (Isaacs et al. 1992). Likewise, both nonsynonymous substitutions of one paralog (U05003) of the *FTZ-F1 related nuclear receptor* gene are in the "E domain III" which is involved in dimerization and transcriptional activation or suppression, whereas the other paralog (U05001) has eight substitutions including two in an otherwise highly conserved zinc finger-containing C-domain that is responsible for DNA binding and one substitution in the FTZ-F1 box (Figure 1.4D; Ellinger-Ziegelbauer et al. 1994). These paralogs are also differently expressed during embryogenesis (Ellinger-Ziegelbauer et al. 1994).

CONCLUSIONS

Thus, evolution of some paralogs (6%) is consistent with neofunctionalization in that they have different rates of nonsynonymous substitution

with one of them evolving faster than a singleton; this is most obviously suggested by substitutions in paralogs of *liver-type arginase*. There remains a lack of consensus regarding the significance of neofunctionalization (Conant and Wagner 2003; Dermitzakis and Clark 2001; Kondrashov et al. 2002; Robinson-Rechavi and Laudet 2001; Van de Peer et al. 2001; Zhang et al. 2003) and further characterization of the expression domains of these asymmetrically evolving paralogs is of interest. With the caveat that substantial functional transitions could be achieved by a small number of amino acid changes, complementary degeneration or enhancement of complementary protein functional domains appears rare in these relatively young paralogs (~30 million year old). Functional constraints on most of these paralogs are similar to singletons. Synthesis of molecular evolution and expression of these paralogs indicates that multiple mechanisms operate sequentially or concurrently to promote their expression within the same genome, in genes of the same functional class, and over the same period of time following duplication.

MATERIALS AND METHODS

Identification of Paralogs

We used multiple approaches to test whether *X. laevis* sequences were derived from genome duplication (tetraploidization) as opposed to another gene duplication event, and to test whether these sequences were paralogous rather than allelic. Outgroup sequence from another amphibian, a reptile, a mammal, or a fish were selected from Genbank in order to maximize the number of bases with unambiguous homology and phylogenetic proximity to clawed frogs. A rooted genealogy of the *X. laevis* paralogs, the *S. tropicalis* ortholog(s), and the outgroup was estimated using maximum likelihood with PAUP* (Swofford 2002) and a model of substitution selected with Modeltest version 3.06 (Posada and Crandall 1998). We included *X. laevis* paralogs that formed a clade with respect to the *S. tropicalis* ortholog(s), as expected because tetraploidization of *X. laevis* occurred after divergence of *Xenopus* and *Silurana* (Evans et al. 2005). We excluded genes that were duplicated in *Silurana* after the divergence of *Xenopus*. To explore the possibility that the sequences were actually allelic variants of one gene rather than alleles of separate paralogous genes, we compared the patristic distance between *X. laevis* paralogs to the average patristic distance between each paralog and the *S. tropicalis* ortholog. We applied a rule of thumb based on our estimates of the divergence times, that *X. laevis* paralogs should be at least one third as divergent from each other as they were from the *S. tropicalis* sequence. One possibility that we could not rule out is that duplication of one of the paralogs occurred in one *X. laevis* paralog after tetraploidization, which would result in more than two post-tetraploidization paralogs in *X. laevis*. However, we expect this possibility to comprise a small portion of the genes that we analyzed, and to not substantially compromise conclusions drawn regarding the impact of gene duplication in *X. laevis* relative to a singleton ortholog in *S. tropicalis*. We included all genes analyzed by Hughes and Hughes (1993) except calmodulin because our analyses suggested that

the sequences from this gene (accession numbers K01944 and K01945) are not paralogs derived from the tetraploidization of *X. laevis*.

We identified some expressed putative paralogs in which phylogenetic analysis did not reveal the expected relationship between the *Xenopus* paralogs and a closely related *S. tropicalis* ortholog, but instead provided weak support for an alternative relationship (Table 1.1), even though these genes had only one closely-related ortholog in the *S. tropicalis* genome, and the ratio of patristic distances was within our expectations. We used parametric bootstrapping (Goldman et al. 2000; Huelsenbeck et al. 1996) to test the null hypothesis that each of these genealogies is consistent with the expected topology depicted in Figs 2 and 3, and included those duplicates that did not reject this null hypothesis. For subsequent analyses without an outgroup, we sometimes included more data because homology within clawed frogs was unambiguous for all nucleotides.

Models for the Retained Expression of Duplicate Genes

We compared alternative models with different branch-specific *ka/ks* ratios and rates of nonsynonymous substitution using the codeml program of PAML, version 3.14 (Yang 1997). A model of codon evolution was assumed in which sites have a *ka/ks* ratio equal to zero, one, or another value estimated from the data, and the proportion of sites in each of these rate ratio categories is estimated from the data (model M2 in Yang et al. 2000). One ratio or rate was estimated over all sites and the transition/transversion ratio was estimated from the data. Equilibrium amino acid positions of the codon substitution model were calculated from the average nucleotide frequencies at each codon position. Significance of improvement in likelihood of the more parameterized model was assessed with a χ^2 test with degrees of freedom equal to the difference in the number of free parameters of each model. We performed five independent estimations of the maximum likelihood of each model for each gene.

The baseml program of PAML was used to perform marginal reconstruction of the sequences of the node ancestral to the *X. laevis* paralogs based on a model partitioned by each codon position with a different transition/transversion rate ratio, different base frequencies, and branch lengths proportional for each partition. The ancestral reconstruction and the extant sequences were used to estimate the number and positions of synonymous substitutions with DNAsp, version 4.0 (Rokas et al. 2003).

Codon Bias

If codon bias is positively correlated with expression levels, the rate of synonymous substitution would be underestimated to a greater degree in duplicated genes, and this could inflate estimates of the post-duplication *ka/ks* ratio (Li 1997; Shields et al. 1988). To explore this possibility, we compared the codon bias of each *X. laevis* paralogs to the codon bias of the *S. tropicalis* sequence and a maximum likelihood reconstruction of the sequence of the diploid ancestor of *X. laevis* (Node 1 on Figure 1.2). Codon bias of each pair of sequences was quantified with the scaled

χ^2 statistic (Shields et al. 1988) as calculated by DNAsp. Significance of the partial correlation coefficients between this estimate of codon bias and number of extra gene copies (zero or one) was assessed while holding constant the impact of the number of synonymous substitutions on the branches connecting each pair of sequences constant (Davis and Petrov 2004; Sokal and Rohlf 2003).

Equal Means Skellam Distribution

We used an approach described by Lynch and Katju (2004) to evaluate the null hypothesis of equal evolutionary rates. This test is based on a special instance of the Skellam distribution that describes the probability distribution of differences between two samples drawn from the same Poisson distribution (Irwin 1937). For each pair of duplicates, the number of sites that experienced a nonsynonymous substitution or a synonymous substitution since divergence was estimated by comparing each sequence to the reconstructed ancestral sequence. This is a conservative estimate of the magnitude of the difference in evolutionary rates because multiple substitutions in the same site are not counted.

In order to improve the statistical power for this test (Lynch and Katju 2004) we concatenated data from multiple loci into two "superparalogs". A randomly chosen paralog from each locus was concatenated into one of the superparalogs and the other paralog from each locus was concatenated into a second superparalog. The difference in the number of mutations in each superparalog (d_{SP}) was then calculated, and superparalog construction was pseudoreplicated for 10,000 iterations to generate a probability distribution of d_{SP}. Under the null hypothesis of equal rates of nonsynonymous substitution, this probability distribution approximates an equal mean Skellam distribution with the expected number of substitutions equal to the sum of the mean number of substitutions in each superparalog (λ_{SP}). This is the true because the sum of multiple Poisson distributions is a Poisson distribution with mean equal to the sum of the constituent distributions.

To evaluate significance, we compared the fit of the observed and simulated probability distributions of d_{SP} to the expected equal mean Skellam distribution. For each simulated locus, the number of mutations on each paralog was drawn from a Poisson distribution. The mean of this Poisson distribution was drawn from another Poisson distribution with a mean equal to the average number of substitutions at the locus being simulated. This approach accommodates uncertainty in the expected number of substitutions at each locus in the test, as well as stochastic sampling of the number of mutation from the distribution defined by this mean. Superparalogs were constructed out of the simulated paralogs and a probability distribution of d_{SP} was obtained in the same way as for the observed data. Fit of the observed and 1,000 simulated probability distributions relative to the expected equal mean Skellam distribution were compared with the χ^2 statistic; the median χ^2 value from nine iterations of the observed data was used as the test statistic.

Because variation in the nonsynonymous substitution rate could stem from different evolutionary rates in different genomic regions rather than different functional constraints at the amino acid level, we excluded from the equal means

Skellam test loci in which the number of synonymous substitutions in each paralog did not meet the Poisson expectations that the mean number of substitutions in each paralog equal the variance. This deviation could stem from variation in the genome-wide rate of evolution, sequencing errors, or other unknown factors, and this would confound efforts to test whether nonsynonymous substitutions have different evolutionary rates due to differential selection on nonsynonymous sites of each paralog. Substitutions in arginase, for example, are suggestive of different evolutionary rates that affect both classes of substitutions (Table 1.1), because one paralog has many nonsynonymous and synonymous substitutions whereas the other is identical to the reconstructed ancestral sequence (i.e. Node 1 in Figure 1.2). Significance of the departure of the variance in the number of synonymous substitutions from the Poisson expectation that it equal the mean was tested with a χ^2 test with Yates correction for small sample size, a d value for infinite degrees of freedom and a liberal rejection criterion ($\alpha = 0.20$). Using this criterion, we eliminated 32 of the 290 genes from this analysis (Table 1.1).

Complementary Substitutions in Each Paralog

We used two methods to test whether substitutions occurred in complementary locations in each paralog. The first method is the paralog heterogeneity test of Dermitzakis and Clark (2001), which was derived from a test for heterogeneous substitution in a singleton protein (Goss and Lewontin 1996; Tang and Lewontin 1999). We applied this approach to the two paralogs in *X. laevis* by considering new mutations in each paralog as opposed to variable mutations between paralogous pairs of orthologs (2001). Significance of the absolute differential of the longest region of different substitution heterogeneity between paralogs ("R" from Dermitzakis and Clark 2001) was assessed by comparison to a null distribution of absolute differentials. This distribution was generated from 1000 simulated paralogs with the same number of substitutions as the observed paralogs and the locations drawn from a permutated set of all observed variable sites in either paralog or in the diploid lineage. This is a more conservative approach than generating a null distribution of R values from a random assignment of mutations (Dermitzakis and Clark 2001) because it assumes that substitutions in a homologous singleton are also heterogeneous.

The second test we used is the runs test for dichotomous variables (Sokal and Rohlf 2003) which tests whether substitutions occur adjacently on the same paralog more frequently than expected by chance. Mutations on each paralog were ordered and converted to a string of binary variables to indicate whether they were on the α or β paralog. We assumed that mutations in the same position on both paralogs interrupt a run. Significance was estimated as the rank of the observed number of runs relative to the number of runs in 100,000 permutations. The paralog heterogeneity test and the runs test were performed only on paralogs that both had at least three mutations. Perl scripts that perform these tests are available by request.

Estimation of the Age of *X. laevis* Paralogs

The age of genome duplication in *Xenopus* (between Nodes 1 and 3 in Figure 1.2) was estimated from two nongapped sequences using a relaxed molecular clock with r8s version 1.7 (Sanderson 1997; Sanderson 2002). To minimize the impact of duplication on our estimates, we analyzed only synonymous substitutions at fixed amino acid positions and four-fold degenerate synonymous substitutions in pipoid frogs. We included 302 variable sites from RAG1 and 162 from chemokine receptor 4. Calibration points were obtained from fossil evidence (23.8 million years as a minimum age of *Xenopus* based on the derived morphology of the fossil species *Xenopus arabiensis* (Henrici and Báez 2001)) and geological evidence (112 million years for the separation of *Pipa* and *Hymenochirus* due to the separation of Africa and South America (Maisey 2000; McLoughlin 2001)). The maximum age of the root of the topology was limited to the age of the earliest frog fossil, 195 million years (Shubin and Jenkins 1995). For *cytochine receptor 4*, we assumed that an unidentified species (AY523691) is either *X. borealis*, *X. muelleri*, or *X. "new tetraploid"* (2004), thereby providing an estimate for Node 3 in Figure 1.2B. If this unknown species is actually another tetraploid with a closer relationship to *X. laevis*, the estimated time of this node would be younger than the actual age of Node 3 in Figure 1.2B. Confidence intervals were obtained by bootstrapping the data as in Evans et al. (2004); an appropriate outgroup (*Scaphiopus* or *Spea*) was used to root Pipoids and then pruned from the topology. Average dates and confidence intervals weighted by the number of variable sites analyzed from each gene are reported.

Accession Numbers

Most of the Genbank (http://www.ncbi.nlm.nih.gov/Genbank) accession numbers for the nucleotide sequences are listed in Table 1.1. The Genbank accession numbers for sequences specifically mentioned in the text are as follows: *calmodulin* (K01944) and (K01945); *cytochine receptor 4* (AY364174, AY523685, AY523691, AY523699, AY523701, BC044963, CR942369), and (Y17895); *estrogen receptor* α1 (AY310906); *estrogen receptor α2* (AY310905); *fibroblast growth factor receptor* (M55163) and (U24491); *embryonic fibroblast growth factor* (X62593) and (X62594); *FTZ-F1–related nuclear receptor* (U05001) and (U05003); *liver-type arginase* (BC043635) and (X69820); *RAG1* (AY874301), (AY874302), (AY874303), (AY874305), (AY874306), (AY874315), (AY874328), (AY874341), and (AY874357); and *transcription factor XCO2* (AF041138).

ACKNOWLEDGEMENTS

We thank E. Dermitzakis, J. Huelsenbeck, M. Lynch, J. Stone, and Z. Yang for advice on analysis, B. Tracey, M. Huntley, E. Wojcik and M. Zubari for assistance with sequence alignments, and D. Kelley, F. Kondrashov, B. Golding, A. Putnam, R. Morton, and four anonymous reviewers for discussions and comments. This research was supported by the Canadian Foundation for Innovation, the National Science and Engineering Research Council, the Ontario Research and Development Challenge Fund, and McMaster University.

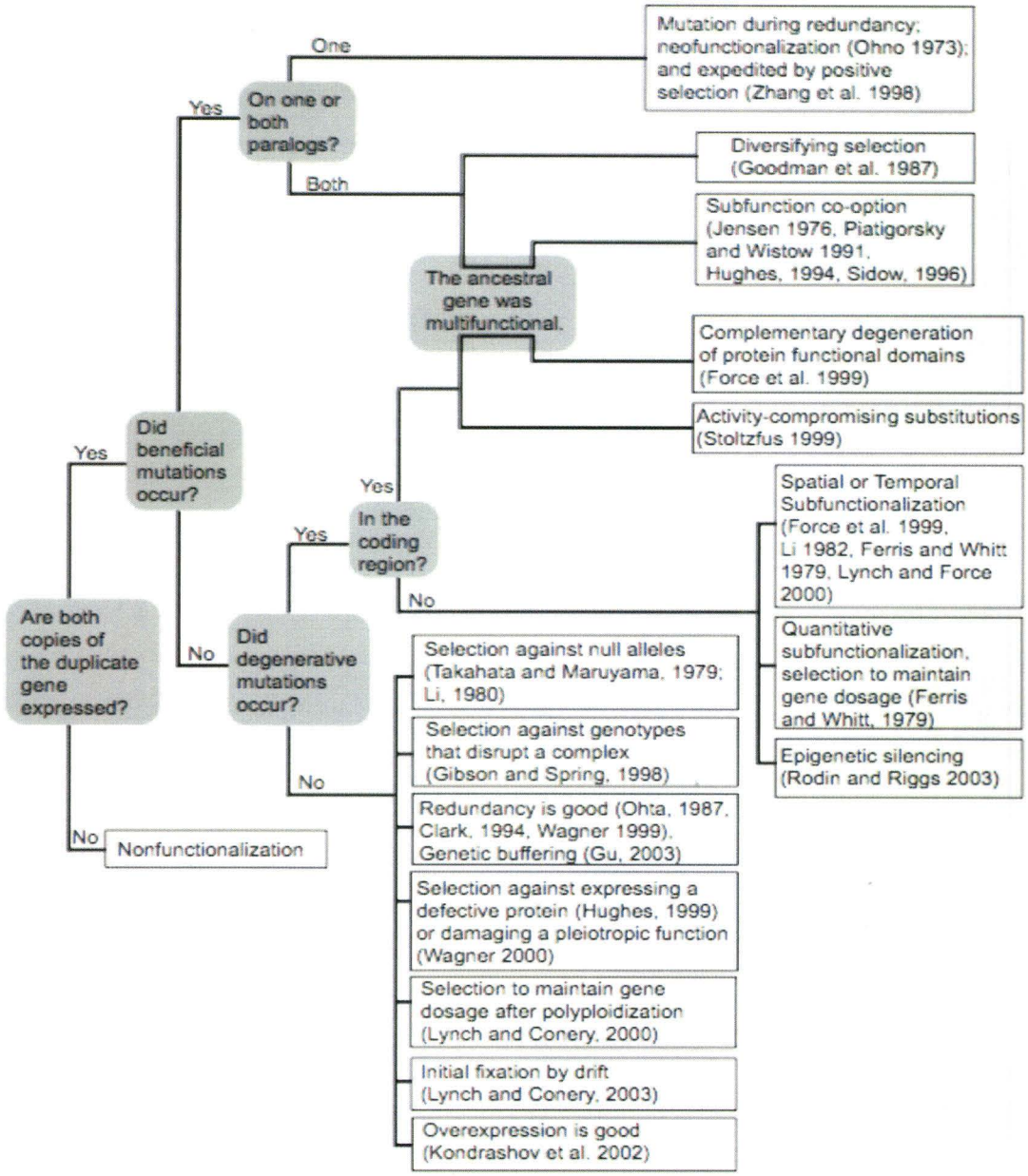


Figure 1.1. A non-exhaustive diagram relating various models for the fate of duplicate genes. Some citations that either propose mechanisms or discuss them are listed (Clark 1994; Ferris and Whitt 1979; Force et al. 1999; Gibson and Spring 1998; Goodman et al. 1987; Gu et al. 2003; Hughes 1994; Jensen 1976; Kondrashov et al. 2002; Li 1980; Li 1982; Lynch and Conery 2000; Lynch and Conery 2003; Lynch and Force 2000; Ohno 1973; Ohta 1987; Piatigorsky and Wistow 1991; Rodin and Riggs 2003; Sidow 1996; Stoltzfus 1999; Takahata and Maruyama 1979; Wagner 1999; Wagner 2000b; Zhang et al. 1998).

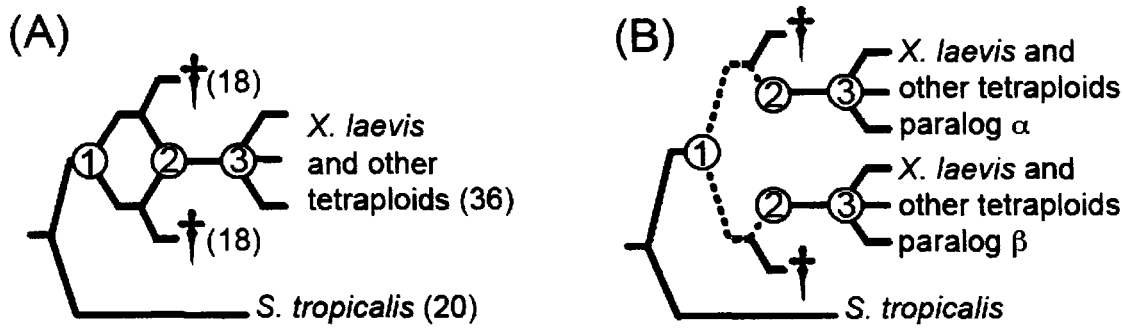
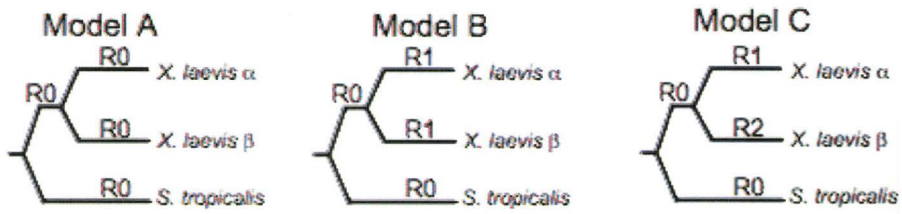


Figure 1.2. Putative allopolyploid evolution of the tetraploid *X. laevis*. Daggers indicate extinct diploid ancestors or genes. Nodes 1 and 2 correspond with the divergence and union, respectively, of two diploid genomes and Node 3 marks the diversification of *Xenopus* tetraploids. (A) A reticulate phylogeny with chromosome number in parentheses. (B) Nuclear genealogy assuming no recombination and no gene conversion between alleles at different paralogous loci (α and β). The dashed portion of the paralogous lineages in (B) evolved independently in different diploid ancestors.



Model A vs. B	Model B vs. C	Complementary substitutions?	Putative mechanism for retained expresson
-	-	-	Selection on regulation
+	-	-	Relaxed purifying selection*, diversifying selection, activity compromising mutations
+/-	+	-	Neofunctionalization
+/-	+/-	+	Enhancement or degradation of complementary protein functional domains

Figure 1.3. Assignment of putative retention mechanisms based on molecular changes in the coding region. We assigned a retention mechanism to paralogs based on the results of three analyses. The first one compared a model with no change in the *ka/ks* ratio after duplication (Model A where the *ka/ks* ratio on all branches is indicated by R0) to a model with a higher *ka/ks* ratio after duplication (Model B with *ka/ks* ratios R1 > R0). The second one compared a model with no difference in the nonsynonymous substitution rate (B, where R0 and R1 are nonsynonymous rates on each branch) to a model with different rates of synonymous substitution in each paralog (Model C where R0, R1, and R2 are nonsynonymous rates on each branch), with the stipulation that the faster paralog also have a higher *ka/ks* ratio than the slower paralog and also than the diploid lineage based on another test. The third analysis tested complementary pattern of amino acid substitution in each paralog. In the table, a minus sign indicates either no significant difference between the models or no significant complementary pattern of substitution. A plus sign indicates a significant improvement in likelihood of the more parameterized model or significant complementarity of substitutions in each paralog. An asterisk denotes the caveat that an increased substitution ratio could stem from relaxed purifying selection even though this is not a mechanism for retention.

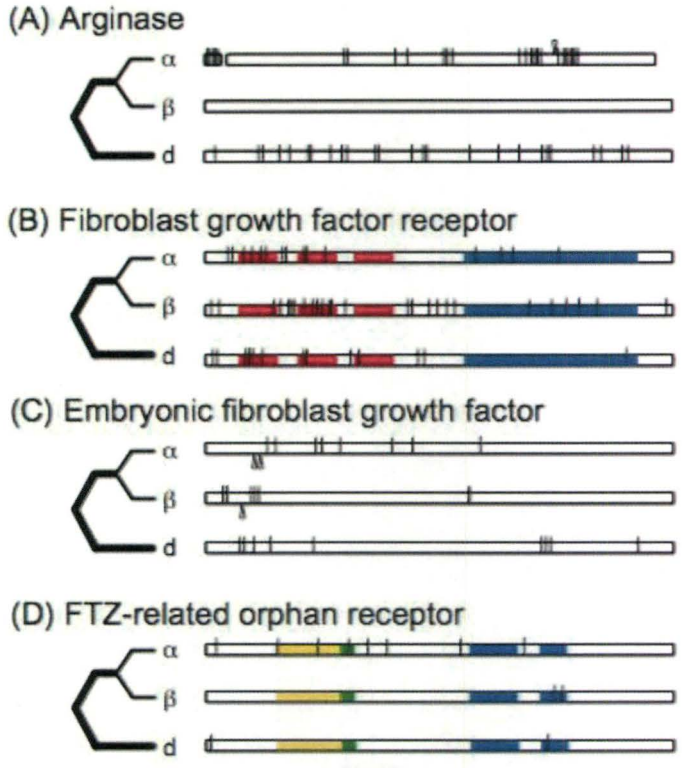


Figure 1.4. Nonsynonymous substitutions in each *X. laevis* paralog (α and β) and the diploid lineage (d) in representative genes. Substitutions in the diploid lineage occurred on the thick branches in the rooted topologies to the right of each locus. The length of each gene is arbitrary. (A) liver-type arginase, (B) fibroblast growth factor receptor, (C) embryonic fibroblast growth factor, and (D) FTZ-F1-related orphan receptor. In (A) a gap indicates a single amino acid deletion, an arrow above the paralog indicates a single amino acid insertion, and this paralog is shortened due to an early stop codon. In (B) three red boxes and a blue box indicate three immunoglobulin domains and a tyrosine kinase domain. In (C) arrows below the paralog indicate predicted cleavage sites in each paralog (Isaacs et al. 1992). In (D) yellow, green, and two light blue boxes indicate the DNA-binding C-domain, FTZ-F1 box, and DNA binding domain regions II and III (Ellinger-Ziegelbauer et al. 1994).

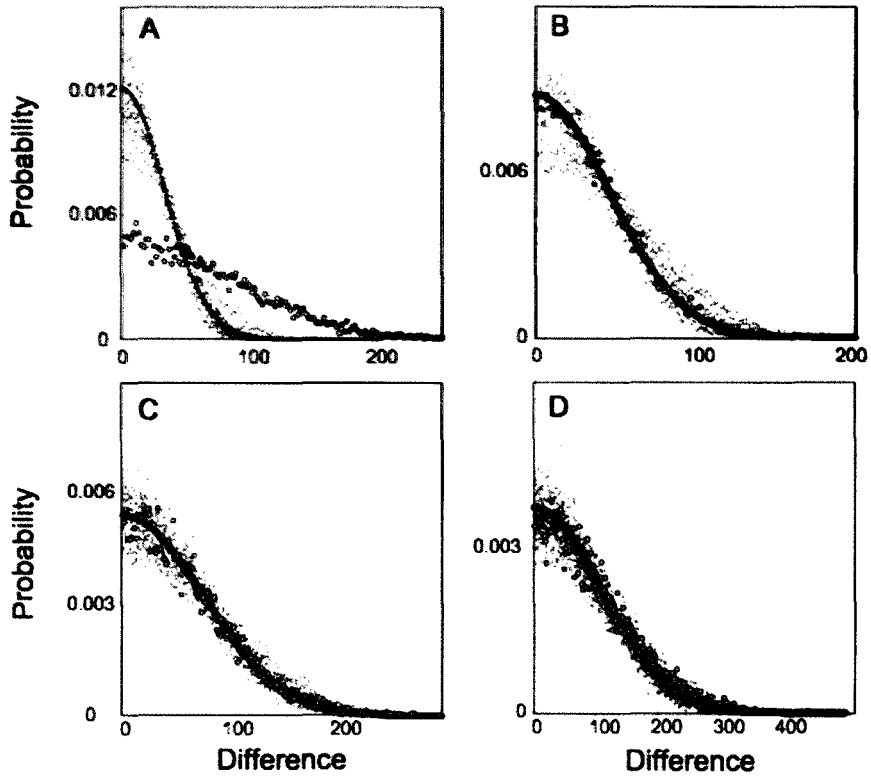


Figure 1.5. Probability versus distribution of the number of differences between superparalogs constructed from (A) nonsynonymous substitutions from paralogs identified by the likelihood analysis as having asymmetric rates of evolution, and (B) synonymous substitutions from these paralogs, and (C) nonsynonymous substitutions from the other paralogs that were not identified as having asymmetric rates and (D) synonymous substitutions from these paralogs. Black circles are the expected Skellam distribution, gray dots are d_{SP} distributions from 10 example simulations (out of 1000 total), and white circles are the observed distribution of superparalog differences.

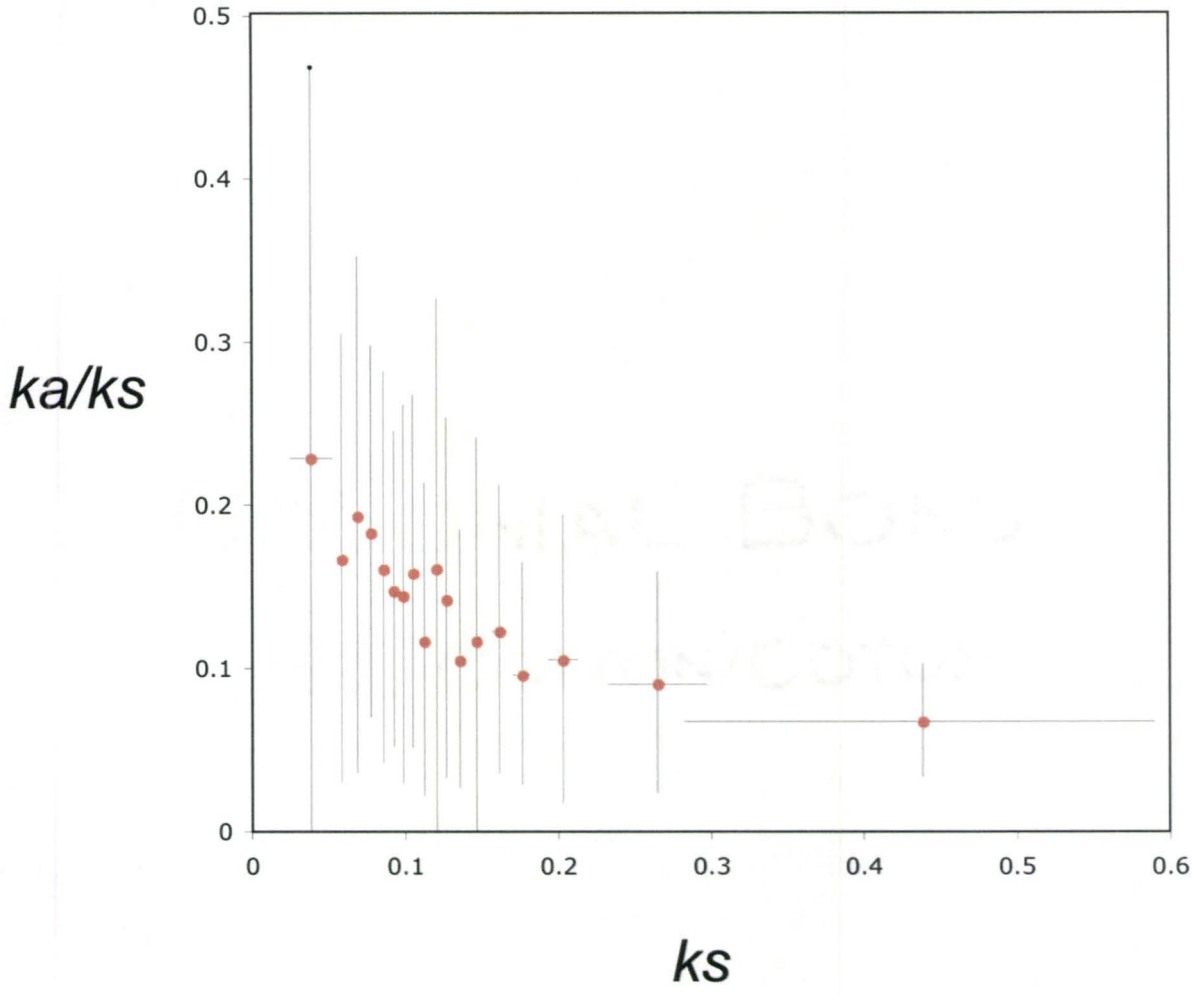


Figure 1.6. The relationship between *ka/ks* and *ks* corresponds with simulations that predict a negative relationship under neutral or near-neutral evolution of synonymous substitutions because of stochastic sampling of synonymous substitutions in slowly evolving genes (Wyckoff et al. 2005). The plot shows the average *ka/ks* ratio on each branch of 290 genealogies versus average *ks* of bins of 50 lineages ranked by *ks* of each one. The last bin has only 20 lineages. Bars indicate the standard deviation of each bin.

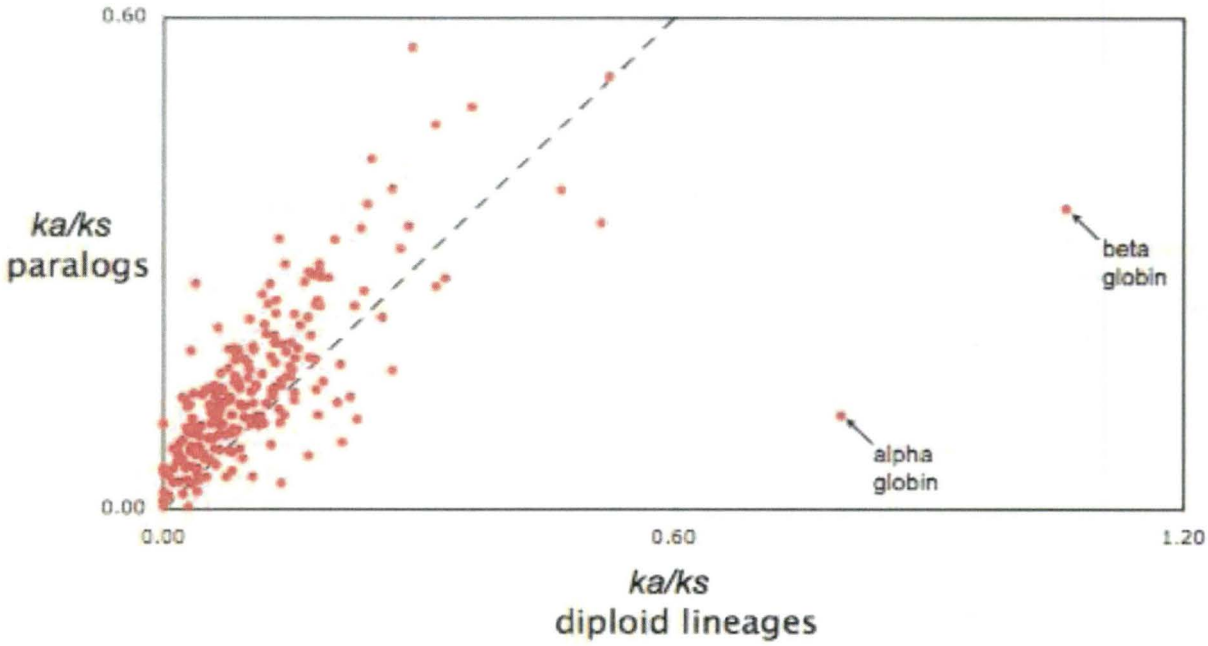


Figure 1.7. The *ka/ks* ratio is often slightly higher in the paralogs, even though it was not significantly higher than the diploid lineage. Only ratios from genes with no significant difference are shown (226 out of 292 genes). A diagonal line indicates no difference between the *ka/ks* ratio in either type of lineage.

Table 1.1: Information on genes including base pairs (bp) analyzed, accession numbers of *X. laevis* paralogs and outgroup, number of polymorphic nonsynonymous and synonymous sites on each paralog and the diploid lineage, and the expected (mean) number of polymorphic sites (lML). Note that the number of polymorphic sites is less than or equal to the number of substitutions because multiple substitutions could occur at a single site. Genes analyzed by Hughes and Hughes (1993) are indicated with an asterisk after the name. Genes which only partial coding sequences were analyzed are indicated with an asterisk after the bp analyzed. Accession numbers with an asterisk are *Silurana tropicalis* sequences obtained from Genbank instead of the genomic scaffolds. Genes not included in the equal means Skellam test due to overdispersion of synonymous substitutions have an asterisk after the synonymous (lML). If the expected rooted topology between *X. laevis* paralog α (XLA), β (XLB), *S. tropicalis*, and the outgroup was not estimated with phylogenetic analysis, we report the P values of a parametric bootstrap test of this null hypothesis; nonsignificant values indicate that the expected topology cannot be rejected by the data. Significance after Bonferroni correction for Analyses 1, 2 and 3 is identified identified by a "+"; P values for these tests are in Tables 1.2, 1.3, and 1.4.

Table 1.1 (continued 1)

Xenopus Gene Name	bp	Xenopus & Outgroup Accession#	replacement substitutions: α	β	Dip	λ_{ML}	synonymous substitutions: α	β	Dip	λ_{ML}	Bootstrap P value	Analysis number: 1	2	3
Actin (skeletal, alpha 3)*	1131	X12525 X03470 J00068 AB001073	0	0	2	0	23	15	24	19				
Activin Receptor-Like Kinase-2 (ALK-2)	1455	AF012245 AJ318064 BC077946	9	8	5	8.5	37	29	56	33				
Activin receptor II	1530	M88594 NM_204317 AF184090 BC082351	9	15	5	12	45	31	64	38		+		
Adipophilin (fatvg)	1194	XM_424822 X62771 M19032	22	15	18	18.5	30	28	39	29	0.105			
AE (Amidating Enzyme)	2610	XM_424857 M18350 M21442	24	30	26	27	71	71	83	71				
Albumin (serum)*	1809	U40452 AB016717	38	29	88	33.5	39	20	85	29.5*				
ALDH (Aldehyde dehydrogenase class1)	1503	AB016718 X58869 X14259 X14261	14	12	12	13	17	24	41	20.5				
Alpha Globin	426	M63797 AF095569 AF095570	3	6	23	4.5	12	8	15	10				
Amelogenin	489	AF095568 X60099 U28370	11	18	15	14.5	10	15	16	12.5				
Xenopus Anterior Neural Folds, Homeobox gene	555	XM_541834 AJ608932 AJ608933	11	10	13	10.5	14	20	27	17				
Amyloid-Beta-like protein precursor	2244	BC000373 BC043906 BC077529	30	13	10	21.5	39	43	56	41		+	+	
Apoptosis Inhibitor 5	1485	BC007133 U67129 X58955	6	11	6	8.5	39	46	78	42.5				
AR (Androgen Receptor)	471*	AY324231 BC043635 X69820	0	2	3	1	0	5	20	2.5*				
Liver L-arginase	945	D38303 U08406 U08408	0	25	25	12.5	0	19	64	9.5*		+	+	
Arginase Type 2	1080	AY074489 BC076815 BC072973	3	7	1	5	25	27	31	26				
Arrestin	1128	Z11501 BC042227 BC072839	10	7	6	8.5	32	34	58	33				
Aspartyl tRNA synthetase	1590	NM_001006528 U93170 U93171	7	4	9	5.5	39	35	90	37				
Atonal Homolog 5	414	AJ630209 AF187862 BC082702	9	4	5	6.5	7	10	11	8.5				
ATP synthase subunit B	750	AY522571 AJ243576 BC079987	9	7	9	8	13	29	33	21*				
Bambi (TGF-beta signalling)	780	BX934061 BC056095 BC068643	4	13	13	8.5	15	27	28	21				
Barren (brrn1, 13S condensin XCAP-H subunit)	2064	BC024211 AY273825 AY273826	23	24	23	23.5	51	59	76	55		+		
Bestrophin-2 (VMD2L1)	1530	NM_017682 J00978 BC071139	25	32	31	28.5	28	37	75	32.5				
Beta Globin	438	DQ091201 D29796 D49373	10	13	29	11.5	9	13	14	11				
Complement factor B (Bf B) (MHC class III gene)	2235	XM_532086	75	73	70	74	59	48	77	53.5				

Table 1.1 (continued 2)

Biglycan	1098	AB037269 BC074403 BT007323 AF224746	4	7	17	5.5	26	20	46	23			
Bicaudal-C	1908*	BC084957 XM_421490 BC046852	16	10	10	13	53	35	80	44			
Bridging integrator 1 (Amphiphysin II)	1389	BC073530 NM_001006524 BC078006	22	13	16	17.5	36	34	69	35			
Bing4	1632*	BC077337 AY648746 D32066	14	26	31	20	28	24	58	26			
BMP (Bone Morphogenetic Protein) receptor	1431*	D32067 BC028383 BC044074	6	7	10	6.5	35	33	44	34			
Block of proliferation 1	960*	BC056015 NM_015201 BC072031	5	12	28	8.5	15	11	61	13			
Brachyury (T)	1296	BC072224 AB001939 AB113401	6	12	9	9	28	30	51	29			
Serine/Threonine protein kinase (c-RMIL)	2217*	AY728383 X67052 U35408	11	17	19	14	43	63	104	53			
Basic transcription element binding protein	864	U35409 Y14296 AB028243	1	3	15	2	4	11	24	7.5			
B-cell translocation gene 1, anti-proliferative	507	BC041244 NM_205350 BC046715	8	6	3	7	5	8	15	6.5		+	+
Calcium homeostasis endoplasmic reticulum protein	2772	BC080042 NM_138585 BC044970	38	16	13	27	66	60	112	63		+	+
Calnexin	1800	BC041719 D78590 BC046257	28	18	37	23	29	40	72	34.5			
Calponin H3 (clpH3)	876	BC073467 CR391451 BC046699	5	6	8	5.5	12	29	39	20.5*			
Calreticulin	1236	BC044068 D78589 BC042287	10	11	20	10.5	41	28	44	34.5			
Carbonic anhydrase II	780	BC041213 BC004897 Y08817	17	7	18	12	13	14	33	13.5			
Casein kinase 1 alpha S (Csnk1a1)	1011	BC043956 U80822 BC077777	0	1	0	0.5	23	21	13	22			
CASK interacting protein 2	3609	BC092148 XM_420128 BC044689	60	47	68	53.5	87	113	155	100			
Procathepsin B	999	BC046667 XM_429301 BC077191	7	13	15	10	26	32	38	29			
Beta Catenin interacting protein 1 (catnbip1)	243	BC082479 BC014300 BC077504	0	1	0	0.5	6	2	3	4	0.06		
Cystathionine-beta-synthase	669	BC044091 NM_178224 U33217	4	3	5	3.5	19	22	38	20.5			
voltage-dependent Calcium channel beta subunit	1443	U33218 BC041811 M60680	6	15	8	10.5	23	28	44	25.5			
CDC2 (cell division cycle 2, kinase)	906	M60681 AF159158 AB080684	3	9	1	6	25	27	43	26		+	
Cathepsin E	1191	AB080685 AB093036 BC054271	10	14	15	12	28	33	40	30.5			
Carboxyl ester lipase	1419	BC054227 NM_009885 BC078053	13	9	16	11	47	38	82	42.5			
Carboxyl ester lipase (bile salt-stimulated lipase)	1641	BC085225 NM_001012997 U37538	20	22	33	21	42	56	97	49			
Centrin	516	BC054948 NM_019405 BC054217	2	1	4	1.5	11	8	12	9.5			
Cerebellin 2 precursor protein	645	BC074247 NM_172633 X59958	9	12	4	10.5	11	13	22	12		+	
Complement factor I (C3b/C4b inactivator)	1755	BC041753 Y18965 U60209	53	35	50	44	36	40	53	38			
Cystic fibrosis transmembrane conductance regulator	4407	X65256 AY026761 X82626	18	21	54	19.5	41	33	184	37		+	
Cortical granule lectin	939	BC085052 AY157364	15	22	26	18.5	26	40	44	33			

Table 1.1 (continued 3)

Choroideremia (Rab escort protein 1)	1884	BC061662 BC078011 NM_017067 BC056117	22	45	41	33.5	39	41	42	40	0.12	
Carbohydrate sulfotransferase 11 (Chst11)	993	BC089137 XM_414372 BC077471	7	4	6	5.5	22	20	31	21		+
Cell death-inducing DFFA-like effector c (CIDE-3alpha)	708	BC082372 AY364640 AJ243955	10	9	14	9.5	15	22	33	18.5		
C-Jun proto-oncogene (AP-1, Activator Protein)	936	AJ243954 X15547 BC075171	12	13	14	12.5	5	21	37	13*		+
Dipeptidase 2 (metallopeptidase M20 family)	1422	BC056069 NM_001006385 M63595 M63596	17	10	12	13.5	31	24	52	27.5		
alpha-1 Collagen type II	4458	AB022046 AY057997	22	39	30	30.5	48	67	86	57.5		
Connexin 31 (Gap junction beta-3 protein)	801	BC072821 NM_001009780 D86505	11	7	4	9	20	14	30	17		+
Contactin/F3/F11 (Contactin A)	3015	AB015205 NM_001004381 AB025245 AB025246	39	43	30	41	56	62	95	59		+
Coronin	1440	AK223073 AB027611 BC073454	16	15	15	15.5	33	31	42	32		
Cortactin	1590	BC011434 U14169	18	15	26	16.5	46	39	69	42.5		
Cytoplasmic polyadenylation element binding protein	1704	BC077702 AF329403 AY049034 AY049035	13	7	6	10	38	36	52	37		
CRY2 (cryptochrome 2)	1503*	AY256684 BC053794 BC082450	22	18	22	20	46	59	89	52.5		
Crystallin, beta A1	633	X87759 BC056059 BC077285	11	6	5	8.5	8	6	11	~		
Cathepsin S (CTSS)	999	M90696 BC077239 BC073186	16	17	25	16.5	14	23	37	18.5		
Cullin3 (Cul3)	2304	AF129738 L23857 L43513	1	2	0	1.5	46	56	65	51		
CyclinE	1224	U28981 X59500 X72902	22	14	14	18	31	28	27	29.5	0.22	
Brain Dopamine receptor D2	1035*	X17458 BC077380 AF393622	9	13	6	11	24	19	18	21.5		
Dapper 1, antagonist of beta-catenin (Frodo)	2454	NM_016651 BC077360 BC070744	33	51	23	42	52	89	74	70.5*		+
Death-associated protein kinase 1	4281	X76104 BC044296 BC074277	19	33	16	26	77	87	133	82		
Drebrin-like	1125	NM_013810 BC046698 BC073223	21	22	17	21.5	29	15	40	22		
Debranching enzyme homolog 1	1596	NM_031403 U10986 U10954	26	14	16	20	44	34	59	39		
Deleted in colorectal cancer tumor suppressor	486*	X76132 X16842 BC077922	5	14	19	9.5	6	7	26	6.5		+
Desmin	1374	AB011672 AF286645 AF286646	13	18	10	15.5	33	26	66	29.5		+
Hand2	396*	NM_133803 AF317841 BC045030	1	5	0	3	6	8	10	~		+
Cytoplasmic dynein light-intermediate chain 1 (DLIC1)	1524	AK222653 BC060495 BC077244	11	12	8	11.5	29	29	36	29		
Dipeptidylpeptidase 3	2187	NM_133803 AB084264 BC082639	36	33	36	34.5	68	37	80	52.5*		
Dullard	732	XM_536616 BC046260 BC073500	0	2	2	1	15	7	48	11		
Dystroglycan (DAG1)	2655	XM_343483 X99700 BC082429	24	16	35	20	60	55	68	57.5		
Dystrophin	1761*	NM_004007 U25959 U25960	4	0	5	2	11	0	57	5.5*		
Helix-loop-helix transcription factor XE1	441	MB3233	4	7	3	5.5	7	9	15	8		

Table 1.1 (continued 4)

E2 (transcription factor E2)	1962	BC044064 X66959 NM 003200 AF040993	25	8	15	16.5	42	27	54	34.5			+
met-mesencephalon-olfactory transcription factor 1 (Ebf2)	1713	AF041138 U71189 BC041714	11	19	2	9.5	25	37	58	31		+	
CCAAT/enhancer binding protein (C/EBP), alpha	891	BC072168 U08604 BC044316	7	19	21	15	13	25	35	19			
Endothelin receptor type A	1242	U06633 NM 204119 X52976	14	9	12	11.5	24	44	46	34*			
EF (Elongation Factor-1 alpha, 42Sp48)	1362*	X52977 AB124568 Z17206 Z17207	2	7	2	4.5	32	39	52	35.5			
Aurora kinase A (EG2)	1218	BC005425 X62973 X62974	8	16	29	12	37	43	69	40			
Engrailed 2 (EN2)	789	NM 010134 X00852 X00853	3	8	18	5.5	14	15	23	14.5			
Enkephalin A (proenkephalin A)*	672	AF195886 BC054169 Y00718	8	4	4	6	8	10	19	9			
ENO (2-phosphoglycerate dehydratase, enolase)	1302	D37900 AY310906 AY310905	0	6	24	3	37	26	55	31.5			
Era (Estrogen Receptor alpha)	1281	AB055221 BC077170 AF289027	3	8	8	5.5	17	32	37	24.5*			
Enhancer of split groucho	2268	AF186092 AF351126 BC097526	0	5	9	2.5	34	33	92	33.5		+	
Enhancer of zeste	2244	XM 418879 L33920 BC071046	8	19	2	13.5	44	53	46	48.5			
Focal adhesion kinase	3147	M86656 X64759 X64760	10	20	20	15	45	76	128	60.5*			+
Transcription factor (XLFB1)	321*	NM 000545 AJ249225 AJ249224	4	5	1	4.5	4	8	4	6		+	
XFD-4	1377	Y08223 AF036327 U68141	13	16	8	14.5	14	29	41	21.5*			
Flap endonuclease-1	1146	X76771 BC043891 BC078490	9	9	14	9	28	46	52	37*			
FetuinB	1326	XM 422765 U05001 U05003	36	43	57	39.5	28	41	35	34.5			
Ftz-F1-related orphan receptor (xFF1r)	1392	AB035498 X62593 X62594	8	2	2	5	23	23	36	23			
FGF (embryonic fibroblast growth factor 4)	561	U76998 M55163 U24491	8	6	10	7	14	4	13	9			+
Fibroblast growth factor receptor	2436	X59380 BC041754 BC073502	15	27	13	21	49	54	93	51.5		+	+ +
Fibrinogen alpha	2244	U44830 AF545659 AF545661	26	52	46	39	43	38	60	40.5			+
Flotillin	1254	NM 005803 BC044264 BC056023	16	17	13	16.5	25	29	45	27	0.125		
fms-related tyrosine kinase 1	2043	XM 534520 BC045075 BC076720	50	48	53	49	36	55	73	45.5			
Fms-interacting protein (NF2/meningioma region)	2034	NM 001006195 X06041 X06042	25	19	14	22	42	51	91	46.5		+	
alpha-fodrin (Xen alpha 1) spectrin, non-erythrocytic 1	705*	BC034956 AJ224511 BC079689	6	8	7	7	20	11	31	15.5			
c-fos proto-oncogene	261	X06769 BC047261 BC060446	2	5	4	3.5	3	5	7	4			
Succinate dehydrogenase complex, Flavoprotein	1995	XM 419054 U27274 BC072084	18	13	18	15.5	33	44	60	38.5			
Frequenin	570	X84048 BC077847 BC097600	1	1	9	1	16	14	23	15			
Fascin	1446	AY044229 M80471 AY901983	18	8	19	13	50	31	90	40.5			
Funn*	1779	X54056	7	14	17	10.5	35	37	65	36			

Table 1.1 (continued 5)

Fused toes homolog	876	BC077318 BC084320 NM_001005838	4	8	4	6	16	21	29	18.5	0.5	
FYN (proto-oncogene c-fyn)	1608	BC043749 M27502 NM_002037 AB060970	2	2	2	2	23	27	53	25		
Galectin	858	BC081109 X79303 L05540	14	16	14	15	18	19	34	18.5		+
alpha subunit of Gq Gtp-binding protein (G protein)	1077	BC081126 U43083 M76566	3	1	5	2	14	13	22	13.5		
GATA-binding protein transcription factor GATA-1	1077	M76563 NM_002049 L13701 L13702	20	23	29	21.5	22	35	47	28.5		
Transcription factor xGata5	1155	NM_205421 AF193797 X14601	11	6	15	8.5	26	21	68	23.5		
Growth hormone A	417*	M33697 BC078071 BC077482	2	8	4	5	11	11	20	11		
Guanylate kinase 1	594	XM_425960 BC077236 AF419148	10	5	14	7.5	7	12	34	9.5		
Glycogenin 1 (mitotic phosphoprotein 45)	924	NM_001006558 BC045005 BC078076	12	16	19	14	23	36	31	29.5		
Holocytochrome c synthase (heme-lyase) (hccs-prov)	897	AJ851811 U26349 U26350	11	27	19	19	18	15	26	16.5		+
cephalic Hedgehog, sonic hedgehog protein 4	1188	NM_009170 U94992 BC084434	13	19	17	16	23	29	36	26		
Transcription factor XHEN1	384	XM_424510 AY189821 BC043769	3	3	6	3	10	7	24	8.5		+
Hypoxia-inducible factor 1 alpha	1461*	NM_204297 AF068847 BC072816	17	10	7	13.5	32	23	43	27.5		
SafA - scaffold attachment factor A	2313	XM_419539 X06592 M23916	41	26	34	33.5	71	57	104	64		
Homeobox 2/2.3*	474*	M16937 M24442 M24443	5	3	8	4	9	3	19	6		
Insulin*	318	V00565 M20140 M20180	3	5	1	4	5	6	9	5.5		
Integrin beta-1 subunit*	2394	X07979 AF321228 AF321229	9	10	17	9.5	58	74	92	66		
Inversin	2982	NM_204551 BC079778 AY376839	73	58	91	65.5	80	74	139	77		
Ubiquitin carboxyl-terminal hydrolase 5 (Isopeptidase T)	2562	U47927 AB083246 AB083247	19	16	15	17.5	44	57	84	50.5		
Kf-1 protein (Adgr34)	2001	AF306394 BC061947 Z48770	24	26	23	25	41	53	79	47		
Kit receptor tyrosine kinase (c-kit)	2739*	X06182 X94082 BC071083	55	52	56	53.5	76	42	108	59*		
Kinesin-like protein 2	2712	AB035898 X05216 X05217	25	23	33	24	45	54	76	49.5		
L1 (ribosomal protein L1)	1173	NM_001007479 X06222 X06223	7	1	5	4	24	42	36	33*		
L14 (ribosomal protein L14)	564	BC021743 AF077838 X06344	4	5	6	4.5	9	12	12	10.5		
Lamin B	1749	M34458 Y17861 AF048817	23	25	28	24	42	53	66	47.5		
Lamina associated polypeptide 2	1548	BC053675 BC077312 BC060412	31	22	45	26.5	33	20	43	26.5		
Clathrin, light polypeptide (Lcb)	615	AJ720113 U07179 U07176	9	6	3	7.5	17	15	19	16	0.32	
Lactate dehydrogenase	1002	XM_534868 AF287147 AF287148	11	9	24	10	32	18	41	25		+
LEF-1 (lymphoid enhancer factor)	1116	AF288571 AF283562 AF283563	7	8	2	7.5	15	14	27	14.5		
TGF-beta family member Lefty-A	1032*	NM_130960	11	6	29	8.5	25	18	63	21.5		

Table 1.1 (continued 6)

LIM domain binding protein	1125	U74360 BC044043 BC013624 BC074438 BC074439	0	1	0	0.5	16	18	14	17	0.115		
Lipocalin (Ptgds)	447*	L06806 AJ249843 AJ249844	7	3	13	5	7	6	21	6.5			
Lpa1R (lysophosphatidic acid receptor)	1098	U70622 BC053822 BC078594	4	2	0	3	16	15	26	15.5		+	
LR (Leptin Receptor)	393	BC056250 AF276084 AF508961	3	2	1	2.5	9	6	14	7.5			
Lipoprotein (LDL) receptor-related protein 6	636*	AF074265 X68817 X68818	2	2	5	2	19	14	44	16.5			
Autoantigen La (La protein)	1275	BC081780 BC044036 BC090193	21	23	34	22	39	30	48	34.5			
Microfibrillar-associated protein 1	1320	BC050742 AF187864 BC084888	9	9	7	9	28	36	57	32			
Myristoylated alanine-rich C kinase substrate	561	NM_205480 L09738 L09739	20	9	14	14.5	16	13	23	14.5			
XMax2 and XMax4	381	NM_001009866 AY046531 AY046532	1	0	1	0.5	6	3	4	4.5			
Myogenin	705	XM_547345 BC042928 BC054224	4	6	1	5	12	15	16	13.5			
Myozenin1	933	XM_573691 M25696 M76710	22	15	12	18.5	20	23	21	21.5			
N-CAM (neural cell adhesion molecule)*	3252	D85084 U85969 U85970	51	30	46	40.5	59	51	75	55			+
NF-M1 (middle molecular neurofilament)	2619	NM_005382 U67778 U67779	31	35	38	33	52	42	89	47			
neurogenin-related 1 (X-NGNR-1)	594	NM_204796 BC077476 M86653	11	14	21	12.5	15	20	31	17.5			
Internexin neuronal intermediate filament protein	1380	NM_199534 BC047968 AF127225	21	13	18	17	33	23	58	28		+	+
NK3 transcription factor related, koza	633	BC074863 BC057729 AF055895	16	16	22	16	11	14	25	12.5			
Nonmuscle myosin II heavy chain A	2655*	NM_022410 L09740 L11231	14	11	17	12.5	48	54	80	51			
Nonmuscle myosin heavy chain B	1062*	NM_175260 X05496 X56039	10	1	9	5.5	22	29	56	25.5			+
Nucleolar-localized protein NO38	882	NM_205267 BC041205 BC068842	9	22	18	15.5	22	22	29	22			
Nucleobindin 1	630*	AF450266 BC068668 X04766	2	4	1	3	9	20	11	14.5*	0.55		
Nucleoplasmin	585	BC068078 AJ617672 AJ617673	5	7	12	6	14	17	19	15.5			
Nucleoporin (Nup88)	2178	Y08612 AY188504 AY188503	19	34	36	26.5	54	58	84	56			
OLPA (Dorphin)	696*	AY188505 AJ010978 AJ010979	16	19	17	17.5	16	20	47	18		+	
Olfactory marker protein (XOMP)	474	NM_011010 X52692 BC075161 (used instead of X52691)	8	6	16	7	13	7	29	10			
OncogenesC-ets-1 (c-ets-1 proto-oncogene)*	1311	J04101 M81683 X52635	5	4	5	4.5	29	37	29	33			
OncogenesC-ets-2 (c-ets-2 proto-oncogene)*	1392*	J04102 BC041189 X56870	15	13	30	14	35	48	67	41.5			
OncogenesC-myc (myelocytomatosis)*	1263	V00568 BC044069 BC081081	11	13	10	12	30	20	54	25			
Dynactin 2 (p50)	1206	AK222693 AJ277159 BC085213	9	7	10	8	42	27	37	34.5			
PACSIN2	1026*	CR456536	19	11	15	15	26	22	57	24		+	

Table 1.1 (continued 7)

		X66493											
		BC074270											
Convertase PC2	1842*	AB105176	8	12	8	10	30	54	70	42*		+	
		AF239760											
		BC077181											
Prolyl isomerase (Pin1)	477	BX950453	3	3	5	4	9	15	18	12			
		AB109739											
		AB109740											
PKC (protein kinase C,delta)	2049	BC043350	17	20	19	18.5	53	38	84	45.5			
		M95593											
		BC068787											
Plakoglobin	2205	Z68228	11	12	15	11.5	57	59	111	58			
		BC077880											
		BC041727											
Peripheral myelin protein 22	474	NM 008885	5	2	9	3.5	13	11	23	12			
		X03843											
		X03844											
POMC (pro-opiomelanocortin)*	777	AF115251	10	11	8	10.5	12	32	42	22*			
		X59056											
		BC093556										+	+
POU domain Gene 1	1068	NM 131161	11	2	2	6.5	17	13	30	15			
		X64835											
		X96423											
POU3	840*	XM 539052	7	1	7	4	10	9	19	9.5			
		BC043624											
Phosphorylase phosphatase (Ppp2B)	1767	BC073612	7	1	1	4	26	49	64	37.5*			
		CR860766											
		BC077952											
Protein phosphatase 4, regulatory subunit 2 (Ppp4r2)	1191	BC088694	26	31	22	28.5	21	43	37	32*			
		NM 174907											
		AF387815											
		AY055473											
LIM protein Prickle	2484	XM 416036	25	17	21	21	42	51	76	46.5			
		AF193800											
		AF193801											
Prolactin Receptor	1821	AB158367	38	42	37	40	41	54	66	47.5			
		BC054174											
		BC045213											
Prothymosin	324	AJ312835	7	6	1	6.5	6	8	5	7			
		BC047245											
		BC072163											
Phosphorylase, glycogen; brain	2529*	BC030795	18	15	14	16.5	45	74	96	59.5*			
		BC074233											
RAB18 (member RAS oncogene family)	615	BC056054	2	6	2	4	13	15	18	14			
		AY357728											
		AF174644											
		BC092101											
Rac GTPase	576	AY279384	1	2	0	1.5	10	13	18	11.5			
		D38488											
		D38489											
Rad51	1008	AB020740	4	2	2	3	29	32	49	30.5			
		AY874341											
		AY874315											
Rag-1	1134*	AY874303	8	10	13	9	28	25	55	26.5			
		AJ304845											
Ral interacting protein (rlip) - RalA (RalB-binding protein)	1833*	AJ252165	5	16	11	10.5	30	39	43	34.5	0.53		+
		AB209924											
		Y16259											
		BC072046											
RalB	570	X15015	2	0	8	1	18	12	61	15			
		X87365											
		L11445										+	+
Retinoic acid receptor alpha	1185	X73972	5	18	8	11.5	37	16	71	26.5*			
		L79913											
RDS35 (retinal degradation slow/peripherin)	1035	L79914											
		AF031238	11	23	9	17	19	33	28	26	0.16		
		L79915											
		BC054965										+	+
RDS38/peripherin	972	J02884	19	4	9	11.5	20	18	44	19			
		AB021737											
		BC082478											
Requiem	1155	XM 341998	10	11	8	10.5	16	33	32	24.5*			
		BC054145											
		L04692											
Rhodopsin	1062	U59922	1	5	13	8	6	5	25	5.5			
		AJ133499											
		AJ133500											
Ringo (p33 ringo, ls26)	888	XM 128768	20	11	15	15.5	33	13	31	23*			
		BC077472											
		BC084165											+
RIO kinase 2	1626	NM 001006581	28	28	23	28	40	36	52	38			
		BC080086											
		BC073326											
Rwdd1 (RWD domain containing 1)	717	AJ720663	15	9	11	12	16	14	30	15			
		AF001048											
		AF001049											
Retinal homeobox A	966	BC058757	5	9	10	7	22	23	35	22.5			
		BC073179											
		BC072132											
Rxrb (retinoid X receptor beta)	1335	BC001167	4	7	6	5.5	23	19	29	21	0.13		

Table 1.1 (continued 8)

Sister chromatid cohesion establishment factor (SCC2)	759	AY661732 AY661733 BC063859 BC071080	7	6	8	6.5	15	9	17	12			
Syndecan 2 (heparan sulfate proteoglycan 1)	570	BC041490 NM_001001462 BC043626	5	4	5	4.5	9	6	26	7.5			
Sek-1 receptor tyrosine kinase	2958	X91191 X65138 BC076643	7	8	2	7.5	35	50	109	42.5		+	
Selenoprotein I	1101*	BC071114 BC021229 BC043894	12	18	29	15	16	31	100	23.5*		+	
Selenoprotein T	414	BC044996 AY358096 BC077941	3	4	0	3.5	6	11	11	8.5		+	
Septin 11	981*	BC073250 XM_420341 AF212298	6	6	3	6	22	24	19	23			
Septin A (XISeptA)	1056	BC082859 NM_010891 BC073077	3	3	0	3	21	16	31	18.5		+	
serum/glucocorticoid regulated kinase	1302	BC074305 NM_204476 AF279245	1	11	6	6	29	24	38	26.5			+
Shab12	540*	U20342 AF450111 BC046706	3	3	1	3	10	14	13	12			
Siah-interacting protein	678	BC074445 NM_009786 X68683	13	8	8	10.5	10	9	25	9.5		+	
Sloan-Kettering viral oncogene homolog	2145	U89999 M28517 U75681	15	13	13	14	38	26	75	32		+	
Histone stem-loop binding protein (SLBP)	762	BC073461 BT007433 AF390895	14	8	23	11	18	17	25	17.5			
Suc1-associated neurotrophic factor target	1527	BC046943 AF036717 D83650	16	19	11	17.5	33	26	31	29.5			
Sox11 (XLS13)	1041	D87209 NM_205187 AJ001730	9	13	21	11	23	11	57	17			
Sox17a (HMG box transcription factor Sox17-alpha)	1122	AB052691 NM_022454 AB052692	12	25	13	18.5	25	27	26	26	0.56		
Sox18 (Transcription factor SOX-18)	897*	AB052693 NM_204309 AF394958	9	12	20	10.5	15	23	54	19			
SP22	489*	DQ406635 AB073864 X62483	2	3	11	2.5	15	11	22	13			
Sparc	897*	BC045013 AB116365 BC077749	14	6	5	10	18	18	34	18		+	
Spats2 (spermatogenesis associated, serine-rich 2)	1641	BC057748 NM_139140 BC078118	32	13	24	22.5	29	42	53	35.5			
Spermatid perinuclear RNA binding protein	1338*	BC089298 BC017732 AF331824	29	16	14	22.5	33	26	40	29.5			
Sprouty-2	921	AF331825 AF176904 BC056037	13	8	7	10.5	17	14	20	15.5			
Sulfide quinone reductase-like	1329	BC054152 BC028247 M24704	14	15	8	14.5	38	24	45	31			
Src (pp60c-src protein)	1596	M23422 V00402 BC076749	6	5	3	5.5	33	26	69	29.5		+	
Stanniocalcin 1	741	BC084424 XM_425760 BC060382	13	7	10	10	21	15	36	18			
Staufen 1	1281*	AY705672 NM_001012831 BC046709	15	11	12	13	32	27	42	29.5			
Stress-induced-phosphoprotein 1	1620	BC078016 U27830 BC042356	18	21	12	19.5	32	47	53	39.5		+	
Stomatin	420*	BC054307 NM_004099 AF387816	4	2	3	3	11	17	16	14			
Strabismus	1563	AY069979 NM_020335 X81986	6	0	1	3	43	33	69	38			+
SUG1	1200	BC045221 XM_425834 AF467942	2	1	1	1.5	39	29	47	34			
translation initiation factor SUI1	339	BC054139 XM_534229	0	0	0	0	5	4	7	4.5			

Table 1.1 (continued 9)

Sumo	303	Z97073 BC090210 XM_516035 AB197247 AF442492	3	0	0	1.5	3	4	4	3.5		
Survivin (Xsvv1)	480	AB182320 AF035016 AF035017	4	3	6	3.5	12	8	14	10		
Synaptobrevin	339	NM_009497 AF035014 AF035015	3	0	2	1.5	4	6	12	5		
Synaptophysin	882	BC064550 AF059570 AF136184	11	9	4	10	16	21	20	18.5		
Xwnt8 inhibitor sizzled (szl) (putative wnt inhibitor frzb3)	840	AF308868 AB022691 AB022692	9	4	11	6.5	19	10	24	14.5		
TAF-Ibeta	828	AJ851581 AF133036 AB032944	2	2	2	2	15	9	14	12		
T-box transcription factor Tbx5	360*	U64433 AF499688 BC098961	1	1	0	1	5	9	16	7		
TCRzeta subunit	477	NM_206879 BC047131 BC079707	13	7	16	10	18	11	21	14.5		
Bax Inhibitor-1, testis enhanced gene transcript	711	BC005588 BC073440 BC044079	10	8	3	9	10	14	27	12		+
TRk-fused protein TFG	1188	U94662 M35343 M35344	8	10	8	9	26	30	25	28	0.18	
Thyroid Hormone Receptor alpha*	1254	L06064 M35359 M35361	2	6	4	4	11	16	33	13.5		
Thyroid Hormone Receptor beta*	1107	L27344 Y14446 Y14447	3	5	1	4	18	16	18	17		
Mesoderm Posterior (Mesp)	924	Y17043 AJ416632 BC080105	20	18	14	19	26	23	17	24.5		
cytotoxic granule-associated RNA binding protein (TIA1)	1164	NM_022173 AJ416631 BC045086	8	3	2	5.5	32	25	36	28.5		
TIAR	1167	NM_003252 M64659 M64660	6	10	3	8	31	19	38	25		+
Tyrosine kinase	468*	M69243 M64661 AF055980	4	6	0	5	15	7	15	11	0.31	
IGF (Insulin-like Growth Factor) Receptor	462	M69244 AB056893 BC078542	1	3	2	2	6	6	17	6		
Transducer of erbB	894	CR456594 BC054950 BC056840	1	7	6	4	10	20	26	15		
Transferrin	2100	U05246 M34699 X64056	42	28	63	35	52	37	86	44.5		
Thyrotropin-releasing Hormone	672	BC069375 AF305620 AJ420782	18	18	29	18	9	11	31	10		
Thyrotropin-releasing Hormone Receptor 1	1188	NM_003301 BC044959 BC087462	7	3	9	5	17	18	68	17.5		
Neurotrophin receptor B xTrkB-alpha	1416	X77251 AB003078 AB003079	21	13	21	17	23	29	39	26		
fast skeletal Troponin C	483	AB110088 AF441126 X93491	1	0	1	0.5	11	4	7	2.5		
unitary non-NMDA glutamate receptor subunit U1	1437	X17314 BC077923 BC077801	14	11	14	12.5	35	30	59	32.5		
Ubiquiton-conjugating enzyme e2e	597	NM_009455 X59863 X57201	4	4	4	4	4	12	13	8*		
xUBF mRNA for upstream binding factor	1911	X53390 AJ506039 BC072878	22	22	30	22	58	48	78	53		
endoplasmic reticulum UDP-Glc/UDP-Gal transporter	1017	BC011888 AY112732 BC084966	2	5	17	3.5	14	12	65	13		
UDP-glucose ceramide glucosyltransferase	1182	BC038711 BC077361 BC077369	3	3	0	3	33	16	34	24.5*	0.25	+
Uroplakin 1A	627*	NM_007000 BC042931 BC077311	5	4	9	4.5	14	12	52	13		
Ubiquinol-cytochrome C reductase complex	1353	NM_025899	15	11	19	13	29	29	51	29		

Table 1.1 (continued 10)

Vasodilator-stimulated phosphoprotein	1101	BC077932 BC072836 NM 003370 AF064601	15	8	13	11.5	18	17	42	17.5			
Ventral anterior homeobox protein (Vax1)	792*	AJ271730 XM 544036 AJ238649	10	4	23	7	18	4	64	11*			
Ventral anterior homeobox protein (Vax2+3)	873	AF113517 Y17791 BC054252	4	6	6	5	10	9	28	9.5			
Von Hippel-Lindau binding protein 1	531	BC092334 XM 420327 AF064633 AF064634	6	8	2	7	9	13	18	11		+	
Vg1 RNA binding protein	1779	NM 001006359 X16843 X16844	6	9	2	7.5	31	25	40	28	0.325		
Vimentin*	1368	X56134 BC046713 BC068695	17	13	27	15	16	25	65	20.5			
Tryptophanyl-tRNA synthetase	1425	XM 421368 U13962 AF035443	15	13	26	14	37	22	58	29.5			
Wee1A kinase	1560*	AK131218 AF358869 AB071983	14	24	29	19	48	35	103	41.5			
Wee1B, Wee1-like protein kinase	957	D30743 AY255636 AY319928 overlap with BC078598	9	9	9	9	39	30	68	34.5			
Uterine sensitization-associated protein-1 (Wise-A)	639	AY319926 M55054 L07538	9	5	11	7	17	8	24	12.5		+	
Xwnt-3	399*	NM 204675 U42011 D82051	2	3	0	2.5	15	8	25	11.5			
Wilms' tumor suppressor (WT1)	1131*	AB033633 U26270 U26269	7	3	8	5	22	27	55	24.5			
Cofilin (XAC)	504	NM 001004406 BC086475 U25961	7	5	10	6	8	14	21	11			
XE2 (helix-loop-helix transcription factor E2)	363*	NM 003199 U63711 BC057721	3	1	2	2	7	4	14	5.5		+	+
Xefiltin	1440	XM 543999 AB018694 BC083024	12	31	24	21.5	26	42	82	34			
Epidermis specific serine protease Prss27 (Xepsin)	1044	NM 031948 U65751 U65750	39	37	47	38	31	34	40	32.5			
Fork head related (XFD1)	1197	NM 204770 X74315 X74316	13	16	11	14.5	14	20	27	17			+
Fork head protein (XFD2)	1074	XM 220287 D89783 D89785	29	12	37	20.5	18	30	53	24			
Interleukin-1 beta-converting enzyme (Caspase 1)	1068*	NM 001223 X73316 X73317	31	43	68	37	23	21	44	22			
Ximb (maternal B9.10 and B9.15 protein)	696	NM 017589 BC084925 L11363	5	14	14	9.5	19	16	28	17.5		+	
L-myc oncogene (xL-myc)	1032	NM 001033082 Z19577 L19566	20	14	6	17	22	18	34	20		+	
Xnot (homeobox protein)	546	NM 205354 AF410800 U79162	5	11	3	8	6	15	22	10.5*			
TGF-beta related growth factor Xnr 4 (Xnr4)	1044*	NM 018055 AB114039 AB114040	37	18	32	27.5	30	15	39	22.5			
XmF12	1668*	XM 228541 BC046663 X91243	18	16	24	17	43	35	48	39			
Xrpf (XrpF1 beta 1) GA binding protein transcription factor	1155	D13317 AF368041 AF368043	7	8	4	7.5	14	34	33	24*			
ZFTF (zinc finger transcription factor SLUG)	798	X77572 U44950 AB037700	0	6	3	3	13	16	24	14.5			+
ZPB (zona pellucida glycoprotein)	1413	AB025428	18	4	53	11	19	26	83	22.5			+

Table 1.2. Comparison of *ka/ks* ratios before versus after gene duplication using a branch test and across diploid and tetraploid lineages (Model A versus B in Figure 1.3). An asterisk indicates individually significant improvement of an additional parameter at α=0.05. Individually significant improvements require a higher *ka/ks* ratio in the diploid lineage in the branch test, which is against our expectation are assigned a P value of 1.0.

Table 1.2 (continued 1)

Is the average substitution ratio higher after duplication?

Gene	ln(L) 1-ratio	ln(L) 2-ratio	tetraploid ω	duploid ω	χ^2 (df = 1)	P value
Actin (skeletal, alpha 3)*	-1797.22	-1795.43	0.00	0.03	3.59	1.0000
Activin Receptor-Like Kinase-2 (ALK-2)	-2625.56	-2622.85	0.09	0.03	5.42	0.0199*
Activin receptor II	-2821.38	-2817.40	0.11	0.03	7.96	0.0048*
Adipophilin (fatvq)	-2339.44	-2339.18	0.19	0.14	0.52	0.4718
AE (Amidating Enzyme)	-4942.15	-4942.00	0.11	0.09	0.30	0.5854
Albumin (serum)*	-3884.53	-3884.29	0.34	0.29	0.47	0.4940
ALDH (Aldehyde dehydrogenase class1)	-2624.70	-2622.80	0.17	0.08	3.79	0.0514
Alpha Globin	-915.38	-908.67	0.11	0.80	13.40	1.0000
Amelogenin	-1019.09	-1018.93	0.49	0.36	0.33	0.5679
Xenopus Anterior Neural Folds, Homeobox gene	-1164.62	-1164.33	0.20	0.14	0.58	0.4456
Amyloid-Beta-like protein precursor	-3941.42	-3936.47	0.18	0.05	9.91	0.0016*
Apoptosis Inhibitor 5	-2756.02	-2753.78	0.06	0.02	4.48	0.0344*
AR (Androgen Receptor)	-770.25	-769.68	0.11	0.03	1.16	0.2823
Liver L-arginase	-1908.63	-1894.70	0.87	0.12	27.87	0.0000*
Arginase Type 2	-1873.94	-1871.78	0.07	0.01	4.32	0.0377*
Arrestin	-2125.88	-2124.55	0.07	0.03	2.66	0.1030
Aspartyl tRNA synthetase	-2881.09	-2880.25	0.05	0.03	1.67	0.1957
Atonal Homolog 5	-778.73	-778.26	0.25	0.13	0.91	0.3409
ATP synthase subunit B	-1463.03	-1462.33	0.15	0.09	1.40	0.2363
Bambi (TGF-beta signaling)	-1514.93	-1514.80	0.11	0.14	0.27	1.0000
Barren (brm1, 13S condensin XCAP-H subunit)	-3943.90	-3943.56	0.12	0.10	0.69	0.4065
Bestrophin-2 (VMD2L1)	-3121.97	-3117.20	0.32	0.13	9.55	0.0020*
Beta Globin	-1021.73	-1019.53	0.36	1.06	4.39	1.0000
Complement factor B (Bf B) (MHC class III gene)	-4899.06	-4897.76	0.47	0.32	2.61	0.1062
Biglycan	-2015.83	-2015.17	0.08	0.13	1.32	1.0000
Bicaudal-C	-3483.38	-3481.16	0.10	0.04	4.43	0.0354*
Bridging integrator 1 (Amphiphysin II)	-2658.59	-2656.34	0.18	0.08	4.51	0.0336*
Bmp4	-3092.97	-3092.56	0.25	0.18	0.84	0.3595
BMP (Bone Morphogenetic Protein) receptor	-2543.23	-2543.17	0.05	0.06	0.13	1.0000
Block of proliferation 1	-1859.71	-1858.84	0.15	0.09	1.74	0.1868
Brachyury (T)	-2373.99	-2373.61	0.10	0.07	0.76	0.3824
Serine/Threonine protein kinase (c-RNIL)	-4141.06	-4140.56	0.09	0.06	0.99	0.3203
Basic transcription element binding protein	-1487.52	-1486.33	0.08	0.21	2.38	1.0000
B-cell translocation gene 1, anti-proliferative	-931.29	-927.82	0.38	0.06	6.93	0.0085*
Calcium homeostasis endoplasmic reticulum protein	-5086.71	-5088.77	0.15	0.04	15.87	0.0001*
Calnexin	-3483.37	-3483.35	0.18	0.17	0.05	0.8293
Calponin H3 (clpH3)	-1616.36	-1616.11	0.08	0.06	0.50	0.4805
Calreticulin	-2295.44	-2294.79	0.06	0.09	1.30	1.0000
Carbonic anhydrase II	-1539.90	-1539.25	0.30	0.18	1.30	0.2534
Casein kinase I alpha S (Csnk1a1)	-1621.15	-1620.87	0.01	0.00	0.56	0.4538
CASK interacting protein 2	-7300.12	-7298.61	0.19	0.16	1.02	0.3135
Procathepsin B	-1949.79	-1949.73	0.10	0.12	0.12	1.0000
Beta Catenin interacting protein 1 (catnbip1)	-387.32	-387.03	0.05	0.00	0.60	0.4389
Cystathionine-beta-synthase	-1263.14	-1263.08	0.05	0.04	0.11	0.7405

Table 1.2 (continued 2)

voltage-dependent Calcium channel beta subunit	-2555.65	-2554.53	0.12	0.06	2.25	0.1336
CDC2 (cell division cycle 2, kinase)	-1658.94	-1654.68	0.08	0.01	8.52	0.0035*
Cathepsin E	-2252.77	-2252.76	0.13	0.12	0.02	0.9001
Carboxyl ester lipase	-2758.16	-2757.41	0.08	0.05	1.51	0.2196
Carboxyl ester lipase (bile salt-stimulated lipase)	-3370.83	-3370.63	0.13	0.11	0.41	0.5225
Centrin	-847.71	-847.38	0.04	0.07	0.67	1.0000
Cerebellin 2 precursor protein	-1229.07	-1226.07	0.30	0.07	6.01	0.0142*
Complement factor I (C3b/C4b inactivator)	-3638.83	-3638.74	0.32	0.28	0.19	0.6669
Cystic fibrosis transmembrane conductance regulator	-7571.35	-7568.44	0.17	0.09	5.82	0.0158*
Cortical granule lectin	-2023.10	-2022.98	0.17	0.21	0.25	1.0000
Choroideremia (Rab escort protein 1)	-3664.97	-3664.80	0.27	0.32	0.34	1.0000
Carbohydrate sulfotransferase 11 (Chst11)	-1759.69	-1759.44	0.08	0.06	0.51	0.4763
Cell death-inducing DFFA-like effector c (CIDE-3alpha)	-1422.90	-1422.89	0.16	0.15	0.02	0.8861
C-Jun proto-oncogene (AP-1, Activator Protein)	-1758.21	-1754.47	0.34	0.10	7.47	0.0063*
Dipeptidase 2 (metallopeptidase M20 family)	-2602.09	-2599.89	0.14	0.06	4.41	0.0357*
alpha-1 Collagen type II	-7211.18	-7209.91	0.20	0.13	2.54	0.1113
Connexin 31 (Gap junction beta-3 protein)	-1482.04	-1479.11	0.17	0.04	5.85	0.0155*
Contactin/F3/F11 (Contactin A)	-5600.92	-5597.88	0.18	0.10	6.07	0.0137*
Coronin	-2675.85	-2675.52	0.14	0.10	0.67	0.4145
Cortactin	-3053.50	-3053.50	0.10	0.10	0.00	0.9685
Cytoplasmic polyadenylation element binding protein	-3044.61	-3042.82	0.09	0.04	3.57	0.0588
CRY2 (cryptochrome 2)	-3092.56	-3091.79	0.10	0.07	1.54	0.2152
Crystallin, beta A1	-1105.00	-1104.02	0.33	0.14	1.95	0.1622
Cathepsin S (CTSS)	-1979.22	-1978.60	0.28	0.18	1.25	0.2641
Cullin3 (Cul3)	-3797.06	-3795.64	0.01	0.00	2.84	0.0918
CyclinE	-2301.62	-2301.56	0.17	0.15	0.12	0.7330
Brain Dopamine receptor D2	-1831.04	-1830.69	0.16	0.10	0.69	0.4049
Dapper 1, antagonist of beta-catenin (Frodo)	-4831.58	-4829.26	0.19	0.11	4.64	0.0312*
Death-associated protein kinase 1	-7489.69	-7484.45	0.11	0.04	10.47	0.0012*
Drebrin-like	-2190.23	-2188.38	0.30	0.14	3.71	0.0542
Debranching enzyme homolog 1	-3027.01	-3025.28	0.16	0.08	3.44	0.0635
Deleted in colorectal cancer tumor suppressor	-993.79	-989.34	0.66	0.14	8.90	0.0029*
Desmin	-2591.27	-2585.40	0.14	0.04	11.74	0.0006*
Hand2	-671.78	-667.77	0.11	0.00	8.01	0.0046*
Cytoplasmic dynein light-intermediate chain 1 (DLIC1)	-2646.83	-2646.06	0.11	0.06	1.54	0.2148
Dipeptidylpeptidase 3	-4323.38	-4322.81	0.20	0.15	1.15	0.2828
Dullard	-1279.69	-1279.28	0.03	0.01	0.83	0.3619
Dystroglycan (DAG1)	-4803.34	-4801.97	0.11	0.18	2.76	1.0000
Dystrophin	-2743.80	-2741.95	0.12	0.03	3.69	0.0547
Helix-loop-helix transcription factor XE1	-813.48	-812.66	0.16	0.07	1.65	0.1993
E2 (transcription factor E2)	-3504.58	-3503.57	0.16	0.09	2.02	0.1552
met-mesencephalon-olfactory transcription factor 1 (Ebf2)	-3053.48	-3042.30	0.15	0.01	22.36	0.0000*
CCAAT/enhancer binding protein (C/EBP), alpha	-1750.38	-1750.30	0.18	0.15	0.17	0.6841
Endothelin receptor type A	-2359.98	-2359.96	0.11	0.10	0.04	0.8434
EF (Elongation Factor-1 alpha, 42Sp48)	-2401.23	-2399.09	0.05	0.01	4.29	0.0384*
Aurora kinase A (EG2)	-2494.99	-2494.63	0.10	0.14	0.71	1.0000
Engrailed 2 (EN2)	-1451.99	-1450.80	0.11	0.23	2.38	1.0000
Enkephalin A (proenkephalin A)*	-1157.55	-1156.75	0.14	0.07	1.59	0.2078
ENO (2-phosphoglycerate dehydratase, enolase)	-2403.59	-2397.77	0.03	0.14	11.65	1.0000
Era (Estrogen Receptor alpha)	-2222.55	-2222.54	0.07	0.07	0.01	0.9051
Enhancer of split groucho	-3843.99	-3843.94	0.03	0.03	0.09	1.0000

Table 1.2 (continued 3)

Enhancer of zeste	-3855 18	-3848 86	0 08	0 01	12 65	0.0004*
Focal adhesion kinase	-5615 85	-5614 25	0 07	0 04	3.19	0 0743
Transcription factor (XLFB1)	-547 56	-547 11	0 27	0 09	0.89	0 3465
XFD-4	-2423 86	-2419 99	0 15	0 05	7.74	0.0054*
Flap endonuclease-1	-2195 71	-2195 68	0 07	0 08	0.07	1.0000
FetuinB	-2933 88	-2933 03	0 35	0 52	1.70	1.0000
Ftz-F1-related orphan receptor (xFF1r)	-2326 77	-2324 91	0 07	0 02	3 71	0.0541
FGF (embryonic fibroblast growth factor 4)	-1041 58	-1041 53	0 25	0 30	0.10	1.0000
Fibroblast growth factor receptor	-4452 12	-4447 33	0 14	0 05	9 58	0.0020*
Fibrinogen alpha	-4305 27	-4304 89	0 27	0 21	0.75	0 3855
Flotillin	-2318 61	-2316 61	0 19	0 08	4.01	0.0453*
fms-related tyrosine kinase 1	-4233 79	-4232 64	0 34	0 23	2.30	0.1291
Fms-interacting protein (NF2/meningioma region)	-3790 03	-3784 42	0 15	0 05	11.23	0.0008*
alpha-fodrin (Xen alpha 1) spectrin, non-erythrocytic 1	-1308 47	-1307 68	0 12	0 06	1.57	0.2097
c-fos proto-oncogene	-479 57	-479 40	0.33	0 20	0.33	0.5628
Succinate dehydrogenase complex, Flavoprotein	-3593 24	-3592 77	0 13	0 09	0.94	0.3323
Frequenin	-972 42	-971 32	0 02	0 00	2.20	0.1376
Fascin	-2870 27	-2869 53	0 15	0 06	1.47	0.2255
Furin*	-3218 64	-3218 54	0 10	0 08	0.20	0.6525
Fused toes homolog	-1529 24	-1528 84	0 12	0 07	0.79	0.3753
FYN (proto-oncogene c-fyn)	-2652 64	-2651 83	0 03	0 01	1.62	0.2036
Galectin	-1725 96	-1725 32	0 26	0 16	1.27	0.2591
alpha subunit of Gq Gtp-binding protein (G protein)	-1751 41	-1751 06	0 04	0 08	0 69	1.0000
GATA-binding protein transcription factor GATA-1	-2240 79	-2239 68	0 29	0 17	2.22	0.1361
Transcription factor xGata5	-2177 64	-2176 57	0 10	0 06	2.14	0.1436
Growth hormone A	-796 28	-795 18	0 14	0 05	2.19	0.1385
Guanylate kinase 1	-1179 75	-1178 88	0 22	0 12	1.75	0.1862
Glycogenin 1 (mitotic phosphoprotein 45)	-1823 42	-1823 21	0 15	0 11	0.43	0.5136
Holocytochrome c synthase (heme-lyase) (hccs-prov)	-1765 31	-1764 56	0 36	0 21	1.51	0.2198
cephalic Hedgehog, sonic hedgehog protein 4	-2285 38	-2285 07	0 23	0 17	0.63	0.4257
Transcription factor XHEN1	-738 55	-738 29	0 13	0 08	0.52	0.4704
Hypoxia-inducible factor 1 alpha	-2603 06	-2599 88	0 15	0 05	6.37	0.0116*
SafA - scaffold attachment factor A	-4544 87	-4543 23	0 15	0 09	3.29	0.0697
Homeobox 2/2.3*	-840 59	-840 21	0 13	0 07	0.75	0.3863
Insulin*	-564 95	-563 04	0 27	0 04	3 82	0.0507
Integrin beta-1 subunit*	-4338 84	-4338 43	0 04	0 05	0.82	1.0000
Inversin	-6329 37	-6327 62	0 28	0 19	3.49	0.0616
Ubiquitin carboxyl-terminal hydrolase 5 (Isopeptidase T)	-4573 10	-4570 87	0 12	0 06	4.47	0.0344*
KT-1 protein (Adqr34)	-3852 67	-3850 69	0 15	0 08	3.95	0.0468*
Kit receptor tyrosine kinase (c-kit)	-5520 53	-5518 07	0 28	0 17	4.92	0.0265*
Kinesin-like protein 2	-4903 83	-4903 75	0 15	0 13	0.16	0.6877
L1 (ribosomal protein L1)	-2083 26	-2083 25	0 04	0 04	0.00	0.9889
L14 (ribosomal protein L14)	-1012 80	-1012 77	0 13	0 15	0.06	1.0000
Lamin B	-3452 00	-3451 93	0 14	0 13	0.15	0.6980
Lamina associated polypeptide 2	-3023 24	-3023 09	0 28	0 33	0.30	1.0000
Clathrin, light polypeptide (Lcb)	-1132 45	-1130 32	0 14	0 04	4.25	0 0391*
Lactate dehydrogenase	-1858 06	-1858 24	0 13	0 22	1.62	1.0000
LEF-1 (lymphoid enhancer factor)	-1681 23	-1677 15	0 14	0 02	8.16	0.0043*
TGF-beta family member Lefty-A	-2039 35	-2039 34	0 11	0 11	0.00	0.9519
LIN domain binding protein	-1758 81	-1758 48	0 01	0 00	0.65	0.4299
Lipocalin (Ptgds)	-847 11	-846 82	0 20	0 13	0 57	0.4488

Table 1.2 (continued 4)

Lpa1R (lysophosphatidic acid receptor)	-1782.40	-1779.11	0.06	0.00	6.58	0.0103*
LR (Leptin Receptor)	-677.63	-676.40	0.13	0.03	2.46	0.1170
Lipoprotein (LDL) receptor-related protein 6	-1183.23	-1183.20	0.03	0.03	0.06	0.8041
Autoantigen La (La protein)	-2564.32	-2564.09	0.14	0.18	0.45	1.0000
Microfibrillar-associated protein 1	-2351.27	-2349.94	0.07	0.03	2.66	0.1026
Myristoylated alanine-rich C kinase substrate	-1150.96	-1150.30	0.42	0.24	1.33	0.2484
XMax2 and XMax4	-577.91	-577.70	0.04	0.10	0.42	1.0000
Myogenin	-1211.22	-1209.30	0.13	0.02	3.83	0.0503
Myozenin1	-1811.81	-1811.24	0.28	0.18	1.15	0.2841
N-CAM (neural cell adhesion molecule)*	-5904.07	-5903.93	0.26	0.22	0.28	0.5972
NF-M1 (middle molecular neurofilament)	-4752.00	-4750.42	0.21	0.13	3.17	0.0750
neurogenin-related 1 (X-NGNR-1)	-1299.88	-1299.88	0.18	0.18	0.00	0.9792
Internexin neuronal intermediate filament protein	-2624.80	-2621.90	0.14	0.06	5.78	0.0162*
NK3 transcription factor related, koza	-1333.40	-1332.89	0.37	0.24	1.03	0.3096
Nonmuscle myosin II heavy chain A	-4598.85	-4598.71	0.07	0.06	0.28	0.5979
Nonmuscle myosin heavy chain B	-1922.40	-1921.78	0.06	0.04	1.24	0.2655
Nucleolar-localized protein NO38	-1716.58	-1716.55	0.16	0.14	0.06	0.8126
Nucleobindin 1	-1021.72	-1021.51	0.05	0.03	0.43	0.5101
Nucleoplasmin	-1110.66	-1110.26	0.13	0.20	0.81	1.0000
Nucleoporin (Nup88)	-4276.52	-4276.49	0.15	0.14	0.07	0.7965
OLPA (Dorphin)	-1525.51	-1520.70	0.37	0.11	9.62	0.0019*
Olfactory marker protein (XOMP)	-966.11	-965.82	0.18	0.13	0.58	0.4458
OncogenesC-ets-1 (c-ets-1 proto-oncogene)*	-2276.75	-2276.67	0.03	0.04	0.14	1.0000
OncogenesC-ets-2 (c-ets-2 proto-oncogene)*	-2682.81	-2680.53	0.10	0.04	4.55	0.0328*
OncogenesC-myc (mvelocytomatosis)*	-2330.86	-2328.73	0.12	0.05	4.26	0.0390*
Dynactin 2 (p50)	-2205.81	-2205.69	0.07	0.09	0.24	1.0000
PACSIN2	-1982.96	-1979.29	0.12	0.04	7.32	0.0068*
Convertase PC2	-3264.99	-3259.68	0.07	0.01	10.63	0.0011*
Prolyl isomerase (Pin1)	-876.12	-876.09	0.07	0.08	0.07	1.0000
PKC (protein kinase C,delta)	-3826.52	-3824.62	0.13	0.07	3.81	0.0509
Plakoglobin	-4098.60	-4098.20	0.06	0.05	0.81	0.3695
Peripheral myelin protein 22	-906.20	-906.18	0.10	0.11	0.04	1.0000
POMC (pro-opiomelanocortin)*	-1541.28	-1538.79	0.14	0.05	4.99	0.0255*
POU domain Gene 1	-1785.93	-1782.32	0.13	0.02	7.23	0.0072*
POU3	-1406.18	-1406.05	0.12	0.09	0.25	0.6143
Phosphorylase phosphatase (Ppp2B)	-3009.31	-3006.81	0.04	0.01	5.00	0.0253*
Protein phosphatase 4, regulatory subunit 2 (Ppp4r2)	-2416.56	-2416.15	0.25	0.18	0.82	0.3642
LIM protein Pnckle	-4439.27	-4437.63	0.13	0.08	3.27	0.0704
Prolactin Receptor	-3771.19	-3769.81	0.29	0.18	2.74	0.0977
Prothymosin	-504.30	-502.95	0.19	0.03	2.71	0.1000
Phosphorylase, glycogen; brain	-4566.04	-4564.71	0.07	0.04	2.66	0.1027
RAB18 (member RAS oncogene family)	-1084.62	-1083.86	0.09	0.03	1.52	0.2179
Rac GTPase	-978.88	-977.19	0.05	0.00	3.38	0.0658
Rad51	-1836.96	-1836.72	0.03	0.02	0.47	0.4912
Raq-1	-2155.88	-2154.89	0.12	0.07	1.99	0.1586
Ral interacting protein (rlip) - RalA (RalB-binding protein)	-3125.04	-3124.88	0.07	0.06	0.31	0.5757
RalB	-1090.30	-1090.21	0.02	0.02	0.19	1.0000
Retinoic acid receptor alpha	-2260.21	-2255.19	0.11	0.03	10.03	0.0015*
RDS35 (retinal degradation slow/peripherin)	-1992.78	-1991.78	0.20	0.10	2.00	0.1577
RDS38/peripherin	-1815.74	-1811.32	0.16	0.04	8.83	0.0030*
Requiem	-2060.52	-2057.77	0.12	0.03	5.49	0.0191*

Table 1.2 (continued 5)

Rhodopsin	-1720.97	-1720.92	0.21	0.17	0.10	0.7516
Ringo (p33 ringo, ls26)	-1777.51	-1776.98	0.24	0.15	1.07	0.3010
RIO kinase 2	-3152.56	-3151.20	0.20	0.12	2.72	0.0992
Rwdd1 (RWD domain containing 1)	-1382.21	-1380.60	0.19	0.08	3.22	0.0727
Retinal homeobox A	-1794.51	-1794.47	0.09	0.08	0.08	0.7833
Rxrb (retinoid X receptor beta)	-2259.35	-2259.29	0.09	0.07	0.13	0.7219
Sister chromatid cohesion establishment factor (SCC2)	-1327.03	-1327.02	0.17	0.15	0.03	0.8609
Syndecan 2 (heparan sulfate proteoglycan 1)	-1015.41	-1013.58	0.22	0.06	3.66	0.0557
Sek-1 receptor tyrosine kinase	-4962.61	-4955.57	0.05	0.01	14.07	0.0002*
Selenoprotein I	-2296.88	-2291.42	0.20	0.07	10.91	0.0010*
Selenoprotein T	-713.18	-710.27	0.14	0.00	5.82	0.0159*
Septin 11	-1615.25	-1615.24	0.06	0.05	0.01	0.9294
Septin A (XlSeptA)	-1830.62	-1825.09	0.08	0.00	11.07	0.0009*
serum/glucocorticoid regulated kinase	-2265.78	-2265.59	0.07	0.05	0.39	0.5308
Shab12	-919.81	-918.95	0.07	0.02	1.71	0.1912
Sieh-interacting protein	-1232.00	-1229.35	0.31	0.09	5.28	0.0216*
Sloan-Kettering viral oncogene homolog	-3769.40	-3765.65	0.14	0.05	7.49	0.0062*
Histone stem-loop binding protein (SLBP)	-1527.82	-1527.18	0.17	0.27	1.29	1.0000
Suc1-associated neurotrophic factor target	-2740.83	-2739.61	0.18	0.09	2.45	0.1174
Sox11 (XLS13)	-1949.11	-1947.91	0.11	0.06	2.39	0.1221
Sox17a (HMG box transcription factor Sox17-alpha)	-2123.67	-2123.43	0.18	0.13	0.47	0.4911
Sox18 (Transcription factor SOX-18)	-1779.30	-1778.21	0.15	0.09	2.18	0.1399
SP22	-944.74	-943.39	0.06	0.17	2.69	1.0000
Sparc	-1627.99	-1624.71	0.17	0.04	6.56	0.0104*
Spats2 (spermatogenesis associated, serine-rich 2)	-3127.53	-3127.04	0.22	0.16	0.98	0.3218
Spermatid perinuclear RNA binding protein	-2583.21	-2580.85	0.26	0.12	4.72	0.0297*
Sprouty-2	-1659.33	-1658.69	0.24	0.13	1.29	0.2562
Sulfide quinone reductase-like	-2476.37	-2474.60	0.14	0.06	3.54	0.0601
Src (pp60c-src protein)	-2798.21	-2795.17	0.06	0.01	6.08	0.0137*
Stanniocalcin 1	-1439.65	-1438.54	0.14	0.07	2.21	0.1372
Staufen 1	-2396.15	-2395.21	0.18	0.10	1.88	0.1701
Stress-induced-phosphoprotein 1	-3014.90	-3011.96	0.15	0.06	5.84	0.0157*
Stomatin	-781.82	-781.78	0.06	0.05	0.07	0.7893
Strabismus	-2694.05	-2692.53	0.02	0.00	3.04	0.0810
SUG1	-2086.85	-2086.63	0.01	0.01	0.44	0.5085
translation initiation factor SUI1	-517.16	-517.16	0.00	0.00	0.00	1.0000
Sumo	-475.03	-473.88	0.10	0.00	2.29	0.1298
Survivin (Xsvv1)	-861.17	-861.15	0.12	0.14	0.04	1.0000
Synaptobrevin	-577.88	-577.73	0.11	0.06	0.29	0.5880
Synaptophysin	-1614.25	-1612.90	0.19	0.08	2.70	0.1002
Xwnt8 inhibitor sizzled (szl) (putative wnt inhibitor frzb3)	-1521.99	-1521.97	0.14	0.15	0.04	1.0000
TAF-Ibeta	-1286.13	-1286.03	0.05	0.03	0.20	0.6576
T-box transcription factor Tbx5	-618.52	-617.00	0.04	0.00	3.04	0.0813
TCRzeta subunit	-1043.09	-1043.06	0.23	0.26	0.05	1.0000
Bax Inhibitor-1, testis enhanced gene transcript	-1287.90	-1283.15	0.24	0.03	9.50	0.0021*
TRK-fused protein TFG	-2105.83	-2105.83	0.11	0.11	0.00	0.9690
Thyroid Hormone Receptor alpha*	-2045.57	-2044.95	0.09	0.04	1.25	0.2636
Thyroid Hormone Receptor beta*	-1810.07	-1809.12	0.08	0.02	1.91	0.1665
Mesoderm Posterior (Mesp)	-1813.70	-1813.68	0.25	0.22	0.04	0.8477
cytotoxic granule-associated RNA binding protein (TIA1)	-2048.49	-2046.97	0.06	0.02	3.04	0.0812
TIAR	-2073.98	-2071.67	0.10	0.03	4.62	0.0316*

Table 1.2 (continued 6)

Tyrosine kinase	-836.17	-831.75	0.14	0.00	8.86	0.0029*
IGF (Insulin-like Growth Factor) Receptor	-782.95	-782.24	0.10	0.03	1.43	0.2312
Transducer of erbB	-1541.06	-1541.02	0.09	0.08	0.08	0.7825
Transferrin	-4313.97	-4313.86	0.26	0.24	0.22	0.6392
Thyrotropin-releasing Hormone	-1467.08	-1465.79	0.56	0.29	2.58	0.1079
Thyrotropin-releasing Hormone Receptor 1	-2128.36	-2127.04	0.10	0.04	2.65	0.1034
Neurotrophin receptor B xTrkB-alpha	-2636.31	-2635.88	0.25	0.18	0.85	0.3568
fast skeletal Troponin C	-729.63	-729.51	0.02	0.04	0.23	1.0000
unitary non-NMDA glutamate receptor subunit U1	-2679.88	-2679.48	0.11	0.08	0.80	0.3716
Ubiquiton-conjugating enzyme e2e	-1004.42	-1004.19	0.17	0.10	0.47	0.4926
xUBF mRNA for upstream binding factor	-3674.88	-3674.85	0.09	0.09	0.06	0.8083
endoplasmic reticulum UDP-Glc/UDP-Gal transporter	-1874.90	-1874.88	0.09	0.08	0.03	0.8629
UDP-glucose ceramide glucosyltransferase	-1972.71	-1969.63	0.05	0.00	6.17	0.0130*
Uroplakin 1A	-1224.91	-1224.53	0.09	0.05	0.76	0.3830
Ubiquinol-cytochrome C reductase complex	-2538.45	-2538.06	0.16	0.12	0.80	0.3720
Vasodilator-stimulated phosphoprotein	-2024.82	-2023.14	0.27	0.13	3.36	0.0669
Ventral anterior homeobox protein (Vax1)	-1566.59	-1564.95	0.19	0.09	3.27	0.0704
Ventral anterior homeobox protein (Vax2+3)	-1510.65	-1509.66	0.23	0.10	1.99	0.1586
Von Hippel-Lindau binding protein 1	-970.30	-966.99	0.18	0.03	6.61	0.0102*
Vq1 RNA binding protein	-2948.84	-2946.52	0.08	0.02	4.64	0.0313*
Vimentin*	-2604.31	-2602.98	0.21	0.12	2.67	0.1024
Tryptophanyl-tRNA synthetase	-2707.49	-2707.46	0.14	0.13	0.05	0.8242
Wee1A kinase	-3164.21	-3162.45	0.13	0.08	3.52	0.0607
Wee1B, Wee1-like protein kinase	-1929.97	-1928.30	0.07	0.03	3.35	0.0672
Uterine sensitization-associated protein-1 (Wise-A)	-1199.95	-1199.93	0.15	0.14	0.04	0.8398
Xwnt-3	-735.61	-731.90	0.06	0.00	7.42	0.0064*
Wilms' tumor suppressor (WT1)	-2042.62	-2042.12	0.06	0.04	1.01	0.3140
Cofilin (XAC)	-963.33	-963.32	0.14	0.13	0.02	0.8986
XE2 (helix-loop-helix transcription factor E2)	-622.22	-621.54	0.14	0.05	1.37	0.2410
Xefiltin	-2914.25	-2909.58	0.18	0.07	9.34	0.0022*
Epidermis specific serine protease Prss27 (Xepsin)	-2417.67	-2417.48	0.39	0.47	0.38	1.0000
Fork head related (XFD1)	-2136.14	-2134.92	0.25	0.12	2.45	0.1176
Fork head protein (XFD2)	-2243.35	-2243.27	0.20	0.18	0.16	0.6938
Interleukin-1 beta-converting enzyme (Caspase 1)	-2492.91	-2492.91	0.53	0.52	0.00	0.9903
Ximb (maternal B9.10 and B9.15 protein)	-1367.53	-1367.30	0.15	0.11	0.46	0.4979
L-myc oncogene (xL-myc)	-1946.26	-1941.20	0.27	0.06	10.13	0.0015*
Xnot (homeobox protein)	-1026.79	-1022.13	0.22	0.03	9.32	0.0023*
TGF-beta related growth factor Xnr-4 (Xnr4)	-2205.24	-2204.65	0.39	0.27	1.18	0.2773
XmF12	-3147.93	-3147.73	0.15	0.19	0.40	1.0000
Xrpf (XrpFI beta 1) GA binding protein transcription factor	-1995.12	-1993.77	0.09	0.03	2.69	0.1008
ZFTF (zinc finger transcription factor SLUG)	-1368.70	-1368.53	0.05	0.03	0.33	0.5678
ZPB (zona pellucida glycoprotein)	-2832.59	-2832.50	0.19	0.21	0.19	1.0000

Table 1.3. Results of test for different nonsynonymous substitution rates in each paralog. A three-rate model was compared to a 2-rate model (Models C and B in Figure 1.3). Individual significance of the rate test is indicated with an asterisk at a = 0.05 and a P value of 1.0 is assigned if the ka/ks ratio (estimated by a separate test) of the paralog with the faster rate was lower than the other paralog and/or lower than the diploid lineage. Maximum likelihood estimates of *ka/ks* ratios are listed; these ratios are not necessarily equivalent to the ratio of the number of nonsynonymous and synonymous sites in Table 1.1 because those listed here consider multiple substitutions at each site. See text for further details and tablewide significance.

Table 1.3 (continued 1)

Gene	Rate test: ln(L) 2-rate	ln(L) 3-rate	χ2(df= 1)	P value	ka/ks ratio α	ka/ks ratio β	ka/ks ratio diploid
Actin (skeletal, alpha 3)*	-1107.14	-1107.14	0.00	1.0000	0.0001	0.000	0.029
Activin Receptor-Like Kinase-2 (ALK-2)	-1558.96	-1558.93	0.05	1.0000	0.0835	0.102	0.028
Activin receptor II	-1687.15	-1686.74	0.83	0.3635	0.0701	0.155	0.028
Adipophilin - Adipose differentiation-	-1464.08	-1463.22	1.72	0.1903	0.2202	0.155	0.145
AE (Amidating Enzyme)	-3041.93	-3041.69	0.48	0.4888	0.0958	0.125	0.094
Albumin (serum)*	-2576.29	-2575.48	1.62	1.0000	0.3126	0.397	0.287
ALDH (Aldehyde dehydrogenase class1)	-1695.87	-1695.86	0.04	0.8463	0.2126	0.136	0.076
Alpha Globin	-567.48	-567.37	0.24	1.0000	0.0761	0.155	0.799
Ameloqenin	-673.41	-671.70	3.41	0.0650	0.3746	0.590	0.367
Xenopus Anterior Neural Folds, Homeobox gene	-722.09	-722.08	0.01	0.9149	0.2289	0.176	0.139
Amyloid-Beta-like protein precursor	-2515.71	-2511.93	7.56	0.0060*	0.269	0.103	0.055
Apoptosis Inhibitor 5	-1545.94	-1545.52	0.84	0.3582	0.0453	0.074	0.022
AR (Androgen Receptor)	-485.22	-483.85	2.75	1.0000	0.0001	0.109	0.034
Liver L-arginase	-1220.90	-1203.75	34.30	0.0000*	0.0001	0.870	0.115
Arginase Type 2	-1111.95	-1111.13	1.65	0.1991	0.0435	0.099	0.011
Arrestin	-1228.15	-1227.86	0.58	0.4455	0.092	0.058	0.033
Aspartyl tRNA synthetase	-1653.61	-1653.35	0.53	0.4685	0.0629	0.031	0.026
Atonal Homolog 5	-508.64	-507.59	2.09	0.1482	0.4483	0.127	0.133
ATP synthase subunit B	-867.41	-867.35	0.13	0.7152	0.2928	0.101	0.085
Bambi (TGF-beta signalling)	-937.47	-935.41	4.12	1.0000	0.0797	0.131	0.144
Barren (brm1, 13S condensin XCAP-H	-2417.75	-2417.75	0.01	1.0000	0.125	0.121	0.096
Bestrophin-2 (VMD2L1)	-2003.05	-2003.03	0.05	1.0000	0.3477	0.292	0.127
Beta Globin	-676.22	-675.85	0.75	1.0000	0.2977	0.431	1.068
Complement factor B (Bf B)	-3289.18	-3289.16	0.05	0.8282	0.4727	0.461	0.320
Biglycan	-1216.00	-1215.56	0.88	1.0000	0.0487	0.111	0.126
Bicaudal-C	-2057.11	-2056.48	1.26	0.2612	0.0988	0.092	0.041
Bridging integrator 1 (Amphiphysin II)	-1612.92	-1611.54	2.77	0.0963	0.2342	0.124	0.084
Bing4	-2028.20	-2025.84	4.72	0.0298*	0.1481	0.365	0.184
BMP (Bone Morphogenetic Protein)	-1548.64	-1548.35	0.59	1.0000	0.0458	0.058	0.061
Block of proliferation 1	-1170.45	-1169.89	1.12	0.2900	0.0739	0.289	0.089
Brachyury (T)	-1431.37	-1430.18	2.39	0.1224	0.0601	0.143	0.067
Serine/Threonine protein kinase	-2461.01	-2460.59	0.85	1.0000	0.093	0.085	0.064
Basic transcription element binding protein	-933.63	-933.08	1.09	1.0000	0.0848	0.079	0.210
B-cell translocation gene 1, anti-proliferative	-595.49	-595.26	0.46	0.4957	0.8109	0.203	0.059
Calcium homeostasis endoplasmic reticulum	-3079.75	-3075.21	9.07	0.0026*	0.206	0.093	0.042
Calnexin	-2220.62	-2220.25	0.75	0.3880	0.2773	0.120	0.171
Calponin H3 (clpH3)	-963.90	-963.88	0.03	1.0000	0.1365	0.063	0.056
Calreticulin	-1390.33	-1390.32	0.02	1.0000	0.0478	0.074	0.093
Carbonic anhydrase II	-990.91	-989.89	2.04	0.1533	0.495	0.161	0.185
Casein kinase I alpha S							
Casein kinase 1, alpha 1	-985.72	-985.02	1.40	0.2361	0.0001	0.014	0.000
CASK interacting protein 2	-4472.90	-4472.40	0.99	0.3194	0.2341	0.160	0.158
Procathepsin B	-1185.41	-1184.99	0.84	1.0000	0.0851	0.117	0.118
Beta Catenin interacting protein 1	-230.19	-229.50	1.39	0.2380	0.0001	0.192	0.000
Cystathionine-beta-synthase	-707.63	-707.56	0.13	0.7134	0.0581	0.039	0.039
voltage-dependent Calcium channel beta	-1570.68	-1569.47	2.43	0.1194	0.0755	0.163	0.062
CDC2 (cell division cycle 2, kinase)	-950.92	-949.35	3.14	0.0766	0.04	0.115	0.007

Table 1.3 (continued 2)

Cathepsin E	-1360.10	-1359.73	0.76	0.3841	0.1044	0.147	0.121
Carboxyl ester lipase	-1601.04	-1600.53	1.02	0.3128	0.1	0.066	0.053
Carboxyl ester lipase (bile salt-stimulated	-2007.85	-2007.67	0.38	1.0000	0.1469	0.112	0.105
Centrin	-514.75	-514.58	0.35	1.0000	0.0407	0.030	0.074
Cerebellin 2 precursor protein	-774.15	-773.86	0.59	0.4411	0.2746	0.326	0.071
Complement factor I (C3b/C4b inactivator)	-2444.96	-2443.89	2.15	0.1425	0.3955	0.243	0.278
Cystic fibrosis transmembrane	-4819.42	-4819.39	0.06	0.8141	0.1428	0.196	0.090
Cortical granule lectin	-1255.47	-1254.36	2.22	1.0000	0.1673	0.179	0.208
Choroideremia (Rab escort protein 1)	-2445.10	-2442.40	5.40	0.0201*	0.1792	0.367	0.326
Carbohydrate sulfotransferase 11	-1072.50	-1071.89	1.21	0.2705	0.0977	0.069	0.056
Cell death-inducing DFFA-like effector c	-872.78	-872.78	0.01	0.9278	0.2069	0.128	0.149
C-Jun (c-jun proto oncogene)	-1110.08	-1110.04	0.08	1.0000	1.1428	0.210	0.101
Dipeptidase 2 (metallopeptidase M20	-1608.81	-1608.34	0.95	0.3294	0.1719	0.110	0.062
alpha-1 Collagen type II	-4303.92	-4302.15	3.53	0.0604	0.1642	0.227	0.131
Connexin 31 (Gap junction beta-3 protein)	-921.02	-920.68	0.67	0.4129	0.1879	0.140	0.042
Contactin/F3/F11 (Contactin A)	-3600.35	-3600.33	0.04	0.8422	0.1801	0.183	0.096
Coronin	-1686.76	-1686.76	0.00	0.9977	0.1536	0.129	0.103
Cortactin	-1867.07	-1866.91	0.32	0.5717	0.1061	0.097	0.103
Cytoplasmic polyadenylation element	-1815.61	-1814.67	1.87	0.1712	0.1266	0.057	0.036
CRY2 (cryptochrome 2)	-1847.24	-1847.20	0.10	0.7555	0.1192	0.088	0.069
Crystallin, beta A1	-765.68	-764.92	1.54	0.2153	0.3642	0.276	0.136
Cathepsin S (CTSS)	-1299.16	-1299.16	0.01	1.0000	0.3478	0.237	0.183
Cullin3 (Cul3)	-2234.99	-2234.82	0.34	0.5618	0.0055	0.010	0.000
CyclinE	-1503.61	-1503.18	0.86	0.3527	0.2048	0.136	0.149
Brain Dopamine receptor D2	-1175.53	-1175.15	0.77	0.3816	0.1149	0.210	0.099
Dapper 1, antagonist of beta-catenin	-3004.40	-3002.70	3.39	0.0655	0.1879	0.194	0.105
Death-associated protein kinase 1	-4610.16	-4608.11	4.11	0.0427*	0.0825	0.125	0.040
Drebrin-like	-1407.14	-1407.12	0.04	0.8408	0.2466	0.388	0.142
Debranching enzyme homolog 1	-1899.47	-1897.85	3.24	0.0719	0.1815	0.133	0.084
Deleted in colorectal cancer tumor	-621.87	-619.61	4.51	0.0338*	0.2898	1.029	0.145
Desmin	-1519.07	-1518.86	0.43	0.5127	0.1035	0.185	0.035
Hand2	-408.75	-407.34	2.81	0.0934	0.0935	0.126	0.000
Cytoplasmic dynein light intermediate chain 1	-1662.79	-1662.78	0.02	0.8756	0.1075	0.117	0.063
Dipeptidylpeptidase 3	-2697.45	-2697.43	0.04	1.0000	0.1638	0.271	0.149
Dullard	-713.62	-712.24	2.77	0.0959	0.0001	0.112	0.011
Dystroglycan (DAG1)	-3002.76	-3002.58	0.35	1.0000	0.1178	0.106	0.182
Dystrophin	-1764.46	-1761.69	5.54	0.0186*	0.1282	0.000	0.028
Helix-loop-helix transcription factor XE1	-505.15	-504.74	0.83	0.3612	0.1279	0.193	0.066
E2 (transcription factor E2)	-2137.36	-2133.77	7.18	0.0074*	0.1941	0.102	0.094
mat-mesencephalon-olfactory transcription	-1839.06	-1838.73	0.66	0.4156	0.1421	0.1624	0.011
CCAAT/enhancer binding protein (C/EBP), alpha	-1114.55	-1113.12	2.86	0.0910	0.1394	0.205	0.156
Endothelin receptor type A	-1430.42	-1429.55	1.75	0.1861	0.208	0.060	0.098
EF (Elongation Factor-1 alpha, 42Sp48)	-1376.60	-1375.17	2.86	0.0906	0.018	0.071	0.011
Aurora kinase A (EG2)	-1498.96	-1498.70	0.52	1.0000	0.0793	0.127	0.138
Engrailed 2 (EN2)	-916.80	-915.64	2.31	1.0000	0.0605	0.151	0.228
Enkephalin A (proenkephalin A)*	-743.61	-743.20	0.82	0.3664	0.2353	0.090	0.067
ENO (alpha enolase) (2-phosphoglycerate	-1424.27	-1422.29	3.97	1.0000	0.0001	0.065	0.140
Era (Estrogen Receptor alpha)	-1374.29	-1372.84	2.91	0.0881	0.0558	0.077	0.066
Enhancer of split groucho	-2263.09	-2260.69	4.80	0.0285*	0.0001	0.056	0.033
Enhancer of zeste	-2401.48	-2399.14	4.68	0.0306*	0.065	0.093	0.010
Focal adhesion kinase	-3396.32	-3394.95	2.75	0.0971	0.052	0.087	0.042

Table 1.3 (continued 3)

Transcription factor (clone XLFB1)	-351.23	-351.18	0.11	1.0000	0.377	0.222	0.086
XFD-4	-1561.18	-1561.03	0.29	1.0000	0.2515	0.116	0.046
Flap endonuclease-1	-1280.68	-1280.52	0.32	0.5725	0.0928	0.054	0.076
FetuinB	-2022.10	-2021.38	1.42	1.0000	0.354	0.341	0.515
Ftz-F1-related orphan receptor (xFF1r)	-1433.94	-1432.01	3.87	0.0491*	0.1088	0.027	0.017
FGF (embryonic fibroblast growth factor	-687.79	-687.62	0.33	1.0000	0.1861	0.478	0.295
Fibroblast growth factor receptor	-2730.49	-2728.46	4.04	0.0443*	0.0976	0.173	0.048
Fibrinogen alpha	-2850.21	-2847.19	6.03	0.0141*	0.1593	0.409	0.215
Flotillin	-1417.32	-1417.29	0.07	1.0000	0.208	0.172	0.085
fms-related tyrosine kinase 1/vascular	-2823.81	-2823.65	0.31	0.5781	0.4305	0.278	0.231
Fms-interacting protein (NF2/meningioma	-2321.61	-2321.04	1.15	0.2837	0.1932	0.119	0.049
alpha-fodrin (Xen alpha 1)	-791.13	-791.09	0.08	0.7796	0.0855	0.201	0.063
c-fos proto-oncogene	-296.58	-295.92	1.32	0.2506	0.2441	0.376	0.202
Succinate dehydrogenase	-2235.14	-2234.75	0.78	0.3767	0.1683	0.093	0.090
Frequenin	-569.42	-569.41	0.04	0.8483	0.0175	0.019	0.000
Fascin	-1647.17	-1646.23	1.87	1.0000	0.091	0.091	0.061
Furin*	-1984.40	-1982.85	3.10	0.0783	0.0666	0.130	0.082
Fused toes homolog	-946.42	-945.74	1.36	0.2441	0.0825	0.143	0.065
FYN (c-fyn, Fyn proto-oncogene)	-1609.57	-1609.57	0.00	1.0000	0.027	0.035	0.011
Galectin	-1064.71	-1064.54	0.34	0.5602	0.2548	0.261	0.163
alpha subunit of Gq Gtp-binding protein	-1100.02	-1099.50	1.05	1.0000	0.0643	0.023	0.080
GATA-binding protein transcription factor	-1375.95	-1375.85	0.18	1.0000	0.3929	0.229	0.169
Transcription factor xGata5	-1297.08	-1296.31	1.54	0.2140	0.124	0.077	0.056
Growth hormone A	-485.58	-483.57	4.01	0.0452*	0.0501	0.268	0.054
Guanylate kinase 1	-732.88	-732.16	1.44	0.2308	0.4128	0.121	0.119
Glycogenin 1 (mitotic phosphoprotein	-1115.05	-1114.90	0.31	0.5788	0.1308	0.160	0.109
Holocytochrome c synthase (cytochrome c	-1200.74	-1197.78	5.93	0.0149*	0.1613	0.705	0.215
cephalic Hedgehog, sonic hedgehog protein	-1444.97	-1444.67	0.61	0.4353	0.2256	0.237	0.170
Transcription factor XHEN1	-427.70	-427.58	0.25	0.6185	0.1222	0.145	0.079
Hypoxia-inducible factor 1 alpha	-1628.67	-1627.60	2.14	0.1438	0.1676	0.123	0.047
SafA - scaffold attachment factor A	-2804.69	-2803.44	2.49	0.1145	0.1648	0.122	0.090
Homeobox 2/2.3*	-547.22	-547.01	0.41	1.0000	0.102	0.209	0.074
Insulin*	-355.49	-354.98	1.03	0.3107	0.2341	0.303	0.037
Integrin beta-1 subunit*	-2559.93	-2559.93	0.01	1.0000	0.0406	0.033	0.051
Inversin	-4147.08	-4146.88	0.40	0.5277	0.3213	0.240	0.195
Ubiquitin carboxyl-terminal hydrolase 5	-2813.14	-2813.09	0.09	0.7679	0.1411	0.097	0.056
Kf-1 protein (Adgr34)	-2413.41	-2413.40	0.03	1.0000	0.1799	0.132	0.083
Kit receptor tyrosine kinase (c-kit)	-3577.21	-3577.20	0.02	1.0000	0.2273	0.354	0.165
Kinesin-like protein 2	-3095.42	-3095.37	0.10	0.7520	0.1627	0.132	0.130
L1 (ribosomal protein L1)	-1188.41	-1187.34	2.15	0.1425	0.0901	0.012	0.041
L14 (ribosomal protein L14)	-622.20	-622.20	0.00	1.0000	0.1338	0.128	0.154
Lamin B	-2056.23	-2056.12	0.21	1.0000	0.151	0.138	0.128
Lamina associated polypeptide 2	-1995.50	-1994.88	1.22	1.0000	0.2711	0.292	0.331
Clathrin, light polypeptide (Lcb)	-679.10	-679.06	0.08	0.7841	0.1609	0.115	0.035
Lactate dehydrogenase	-1194.51	-1194.30	0.41	1.0000	0.1167	0.166	0.219
LEF-1 (lymphoid enhancer factor)	-1179.90	-1179.87	0.07	0.7913	0.1265	0.156	0.020
TGF-beta family member Lefty-A	-1267.13	-1266.99	0.28	0.5978	0.1193	0.102	0.114
LIM domain binding protein	-1103.34	-1102.66	1.37	0.2419	0.0001	0.018	0.000
Lipocalin (Ptgds)	-562.33	-561.87	0.92	0.3366	0.2753	0.129	0.132
Lpa1R (lysophosphatidic acid receptor)	-1094.97	-1094.63	0.69	0.4077	0.0803	0.043	0.000
LR (Leptin Receptor)	-387.45	-387.35	0.20	0.6536	0.1344	0.128	0.026

Table 1.3 (continued 4)

Lipoprotein (LDL) receptor-related protein	-663.39	-663.39	0.00	0.9768	0.0294	0.039	0.028
Autoantigen La (La protein)	-1612.49	-1612.46	0.08	1.0000	0.1291	0.162	0.179
Microfibrillar-associated protein 1	-1361.84	-1361.83	0.01	0.9163	0.0726	0.064	0.031
Myristoylated alanine-rich C kinase substrate	-715.64	-714.24	2.80	0.0945	0.681	0.241	0.247
XMax2 and XMax4	-355.61	-354.92	1.38	1.0000	0.0594	0.000	0.106
Myogenin	-743.15	-742.95	0.40	0.5295	0.1179	0.147	0.021
Myozenin1	-1153.56	-1152.70	1.71	0.1904	0.3979	0.197	0.175
N-CAM (neural cell adhesion molecule)*	-3840.77	-3837.77	6.00	0.0143*	0.3101	0.201	0.226
NF-M1 (middle molecular neurogenin-related 1 (X-NGNR-1)	-2902.80	-2902.75	0.09	0.7667	0.201	0.217	0.130
Internexin neuronal intermediate filament	-816.39	-816.19	0.41	1.0000	0.1907	0.173	0.178
NK3 transcription factor related, locus 1	-1577.68	-1576.87	1.62	1.0000	0.1362	0.144	0.059
Nonmuscle myosin II heavy chain A	-906.58	-906.55	0.06	0.8105	0.4125	0.337	0.240
Nonmuscle myosin heavy chain B	-2838.50	-2838.30	0.40	0.5285	0.0866	0.059	0.059
Nucleolar-localized protein NO38	-1057.96	-1054.81	6.29	0.0121*	0.1306	0.012	0.036
Nucleobindin 1	-1095.93	-1093.39	5.08	0.0242*	0.0822	0.256	0.145
Nucleoplasmin	-617.99	-617.65	0.68	1.0000	0.0545	0.051	0.025
Nucleoporin (Nup88)	-668.03	-667.86	0.34	1.0000	0.1155	0.140	0.205
OLPA (Dorphin)	-2630.12	-2628.34	3.56	0.0591	0.0973	0.203	0.140
Olfactory marker protein (XOMP)	-956.32	-956.12	0.40	0.5265	0.2879	0.473	0.111
OncogenesC-ets-1 (c-ets-1b proto-oncogene)*	-630.77	-630.69	0.16	1.0000	0.1278	0.378	0.122
OncogenesC-ets-2 (ets-2a proto-oncogene)*	-1363.14	-1363.09	0.11	1.0000	0.0418	0.027	0.042
OncogenesC-myc (myelocytomatosis	-1607.82	-1607.81	0.03	0.8681	0.126	0.080	0.042
Dynactin 2 (p50)	-1407.29	-1407.24	0.09	0.7596	0.094	0.171	0.053
PACSIN2	-1303.85	-1303.76	0.19	1.0000	0.0654	0.080	0.090
Convertase PC2	-1232.49	-1230.91	3.14	0.0762	0.148	0.090	0.042
Prolyl isomerase (Pin1)	-1942.75	-1942.33	0.84	1.0000	0.0778	0.072	0.012
PKC (protein kinase C, delta)	-525.01	-525.00	0.02	0.8849	0.08	0.057	0.080
Plakoglobin	-2348.88	-2348.75	0.24	0.6219	0.1042	0.163	0.069
Peripheral myelin protein 22	-2377.19	-2377.19	0.00	1.0000	0.0709	0.058	0.047
POMC (pro-opiomelanocortin)*	-524.87	-524.19	1.36	0.2444	0.143	0.055	0.113
POU domain Gene 1	-936.16	-936.16	0.01	1.0000	0.2699	0.105	0.048
POU3	-1122.05	-1118.60	6.90	0.0086*	0.1843	0.045	0.019
Phosphorylase phosphatase (Ppp2B)	-905.52	-902.93	5.18	0.0228*	0.222	0.027	0.091
Protein phosphatase 4, regulatory subunit 2	-1730.29	-1727.74	5.08	0.0241*	0.1012	0.007	0.005
LIM protein Prickle	-1547.25	-1547.01	0.48	1.0000	0.3325	0.203	0.177
Prolactin Receptor	-2830.32	-2829.73	1.19	0.2744	0.1851	0.095	0.076
Prothymosin	-2450.50	-2450.05	0.90	1.0000	0.3075	0.273	0.183
Phosphorylase, glycogen; brain	-317.57	-317.53	0.08	0.7720	0.2172	0.165	0.032
RAB18 (member RAS oncogene family)	-2761.33	-2761.27	0.13	0.7173	0.1083	0.054	0.041
Rac GTPase	-649.70	-649.19	1.02	0.3122	0.0486	0.130	0.033
Rad51	-570.80	-570.62	0.36	0.5506	0.0361	0.096	0.000
Raf-1	-1008.02	-1007.68	0.69	0.4077	0.0425	0.019	0.018
Ral interacting protein (Rip gene) RalA	-1298.27	-1298.15	0.25	0.6189	0.0902	0.196	0.066
RalB	-1925.97	-1922.94	6.06	0.0138*	0.0363	0.105	0.059
Retinoic acid receptor alpha	-596.95	-596.62	0.67	1.0000	0.0227	0.000	0.024
RDS35 (retinal degradation	-1332.57	-1328.42	8.30	0.0040*	0.0328	0.317	0.029
rds/peripherin (rds38)	-1258.75	-1256.33	4.84	0.0278*	0.2647	0.213	0.106
Reticulon	-1139.23	-1133.82	10.82	0.0010*	0.3136	0.042	0.045
Rhodopsin	-1265.44	-1265.41	0.06	1.0000	0.1896	0.094	0.030
Ringo (p33 ringo, ik26) (rapid inducer of G 2/M	-1138.75	-1138.13	1.24	0.2648	0.0455	0.475	0.174
	-1135.21	-1134.10	2.23	1.0000	0.1928	0.368	0.150

Table 1.3 (continued 5)

RIO kinase 2	-2055.42	-2058.40	0.04	0.8403	0.1941	0.201	0.116
Rwdd1 (RWD domain containing 1)	-863.37	-862.67	1.40	0.2370	0.2657	0.133	0.084
Retinal homeobox A	-1075.35	-1074.97	0.75	0.3849	0.0703	0.114	0.080
Rxrb (retinoid X receptor beta)	-1399.48	-1399.17	0.62	0.4306	0.0581	0.125	0.072
Sister chromatid cohesion establishment	-856.35	-856.30	0.10	1.0000	0.1417	0.214	0.151
syndecan 2 (heparan sulfate proteoglycan 1,	-626.00	-625.95	0.10	1.0000	0.1871	0.278	0.063
Sek-1 receptor tyrosine kinase	-3004.62	-3004.59	0.07	1.0000	0.063	0.048	0.005
Selenoprotein I	-1370.88	-1370.18	1.39	1.0000	0.2496	0.178	0.067
Selenoprotein T	-442.66	-442.58	0.15	1.0000	0.1943	0.120	0.000
Septin 11	-1010.26	-1010.26	0.00	1.0000	0.0516	0.067	0.054
septin A (XlSeptA)	-1096.18	-1096.18	0.00	1.0000	0.0781	0.073	0.000
serum/glucocorticoid regulated kinase	-1376.08	-1369.08	14.00	0.0002*	0.0102	0.143	0.050
Shab12 (delayed rectifier potassium ion	-565.96	-565.87	0.18	0.6756	0.0761	0.058	0.017
Siah-interacting protein	-811.15	-810.14	2.02	0.1555	0.3826	0.230	0.092
Sloan-Kettering viral oncogene homolog	-2307.30	-2307.27	0.07	1.0000	0.128	0.154	0.050
Histone stem-loop binding protein (SLBP)	-987.69	-987.14	1.08	1.0000	0.2134	0.121	0.271
suc1-associated neurotrophic factor	-1757.22	-1757.20	0.05	0.8283	0.1378	0.242	0.094
Sox11 (XLS13)	-1242.28	-1242.11	0.34	0.5600	0.0678	0.196	0.059
Sox17a (HMG box transcription factor	-1376.31	-1373.93	4.77	0.0290*	0.1106	0.245	0.132
Sox18 (Transcription factor SOX-18)	-1103.03	-1102.73	0.60	1.0000	0.1917	0.131	0.085
SP22	-549.87	-549.77	0.20	1.0000	0.0465	0.082	0.171
Sparc	-1017.59	-1015.41	4.35	0.0370*	0.2467	0.101	0.044
Spats2 (spermatogenesis	-1964.55	-1962.01	5.08	0.0242*	0.4294	0.100	0.160
Spermatid perinuclear RNA binding protein	-1633.67	-1631.70	3.93	0.0474*	0.297	0.213	0.116
Sprouty-2	-1055.90	-1055.30	1.19	0.2745	0.2637	0.204	0.132
Sulfide quinone reductase-like (yeast)	-1522.53	-1522.47	0.10	0.7464	0.1095	0.177	0.060
Src (pp60c-src protein)	-1645.55	-1645.50	0.09	1.0000	0.057	0.066	0.013
Stanniocalcin 1	-881.57	-880.69	1.76	0.1847	0.1693	0.108	0.070
Staufen 1	-1450.62	-1450.60	0.06	0.8138	0.1882	0.161	0.100
Stress-induced-phosphoprotein 1	-1837.99	-1837.77	0.44	1.0000	0.1874	0.130	0.061
Stomatin	-455.91	-455.56	0.68	0.4090	0.1148	0.030	0.046
Strabismus	-1558.21	-1554.05	8.33	0.0039*	0.0455	0.000	0.005
SUG1	-1166.75	-1166.58	0.34	0.5597	0.0156	0.011	0.007
translation initiation factor SUI1	-319.22	-319.22	0.00	1.0000	0.0001	0.000	0.000
Sumo	-302.70	-300.61	4.19	0.0407*	0.2404	0.000	0.000
Survivin (Xsvv1)	-548.98	-548.91	0.14	1.0000	0.1112	0.133	0.136
Synaptobrevin	-349.33	-347.27	4.13	0.0422*	0.2793	0.000	0.060
Synaptophysin	-1011.08	-1010.96	0.23	0.6344	0.2607	0.144	0.076
Xwnt8 inhibitor sizzled (szl)	-970.36	-969.81	1.10	1.0000	0.1416	0.135	0.154
TAF-Ibeta	-795.51	-795.51	0.00	0.9956	0.0464	0.054	0.033
T-box transcription factor Tbx5	-365.56	-365.56	0.01	0.9145	0.0548	0.034	0.000
TCRzeta subunit	-656.30	-655.90	0.79	1.0000	0.2344	0.225	0.258
Bax Inhibitor-1, testis enhanced gene	-779.13	-779.03	0.21	0.6462	0.3343	0.179	0.034
TRK-fused protein TFG	-1256.21	-1256.19	0.06	1.0000	0.1152	0.108	0.109
Thyroid Hormone Receptor alpha*	-1309.97	-1308.31	3.32	0.0686	0.0517	0.110	0.043
Thyroid Hormone Receptor beta*	-1137.02	-1136.78	0.48	0.4881	0.0654	0.101	0.022
Mesoderm Posterior (Mesp)	-1190.88	-1190.84	0.09	1.0000	0.2265	0.267	0.224
cytotoxic granule-associated RNA binding	-1215.41	-1214.22	2.38	0.1229	0.0738	0.033	0.016
TIAR	-1259.56	-1258.98	1.17	0.2802	0.0523	0.174	0.026
Tyrosine kinase	-508.30	-508.10	0.39	0.5338	0.0756	0.324	0.000
IGF (Insulin-like Growth Factor) Receptor	-483.55	-483.03	1.05	0.3056	0.0507	0.145	0.033

Table 1.3 (continued 6)

Transducer of erbB	-939.50	-936.92	5.16	0.0231*	0.0319	0.119	0.079
Transferrin	-2744.61	-2742.98	3.25	0.0714	0.2793	0.244	0.236
Thyrotropin-releasing Hormone	-1001.41	-1001.15	0.51	0.4731	0.6228	0.509	0.293
Thyrotropin-releasing Hormone Receptor 1	-1271.59	-1270.76	1.67	0.1961	0.139	0.060	0.044
Neurotrophin receptor B xTrkB-alpha	-1704.67	-1704.04	1.26	0.2620	0.3707	0.169	0.181
fast skeletal Troponin C	-447.63	-446.93	1.39	1.0000	0.025	0.000	0.040
unitary non-NMDA glutamate receptor	-1625.32	-1625.29	0.06	0.8108	0.1173	0.106	0.080
Ubiquiton-conjugating enzyme e2e	-647.60	-647.60	0.00	0.9881	0.3764	0.108	0.099
xUBF mRNA for upstream binding factor	-2230.17	-2230.17	0.01	0.9227	0.0889	0.100	0.087
endoplasmic reticulum UDP-Glc/UDP-Gal	-1107.87	-1107.63	0.48	0.4895	0.0561	0.139	0.083
UDP-glucose ceramide glucosyltransferase	-1212.71	-1212.70	0.03	0.8731	0.0324	0.073	0.000
Uroplakin 1A	-716.72	-716.67	0.09	0.7593	0.095	0.075	0.054
Ubiquinol-cytochrome C reductase complex.	-1530.74	-1530.64	0.19	0.6647	0.202	0.127	0.116
Vasodilator-stimulated phosphoprotein	-1235.12	-1233.74	2.76	0.0968	0.3691	0.181	0.125
Ventral anterior homeobox protein	-975.47	-974.31	2.31	1.0000	0.1796	0.231	0.088
Ventral anterior homeobox protein	-912.94	-912.73	0.43	0.5125	0.1763	0.287	0.102
Von Hippel-Lindau binding protein 1	-603.05	-602.90	0.30	0.5818	0.1759	0.181	0.027
Vq1 RNA binding protein	-1805.61	-1805.29	0.65	0.4216	0.055	0.105	0.018
Vimentin*	-1625.30	-1624.60	1.40	0.2366	0.3532	0.136	0.120
Tryptophanyl-tRNA synthetase	-1700.97	-1700.92	0.09	1.0000	0.1188	0.169	0.128
Wee1A kinase	-1913.25	-1912.26	1.99	0.1586	0.0771	0.228	0.077
Wee1B, Wee1-like protein kinase	-1089.34	-1089.34	0.00	1.0000	0.08	0.066	0.032
Uterine sensitization-associated protein-1	-766.25	-765.82	0.85	1.0000	0.1458	0.173	0.138
Xwnt-3	-415.24	-415.13	0.21	0.6454	0.0341	0.115	0.000
Wilms' tumor suppressor (WT1)	-1209.61	-1209.45	0.32	0.5730	0.0995	0.036	0.037
Cofilin (XAC)	-595.44	-595.26	0.37	0.5422	0.2643	0.0812	0.125
XE2 (helix-loop-helix transcription factor E2)	-362.66	-362.13	1.05	0.3049	0.1487	0.1117	0.045
Xefiltin	-1745.75	-1742.77	5.97	0.0146*	0.1503	0.1932	0.070
Epidermis specific serine protease Prss27	-1622.08	-1622.08	0.00	1.0000	0.4139	0.3615	0.468
Fork head related (XFD1)	-1398.08	-1397.93	0.30	1.0000	0.257	0.2408	0.123
Fork head protein (XFD2)	-1445.08	-1441.53	7.11	0.0077*	0.3813	0.1054	0.176
Interleukin-1 beta-converting enzyme	-1733.46	-1733.32	0.27	0.6020	0.4472	0.6068	0.526
Xlmb (maternal B9.10 and B9.15 protein)	-879.69	-878.73	1.91	0.1668	0.0777	0.2514	0.109
L-myc oncogene (xL-myc)	-1225.21	-1224.63	1.16	0.2809	0.2824	0.255	0.060
Xnot (homeobox protein)	-648.13	-646.91	2.44	1.0000	0.2276	0.223	0.031
TGF-beta related growth factor Xnr-4 (Xnr4)	-1500.96	-1498.69	4.53	0.0334*	0.3928	0.3794	0.269
XmF12	-1907.36	-1907.36	0.00	1.0000	0.1525	0.1535	0.188
Xrpf (XrpFI beta 1) GA binding protein	-1212.90	-1212.86	0.08	1.0000	0.1375	0.0688	0.034
ZFTF (zinc finger transcription factor	-821.66	-817.49	8.35	0.0039*	0.0001	0.1058	0.034
ZPB (zona pellucida glycoprotein)	-1776.40	-1773.26	6.27	0.0123*	0.4215	0.05	0.216

Table 1.4. Tests for complementary patterns of substitution using the paralog heterogeneity test and runs test for dichotomous variables on nonsynonymous and synonymous substitutions. P values of paralog heterogeneity tests are reported for two (P2) or three domains (P3). Tests were not conducted if there were a small number of substitutions in one paralog (see text); individually significant values ($P < 0.05$) are indicated with an asterisk.

Table 1.4 (continued 1)

	nonsynonymous substitutions			synonymous substitutions		
	paralog heterogeneity test		Runs	paralog heterogeneity test		Runs
Gene name	P2	P3	test	P2	P3	test
Actin (skeletal, alpha 3)*	-	-	-	0.7400	0.1220	0.0789
Activin Receptor-Like Kinase-2 (ALK-2)	0.1000	0.0920	0.4093	0.9950	0.9860	0.6453
Activin receptor II	0.9820	0.5800	0.7317	0.9900	0.9770	0.2607
Adipophilin (fatvg)	0.8370	0.9420	0.4892	0.9040	0.0980	0.3024
AE (Amidating Enzyme)	0.4300	0.0150*	0.3314	0.6140	0.2600	0.0325*
Albumin (serum)*	0.0670	0.1480	0.0978	0.0340*	0.0270*	0.0030*
ALDH (Aldehyde dehydrogenase class1)	0.4040	0.5930	0.0245*	0.9310	0.9280	0.7483
Alpha Globin	0.0580	0.1290	0.3911	0.1130	0.1110	0.9449
Amelogenin	0.9460	0.9490	0.7877	0.3210	0.5720	0.6611
Xenopus Anterior Neural Folds, Homeobox gene	0.6340	0.2430	0.8382	0.5055	0.5850	0.1097
Amyloid-Beta-like protein precursor	0.1890	0.0690	0.7964	0.9470	0.8260	0.8311
Apoptosis Inhibitor 5	0.5590	0.7620	0.5438	0.0680	0.0000*	0.2401
AR (Androgen Receptor)	-	-	-	-	-	-
Liver L-arginase	-	-	-	-	-	-
Arginase Type 2	0.9605	0.3690	0.9160	0.0420*	0.1330	0.0263*
Arrestin	0.8600	0.9520	0.0744	0.2410	0.3350	0.5997
Aspartyl tRNA synthetase	0.0390*	0.0500	0.7193	0.8010	0.7460	0.6021
Atonal Homolog 5	0.4560	0.6530	0.9517	0.3490	0.4740	0.2661
ATP synthase subunit B	0.2150	0.3730	0.8601	0.4210	0.4670	0.6387
Bambi (TGF-beta signalling)	0.1640	0.1990	0.2208	0.1110	0.4840	0.7151
Barren (brm1, 13S condensin XCAP-H subunit)	0.2580	0.1580	0.4487	0.6180	0.7850	0.5241
Bestrophin-2 (VMD2L1)	0.6130	0.3680	0.6110	0.3750	0.5800	0.0398*
Beta Globin	0.8320	0.2550	0.6675	0.2020	0.3870	0.0539
Complement factor B (Bf B) (MHC class III gene)	0.1060	0.2570	0.0263*	0.7600	0.3280	0.7251
Biglycan	0.8110	0.8570	0.2370	0.7615	0.9020	0.9060
Bicaudal-C	0.0990	0.1840	0.4466	0.0260*	0.0580	0.2421
Bridging integrator 1 (Amphiphysin II)	0.4140	0.0590	0.0489*	0.2790	0.1000	0.0159*
Bmp4	0.5030	0.3060	0.8368	0.8370	0.7830	0.3028
BMP (Bone Morphogenetic Protein) receptor	0.7950	0.4630	0.0825	0.9640	0.9340	0.5981
Block of proliferation 1	0.8470	0.9680	0.4702	0.9220	0.9720	0.5475
Brachyury (T)	0.7340	0.8290	0.8311	0.1200	0.2140	0.9680
Serine/Threonine protein kinase (c-RMIL)	0.3610	0.3880	0.0789	0.0970	0.3110	0.3378
Basic transcription element binding protein	-	-	-	0.1560	0.0730	0.5217
B-cell translocation gene 1, anti-proliferative	0.0020*	0.0070*	0.2248	0.9360	0.9760	0.4624
Calcium homeostasis endoplasmic reticulum protein	0.3930	0.6020	0.5054	0.9830	0.9280	0.5813
Calnexin	0.1270	0.0410*	0.1412	0.1050	0.3100	0.4393
Calponin H3 (clpH3)	0.6890	0.1180	0.8006	0.5410	0.3940	0.0332*
Calreticulin	0.8660	0.6940	0.5890	0.0820	0.1230	0.3719
Carbonic anhydrase II	0.7540	0.8250	0.8223	0.1570	0.2430	0.2826
Casein kinase I alpha S (Csnk1a1)	-	-	-	0.6480	0.8260	0.3338
CASK interacting protein 2	0.7150	0.3970	0.6820	0.8890	0.9810	0.3709
Procathepsin B	0.2950	0.5270	0.1503	0.6160	0.8300	0.8752
Beta Catenin interacting protein 1 (catnbip1)	-	-	-	-	-	-
Cystathionine-beta-synthase	0.9055	0.9740	0.8859	0.3290	0.2900	0.0220*
voltage-dependent Calcium channel beta subunit	0.2210	0.3700	0.7704	0.0030*	0.0010*	0.0016*
CDC2 (cell division cycle 2, kinase)	0.7550	0.7275	0.6180	0.4510	0.6670	0.8004

Table 1.4 (continued 2)

Cathepsin E	0.7390	0.8230	0.2405	0.6390	0.8170	0.6720
Carboxyl ester lipase	0.4060	0.3060	0.1230	0.9860	0.9820	0.4987
Carboxyl ester lipase (bile salt-stimulated lipase)	0.2600	0.4690	0.1369	0.1220	0.2550	0.6596
Centrin	-	-	-	0.0280*	0.0640	0.7910
Cerebellin 2 precursor protein	0.9960	0.4440	0.2897	0.7420	0.8510	0.5142
Complement factor I (C3b/C4b inactivator)	0.5450	0.1890	0.6225	0.1150	0.3070	0.9191
Cystic fibrosis transmembrane conductance regulator	0.2380	0.1230	0.0391*	0.9970	0.2110	0.7162
Cortical granule lectin	0.9030	0.0590	0.6410	0.1875	0.0600	0.9516
Choroideremia (Rab escort protein 1)	0.9690	0.5580	0.3615	0.9080	0.8290	0.3702
Carbohydrate sulfotransferase 11 (Chst11)	0.0110*	0.0050*	0.0881	0.0850	0.1210	0.1118
Cell death-inducing DFFA-like effector c (CIDE-3alpha)	0.8900	0.0860	0.1297	0.3970	0.4850	0.6535
C-Jun proto-oncogene (AP-1, Activator Protein)	0.5330	0.3180	0.2779	0.8340	0.8780	0.8821
Dipeptidase 2 (metallopeptidase M20 family)	0.0820	0.0680	0.4265	0.9720	0.6230	0.2856
alpha-1 Collagen type II	0.4680	0.6020	0.2774	0.7740	0.4640	0.3927
Connexin 31 (Gap junction beta-3 protein)	0.4560	0.7090	0.5085	0.3150	0.5610	0.1115
Contactin/F3/F11 (Contactin A)	0.6620	0.6330	0.2701	0.4030	0.7100	0.3679
Coronin	0.6960	0.7560	0.1042	0.1840	0.0840	0.6929
Cortactin	0.3380	0.2270	0.1984	0.3250	0.2430	0.2439
Cytoplasmic polyadenylation element binding protein	0.5150	0.5510	0.1509	0.0810	0.1160	0.5035
CRY2 (cryptochrome 2)	0.6050	0.7140	0.9125	0.8280	0.3240	0.3694
Crystallin, beta A1	0.9790	0.2740	0.7393	0.0470*	0.1710	0.3194
Cathepsin S (CTSS)	0.2250	0.3150	0.3667	0.9040	0.9180	0.5787
Cullin3 (Cul3)	-	-	-	0.9600	0.9800	0.1359
CyclinE	0.7740	0.8660	0.4828	0.2905	0.5040	0.9573
Brain Dopamine receptor D2	0.1905	0.3920	0.9106	0.7660	0.9070	0.6155
Dapper 1, antagonist of beta-catenin (Frodo)	0.1110	0.2790	0.0330*	0.1670	0.3930	0.1575
Death-associated protein kinase 1	0.1160	0.2520	0.7031	0.1680	0.2480	0.6978
Drebrin-like	0.5960	0.3390	0.7424	0.4060	0.7530	0.7693
Debranching enzyme homolog 1	0.0840	0.0230*	0.1703	0.2750	0.3600	0.6467
Deleted in colorectal cancer tumor suppressor	0.7980	0.8880	0.2120	0.0390*	0.0870	0.6148
Desmin	0.3950	0.1030	0.7144	0.2150	0.3320	0.3864
Hand2	-	-	-	0.0870	0.1760	0.3609
Cytoplasmic dynein light-intermediate chain 1 (DLIC1)	0.1120	0.3120	0.2669	0.0450*	0.0150*	0.4997
Dipeptidylpeptidase 3	0.4620	0.3870	0.9269	0.4830	0.6660	0.6671
Dullard	-	-	-	0.9970	0.2390	0.7573
Dystroglycan (DAG1)	0.0200*	0.0180*	0.4725	0.9880	0.9720	0.6865
Dystrophin	-	-	-	-	-	-
Helix-loop-helix transcription factor XE1	0.7510	0.8880	0.4700	0.4400	0.7140	0.5255
E2 (transcription factor E2)	0.0750	0.1690	0.4647	0.0810	0.0830	0.6111
met-mesencephalon-olfactory transcription factor 1 (Ebf2)	0.7320	0.1890	0.1397	0.6830	0.8520	0.2250
CCAAT/enhancer binding protein (C/EBP), alpha	0.8830	0.9490	0.8050	0.1540	0.4400	0.7463
Endothelin receptor type A	0.8160	0.8900	0.9590	0.2060	0.0520	0.3882
EF (Elongation Factor-1 alpha, 42Sp48)	-	-	-	0.6735	0.6400	0.6702
Aurora kinase A (EG2)	0.3510	0.6200	0.7224	0.1830	0.0170*	0.6086
Engrailed 2 (EN2)	0.0560	0.1510	0.0385*	0.2880	0.4300	0.1781
Enkephalin A (proenkephalin A)*	0.3460	0.1660	0.1074	0.0060*	0.0230*	0.3346
ENO (2-phosphoglycerate dehydratase, enolase)	-	-	-	0.6460	0.6550	0.0764
Era (Estrogen Receptor alpha)	0.8590	0.9560	0.8939	0.2690	0.0540	0.1565
Enhancer of split groucho	-	-	-	0.9710	0.9840	0.6392
Enhancer of zeste	0.0590	0.1220	0.6242	0.2060	0.1450	0.2269
Focal adhesion kinase	0.4990	0.6620	0.4403	0.9550	0.9960	0.8084

Table 1.4 (continued 3)

Transcription factor (XLFB1)	0.0065*	0.0135*	0.6450	0.0510	0.0860	0.4077
XFD-4	0.0640	0.2600	0.1903	0.0490*	0.0100*	0.1606
Flap endonuclease-1	0.4820	0.1360	0.1732	0.1800	0.1740	0.7857
FetuinB	0.2430	0.0460*	0.1642	0.0970	0.1785	0.8815
Ftz-F1-related orphan receptor (xFF1r)	-	-	-	0.2580	0.3030	0.7223
FGF (embryonic fibroblast growth factor 4)	0.0370*	0.1190	0.0161*	0.8960	0.9440	0.4253
Fibroblast growth factor receptor	0.1220	0.0390*	0.0096*	0.1170	0.2820	0.3214
Fibrinogen alpha	0.9360	0.8470	0.6547	0.0790	0.1480	0.9566
Flotillin	0.2900	0.3810	0.5994	0.1060	0.1100	0.7183
fms-related tyrosine kinase 1	0.1570	0.0860	0.1745	0.9250	0.9295	0.8388
Fms-interacting protein (NF2/meningioma region)	0.1090	0.2410	0.8749	0.4720	0.1460	0.1974
alpha-fodrin (Xen alpha 1) spectrin, non-erythrocytic 1	0.7290	0.2140	0.8795	0.2800	0.1920	0.4670
c-fos proto-oncogene	-	-	-	0.9110	0.7700	0.5726
Succinate dehydrogenase complex, Flavoprotein	0.2180	0.5340	0.4492	0.2970	0.3850	0.0920
Frequenin	-	-	-	0.5350	0.6050	0.2395
Fascin	0.9360	0.2290	0.6556	0.4370	0.7640	0.4719
Funn*	0.7300	0.8900	0.6233	0.2700	0.0210*	0.2427
Fused toes homolog	0.6440	0.9130	0.4064	0.8050	0.1930	0.2861
FYN (proto-oncogene c-fyn)	-	-	-	0.6440	0.8310	0.6285
Galectin	0.0120*	0.0240*	0.0067*	0.6710	0.7610	0.5696
alpha subunit of Gq Gtp-binding protein (G protein)	-	-	-	0.0850	0.0090*	0.0896
GATA-binding protein transcription factor GATA-1	0.6110	0.7870	0.5850	0.9690	0.7575	0.7088
Transcription factor xGata5	0.1780	0.3270	0.5439	0.8810	0.8680	0.5895
Growth hormone A	-	-	-	0.3020	0.4050	0.9991
Guanylate kinase 1	0.6450	0.5520	0.7104	0.3850	0.0610	0.1831
Glycogenin 1 (mitotic phosphoprotein 45)	0.2920	0.4810	0.0168*	0.4530	0.5320	0.0426*
Holocytochrome c synthase (heme-lyase) (hccs-prov)	0.8270	0.8880	0.6069	0.0000*	0.0030*	0.1192
cephalic Hedgehog, sonic hedgehog protein 4	0.1360	0.2550	0.7561	1.0000	0.2430	0.2291
Transcription factor XHEN1	0.5965	0.8335	0.8002	0.4720	0.3660	0.6467
Hypoxia-inducible factor 1 alpha	0.3860	0.4590	0.4948	0.2880	0.3720	0.3111
SafA - scaffold attachment factor A	0.0830	0.0400*	0.5762	0.0020*	0.0320*	0.9182
Homeobox 2/2.3*	0.2830	0.1520	0.4867	0.0430*	0.0730	0.6195
Insulin*	0.2435	0.3705	0.2870	0.7440	0.1270	0.3927
Integrin beta-1 subunit*	0.2110	0.0120*	0.1276	0.0770	0.2850	0.7990
Inversin	0.3760	0.1510	0.0062*	0.5350	0.8610	0.7457
Ubiquitin carboxyl-terminal hydrolase 5 (Isopeptidase T)	0.3420	0.0890	0.0415*	0.9170	0.3930	0.2310
Kf-1 protein (Adqr34)	0.4400	0.5920	0.4633	0.2820	0.0440*	0.3209
Kit receptor tyrosine kinase (c-kit)	0.9840	0.9720	0.3527	0.3110	0.6370	0.4923
Kinesin-like protein 2	0.3860	0.5980	0.8574	0.7350	0.2540	0.4116
L1 (ribosomal protein L1)	-	-	-	0.2470	0.5680	0.7042
L14 (ribosomal protein L14)	0.3705	0.6370	0.5735	0.8860	0.9290	0.4501
Lamin B	0.1100	0.0380*	0.3943	0.9400	0.9240	0.7441
Lamina associated polypeptide 2	0.9880	0.6750	0.1766	0.5910	0.6820	0.8860
Clathrin, light polypeptide (Lcb)	0.8260	0.9280	0.6637	0.2010	0.2860	0.0051*
Lactate dehydrogenase	0.7910	0.7760	0.4370	0.0770	0.1600	0.0327*
LEF-1 (lymphoid enhancer factor)	0.0390*	0.1130	0.7870	0.0240*	0.0650	0.8252
TGF-beta family member Leftv-A	0.0600	0.1750	0.0729	0.9070	0.1470	0.1087
LIM domain binding protein	-	-	-	0.3020	0.4630	0.1558
Lipocalin (Ptgds)	0.1710	0.2980	0.4336	0.4170	0.3150	0.0834
Lpa1R (lysophosphatidic acid receptor)	-	-	-	0.3450	0.0910	0.0093*
LR (Leptin Receptor)	-	-	-	0.7390	0.5360	0.1189

Table 1.4 (continued 4)

Lipoprotein (LDL) receptor-related protein 6	-	-	-	0.9160	0.1930	0.6215
Autoantigen La (La protein)	0.2540	0.0880	0.3856	0.2680	0.3840	0.8943
Microfibrillar-associated protein 1	0.7820	0.9060	0.9091	0.3995	0.6050	0.8731
Myristoylated alanine-rich C kinase substrate	0.0430*	0.1350	0.4264	0.9825	0.9830	0.4463
XMax2 and XMax4	-	-	-	0.0955	0.1825	0.4937
Myogenin	0.7540	0.8170	0.9276	0.0150*	0.0470*	0.1267
Myozenin1	0.9160	0.4210	0.7304	0.4280	0.3700	0.4516
N-CAM (neural cell adhesion molecule)*	0.3210	0.0610	0.2543	0.8070	0.2460	0.1001
NF-M1 (middle molecular neurofilament)	0.7300	0.6570	0.7659	0.1270	0.3350	0.7672
neurogenin-related 1 (X-NGNR-1)	0.8650	0.9730	0.9218	0.3230	0.3000	0.7386
Internexin neuronal intermediate filament protein	0.0510	0.0160*	0.0053*	0.7430	0.7570	0.2816
NK3 transcription factor related, koza	0.4280	0.6400	0.0406*	0.0370*	0.0590	0.0399*
Nonmuscle myosin II heavy chain A	0.9980	0.9470	0.3710	0.1110	0.0800	0.2255
Nonmuscle myosin heavy chain B	-	-	-	0.0930	0.0580	0.9894
Nucleolar-localized protein NO38	0.9130	0.9320	0.7492	0.8350	0.7310	0.3799
Nucleobindin 1	-	-	-	0.4760	0.6070	0.4252
Nucleoplasmin	0.9280	0.8990	0.9963	0.9600	0.9230	0.8325
Nucleoporin (Nup88)	0.7300	0.8010	0.4510	0.2570	0.5520	0.2907
OLPA (Dorphin)	0.4560	0.6580	0.1583	0.9470	0.9780	0.5285
Olfactory marker protein (XOMP)	0.1330	0.2650	0.5306	0.5180	0.7460	0.0215*
OncogenesC-ets-1 (c-ets-1 proto-oncogene)*	0.4370	0.5880	0.8562	0.3230	0.4850	0.7489
OncogenesC-ets-2 (c-ets-2 proto-oncogene)*	0.1075	0.3050	0.2922	0.9440	0.6940	0.9549
OncogenesC-myc (myelocytomatosis)*	0.8950	0.8600	0.6704	0.8700	0.3820	0.6145
Dynactin 2 (p50)	0.7620	0.8890	0.4371	0.6740	0.1360	0.1604
PACSIN2	0.6450	0.5570	0.4387	0.1490	0.0300*	0.7350
Convertase PC2	0.9310	0.9200	0.8691	0.9770	0.7990	1.0000
Prolyl isomerase (Pin1)	0.9520	0.9340	0.4994	0.8320	0.9180	0.9834
PKC (protein kinase C delta)	0.8900	0.9675	0.3247	0.5820	0.7520	0.7905
Plakoglobin	0.9010	0.9790	0.8255	0.9120	0.9540	0.9874
Peripheral myelin protein 22	-	-	-	0.0490*	0.0170*	0.0531
POMC (pro-opiomelanocortin)*	0.1680	0.2990	0.9937	0.0590	0.1780	0.5733
POU domain Gene 1	-	-	-	0.6580	0.8700	0.5370
POU3	-	-	-	0.0730	0.1500	0.5944
Phosphorylase phosphatase (Ppp2B)	-	-	-	0.4770	0.0750	0.8485
Protein phosphatase 4, regulatory subunit 2 (Ppp4r2)	0.3360	0.4300	0.3293	0.8360	0.8900	0.1810
LIM protein Prickle	0.5210	0.5850	0.2356	0.6660	0.7960	0.1445
Prolactin Receptor	0.8950	0.9400	0.5260	0.3670	0.6460	0.0834
Prothymosin	0.6825	0.7520	0.9170	0.9450	0.2720	0.1542
Phosphorylase, glycogen, brain	0.7800	0.2840	0.8804	0.0720	0.2070	0.5815
RAB18 (member RAS oncogene family)	-	-	-	0.1870	0.3570	0.5102
Rac GTPase	-	-	-	0.4420	0.4800	0.4481
Rad51	-	-	-	0.0780	0.2670	0.0483*
Raq-1	0.4860	0.7290	0.4375	0.1530	0.3000	0.0195*
Ral interacting protein (rlip) - RalA (RalB-binding protein)	0.6390	0.7060	0.9260	0.3500	0.6870	0.6030
RalB	-	-	-	0.4960	0.7190	0.8398
Retinoic acid receptor alpha	0.8690	0.9370	0.0443*	0.1050	0.0170*	0.5800
RDS35 (retinal degradation slow/peripherin)	0.1450	0.0480*	0.0889	0.8570	0.9360	0.5515
RDS38/peripherin	0.4350	0.6150	0.2368	0.2150	0.0060*	0.5074
Requiem	0.8220	0.8800	0.9277	0.8580	0.7810	0.1581
Rhodopsin	-	-	-	0.7550	0.1360	0.8260
Ringo (p33 ringo, ls26)	0.4890	0.6930	0.3147	0.9060	0.8100	0.2189

Table 1.4 (continued 5)

RIO kinase 2	0.0040*	0.0080*	0.1706	0.1570	0.3850	0.7642
Rwdd1 (RWD domain containing 1)	0.9650	0.9300	0.6886	0.9830	0.3590	0.5107
Retinal homeobox A	0.6920	0.7470	0.8166	0.1990	0.4380	0.9749
Rxrb (retinoid X receptor beta)	0.0690	0.1790	0.1411	0.6290	0.8350	0.9051
Sister chromatid cohesion establishment factor (SCC2)	0.1740	0.3000	0.1234	0.8160	0.7580	0.9521
Syndecan 2 (heparan sulfate proteoglycan 1)	0.0560	0.1240	0.6430	0.6160	0.7800	0.0112*
Sek-1 receptor tyrosine kinase	0.4450	0.2700	0.7852	0.6240	0.7470	0.5719
Selenoprotein I	0.9560	0.3960	0.6264	0.0890	0.1000	0.0917
Selenoprotein T	0.5485	0.4950	0.6713	0.5470	0.5950	0.7394
Septin 11	0.7220	0.8160	0.4990	0.8600	0.8290	0.7254
Septin A (XlSeptA)	0.4365	0.6625	0.8783	0.3250	0.2040	0.1425
serum/glucocorticoid regulated kinase	-	-	-	0.7270	0.6350	0.8501
Shab12	0.5655	0.6585	0.7982	0.2110	0.4150	0.5980
Siah-interacting protein	0.7560	0.9400	0.7691	0.8070	0.8500	0.9484
Sloan-Kettering viral oncogene homolog	0.2695	0.5020	0.9736	0.5360	0.6530	0.3125
Histone stem-loop binding protein (SLBP)	0.6610	0.6700	0.5797	0.0030*	0.0020*	0.8836
Suc1-associated neurotrophic factor target	0.9550	0.6790	0.3862	0.7130	0.8240	0.6105
Sox11 (XLS13)	0.9730	0.7600	0.8218	0.9280	0.8150	0.6637
Sox17a (HMG box transcription factor Sox17-alpha)	0.9110	0.9370	0.6604	0.1985	0.3790	0.0410*
Sox18 (Transcription factor SOX-18)	0.3730	0.3430	0.8661	0.1230	0.2990	0.0399*
SP22	-	-	-	0.8350	0.5630	0.3881
Sparc	0.7860	0.2890	0.1551	0.8310	0.9340	0.7497
Spats2 (spermatogenesis associated, serine-rich 2)	0.2910	0.2240	0.1836	0.2930	0.4000	0.5665
Spermatid perinuclear RNA binding protein	0.2420	0.1660	0.2982	0.8430	0.9310	0.7468
Sprouty-2	0.0490*	0.1910	0.1259	0.5870	0.7770	0.3808
Sulfide quinone reductase-like	0.0730	0.1980	0.8676	0.2700	0.2460	0.0742
Src (pp60c-src protein)	0.2630	0.4510	0.6295	0.4295	0.7140	0.5936
Stanniocalcin 1	0.5290	0.3150	0.2126	0.6770	0.7030	0.9394
Staufen 1	0.6000	0.8130	0.2474	0.2915	0.4800	0.8918
Stress-induced-phosphoprotein 1	0.4100	0.5790	0.6479	0.0090*	0.0680	0.6135
Stomatin	-	-	-	0.8120	0.3930	0.8533
Strabismus	-	-	-	0.7430	0.8270	0.6505
SUG1	-	-	-	0.8220	0.8380	0.7509
translation initiation factor SUI1	-	-	-	0.2230	0.4310	0.3848
Sumo	-	-	-	0.7760	0.6540	0.8864
Survivin (Xsvv1)	0.2610	0.3770	0.1277	0.0610	0.0760	0.0150*
Synaptobrevin	-	-	-	0.0190*	0.0550	0.0049*
Synaptophysin	0.0700	0.1200	0.4390	0.0220*	0.0050*	0.6087
Xwnt8 inhibitor szzled (szl) (putative wnt inhibitor frzb3)	0.5600	0.2660	0.1614	0.7100	0.8370	0.6371
TAF-Ibeta	-	-	-	0.2660	0.4250	0.1651
T-box transcription factor TbxS	-	-	-	0.0900	0.1310	0.6234
TCRzeta subunit	0.6870	0.1720	0.8223	0.1470	0.2230	0.8226
Bex Inhibitor-1, testis enhanced gene transcript	0.9190	0.4110	0.0835	0.8650	0.9540	0.5544
TRK-fused protein TFG	0.9500	0.9570	0.7025	0.8050	0.9040	0.7929
Thyroid Hormone Receptor alpha*	-	-	-	0.2570	0.3870	0.0537
Thyroid Hormone Receptor beta*	0.0170*	0.1330	0.4844	0.9530	0.7670	0.7169
Mesoderm Posterior (Mesp)	0.4630	0.3260	0.9666	0.8170	0.7180	0.4615
cytotoxic granule-associated RNA binding protein (TIA1)	0.2810	0.0570	0.1501	0.0990	0.2030	0.1358
TIAR	0.5270	0.6340	0.3010	0.9210	0.7800	0.2212
Tyrosine kinase	0.4330	0.6350	0.1190	0.3170	0.5000	0.5804
IGF (Insulin-like Growth Factor) Receptor	-	-	-	0.1030	0.1840	0.1212

Table 1.4 (continued 6)

Transducer of erbB	-	-	-	0.8875	0.7240	0.6015
Transferrin	0.4500	0.2500	0.5726	0.0370*	0.0290*	0.7927
Thyrotropin-releasing Hormone	0.9500	0.5180	0.2479	0.7560	0.8420	0.3405
Thyrotropin-releasing Hormone Receptor 1	0.5320	0.0480*	0.7070	0.9910	0.9970	0.1205
Neurotrophin receptor B xTrkB-alpha	0.9490	0.9470	0.7177	0.2280	0.1330	0.0964
fast skeletal Troponin C	-	-	-	0.0770	0.2090	0.5216
unitary non-NMDA glutamate receptor subunit U1	0.7040	0.8490	0.2989	0.9430	0.4640	0.3726
Ubiquitin-conjugating enzyme e2e	0.4910	0.8440	0.4993	0.4000	0.3490	0.4857
xUBF mRNA for upstream binding factor	0.7980	0.8860	0.2699	0.8510	0.3910	0.8087
endoplasmic reticulum UDP-Glc/UDP-Gal transporter	-	-	-	0.0960	0.0690	0.7941
UDP-glucose ceramide glucosyltransferase	0.7925	0.8770	0.9496	0.6770	0.6770	0.4455
Uroplakin 1A	0.4600	0.5840	0.0444*	0.0720	0.0870	0.5115
Ubiquinol-cytochrome C reductase complex	0.6230	0.0700	0.1405	0.3330	0.1720	0.6988
Vasodilator-stimulated phosphoprotein	0.5510	0.6800	0.3386	0.1890	0.3460	0.1994
Ventral anterior homeobox protein (Vax1)	0.8400	0.3190	0.7649	0.8860	0.3990	0.8373
Ventral anterior homeobox protein (Vax2+3)	0.6090	0.5930	0.1202	0.7720	0.8220	0.7571
Von Hippel-Lindau binding protein 1	0.1505	0.1650	0.3194	0.7070	0.7910	0.5620
Vg1 RNA binding protein	0.5360	0.6320	0.8041	0.8670	0.5650	0.3260
Vimentin*	0.1470	0.3280	0.3921	0.7860	0.7540	0.5623
Tryptophanyl-tRNA synthetase	0.7810	0.3300	0.5103	0.6000	0.6890	0.0039*
Wee1A kinase	0.0970	0.2020	0.0660	0.2340	0.5120	0.9277
Wee1B, Wee1-like protein kinase	0.9840	0.7280	0.7931	0.7710	0.9380	0.1694
Uterine sensitization-associated protein-1 (Wise-A)	0.0420*	0.0810	0.1216	0.1900	0.1670	0.1914
Xwnt-3	-	-	-	0.3185	0.3690	0.4151
Wilms' tumor suppressor (WT1)	0.7780	0.8320	0.7098	0.1450	0.4540	0.3598
Cofilin (XAC)	0.1640	0.2510	0.9734	0.8820	0.8890	0.6422
XE2 (helix-loop-helix transcription factor E2)	-	-	-	0.7920	0.7570	0.2385
Xefiltin	0.4580	0.1710	0.0494*	0.0220*	0.0270*	0.0595
Epidermis specific serine protease Prss27 (Xepsin)	0.9450	0.2570	0.1439	0.9700	0.9770	0.4597
Fork head related (XFD1)	0.2590	0.3860	0.0902	0.9960	0.7630	0.5750
Fork head protein (XFD2)	0.3555	0.0660	0.7029	0.9400	0.9670	0.8595
Interleukin-1 beta-converting enzyme (Caspase 1)	0.1570	0.1240	0.8462	0.1510	0.3260	0.5056
Ximb (maternal B9.10 and B9.15 protein)	0.2820	0.3970	0.0270*	0.1200	0.1450	0.1260
L-myc oncogene (xL-myc)	0.2650	0.1680	0.1837	0.0250*	0.0070*	0.1856
Xnot (homeobox protein)	0.2560	0.1340	0.8942	0.0950	0.1770	0.5087
TGF-beta related growth factor Xnr-4 (Xnr4)	0.5660	0.0350*	0.2998	0.0730	0.1270	0.9123
XrnF12	0.1590	0.1970	0.7625	0.9740	0.9970	0.6248
Xrpf (XrpFI beta 1) GA binding protein transcription factor	0.1800	0.2550	0.0086*	0.1360	0.1890	0.3775
ZFTF (zinc finger transcription factor SLUG)	-	-	-	0.7740	0.2720	0.5955
ZPB (zona pellucida glycoprotein)	0.9910	0.6300	0.3384	0.7100	0.8575	0.6232

CHAPTER 2

Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization

Chain, F.J.J., D. Ilieva, and B.J. Evans (2008) *BMC Evolutionary Biology* **8:** 43.

PREFACE

The findings from chapter one led us to explore the possibilities (i) that our estimates of functional constraints have become obscured with time, and (ii) that duplicate gene expression divergence might play a prevalent role in preserving paralogs. Because we found most duplicate genes had similar functional constraints yet relaxed compared to a singleton ortholog, we examined the possibility that selective forces acting on them changed over time. We also compared duplicate gene expression in search for patterns consistent with temporal, spatial and quantitative subfunctionalization across 5 distinct developmental tissue types.

ABSTRACT

The mechanism by which duplicate genes originate – whether by duplication of a whole genome or of a genomic segment – influences their genetic fates. To study events that trigger duplicate gene persistence after whole genome duplication in vertebrates, we have analyzed molecular evolution and expression of hundreds of persistent duplicate gene pairs in allopolyploid clawed frogs (*Xenopus* and *Silurana*). We collected comparative data that allowed us to tease apart the molecular events that occurred soon after duplication from those that occurred later on. We also quantified expression profile divergence of hundreds of paralogs during development and in different tissues. Our analyses indicate that persistent duplicates generated by allopolyploidization are subjected to strong purifying selection soon after duplication. The level of purifying selection is relaxed compared to a singleton ortholog, but not significantly variable over a period spanning about 40 million years. Despite persistent functional constraints, however, analysis of paralogous expression profiles indicates that quantitative aspects of their expression diverged substantially during this period. These results offer clues into how vertebrate transcriptomes are sculpted in the wake of whole genome duplication (WGD), such as those that occurred in our early ancestors. That functional constraints were relaxed relative to a singleton ortholog but not significantly different in the early compared to the later stage of duplicate gene evolution suggests that the timescale for a return to pre-duplication levels is drawn out over tens of millions of years – beyond the age of these tetraploid species. Quantitative expression divergence can occur soon after WGD and with a magnitude that is not correlated with the rate of protein sequence divergence. On a coarse scale, quantitative expression divergence appears to be more prevalent than spatial and temporal expression divergence, and

also faster or more frequent than other processes that operate at the protein level, such as some types of neofunctionalization.

INTRODUCTION

Gene duplication can catalyze the evolution of novel function by providing a respite from purifying selection (Ohno 1970). The most common fate of a duplicated copy, however, is nonfunctionalization (pseudogenization), raising the question of how and why both copies of some duplicates manage to persist as functional entities. Interestingly, duplicate gene longevity is positively correlated with the scale of gene duplication – duplicate genes derived from whole genome duplication (WGD) typically persist for a longer period and evade pseudogenization at a higher frequency than those generated by segmental duplication (Amores et al. 1998; Ferris and Whitt 1979; Nadeau and Sankoff 1997; Wendel 2000). Therefore it appears that mechanisms that promote duplicate gene persistence in polyploid genomes are either different from or more effective than those that operate on duplicated genes generated by segmental duplication. This is probably because mechanisms specific to polyploid genomes, such as stoichiometric requirements / genic balance, increase their longevity (Freeling and Thomas 2006; Lynch and Conery 2000; Papp et al. 2003a; Veitia 2003), whereas characteristics specific to segmental duplicates, such as incomplete coding regions and regulatory elements decrease theirs (Katju and Lynch 2003). Furthermore, prezygotic isolating mechanisms could increase assortative mating within ploidy levels (Husband and Schemske 2000), facilitating speciation of polyploids and fixation of their duplicated genome in a new species. In clawed frogs, for example, second generation backcrossed hybrid females can produce a clutch comprised of fertile polyploid individuals of both sexes (Kobel and DuPasquier 1986; Kobel 1996). Sympatric speciation could be essentially instantaneous if these polyploid siblings interbreed and if reproductive incompatibilities exist between them and the lower ploidy parental species. In contrast, segmental duplicates begin as polymorphisms whose probability of fixation and time to fixation depend on genetic drift and natural selection (Clark 1994).

If stoichiometry is important, then an incentive immediately exists to preserve unadulterated versions of both copies of duplicates generated by WGD. Duplicate genes could also persist without functional change after duplication if overexpression is advantageous (Kondrashov et al. 2002; Larhammer and Risinger 1994), if there is selection against expression of a defective protein (Gibson and Spring 1998), or if neofunctionalized alleles were already segregating prior to duplication (Lynch et al. 2001). However, if neofunctionalizing mutations are rare or not very advantageous, or if population size is small, pre-duplication neofunctionalization is unlikely to be a common mechanism for duplicate gene persistence (Lynch et al. 2001; Walsh 1995), although clearly it has occurred (Dulai et al. 1999). Duplication could also facilitate the resolution of conflicts that arise from gene sharing – when two distinct protein phenotypes arise from the same

transcriptional unit – such as if an altered expression level is advantageous in one tissue but disadvantageous in another (Piatigorsky and Wistow 1991). In duplicates generated via WGD by allopolyploidization, heterosis from interactions between diverged subgenomes could contribute to duplicate gene longevity without necessitating altered function after duplication (Evans 2007).

An alternative explanation is that persistence of duplicates is triggered by genetic modification of one or both paralogs after duplication. For example, duplication could permit each copy of a multifunctional protein to specialize on a subset of the ancestral activities, thereby reducing pleiotropy (Hughes 1994; Hughes 1999). Duplicates might also be preserved if each paralog degrades in a complementary fashion (Force et al. 1999; Stoltzfus 1999) or if one or both paralogs acquire novel function (Goodman et al. 1987; Ohno 1970). The post-duplication neofunctionalization model, for example, posits that one gene copy carries out the ancestral function(s), while the other one evolves neutrally and then acquires beneficial mutations by chance during the early stages of evolution (Ohno 1970). Once a new function is achieved, purifying selection is expected to dominate later stages of evolution. Neofunctionalization could occur with complete loss, partial degradation, or retention of ancestral function (He and Zhang 2005b). The duplication-degeneration-complementation model, also known as subfunctionalization, posits that after duplication each paralog degenerates in a complementary fashion such that the action of both is necessary to accomplish the full suite of ancestral activities (Force et al. 1999; Lynch and Force 2000). Subfunctionalization could occur at the expression level through degeneration of paralogous expression profiles in a spatial, temporal, or quantitative dimension (Force et al. 1999; Lynch and Force 2000; Postlethwait et al. 2004). It could also occur at the protein level through complementary degeneration of different functional domains (Force et al. 1999) or as a consequence of activity compromising substitutions (Stoltzfus 1999). The cellular location of expression also has an impact on protein function, and subcellular relocalization could facilitate or catalyze the evolution of unique functions in paralogs (Byun-McKay and Geeta 2007).

If genetic modification triggers the persistence of both paralogs, it must occur within a few million years after duplication or else one copy will likely become a pseudogene (Lynch and Conery 2000). Moreover, the tempo of genetic modification after duplication may be dynamic, wherein changes that occur when the duplicate is young differ in frequency or nature from those that occur later on. After subfunctionalization or post-duplication neofunctionalization has occurred, for example, purifying selection is expected to increase. Additionally, some of these mechanisms for duplicate gene retention are not mutually exclusive and could operate concurrently or sequentially (Chain and Evans 2006; He and Zhang 2005b) and this could also be associated with temporal changes in functional constraints. To better understand the genetic basis of duplicate gene survival, it is therefore useful to consider their early stages of evolution separately from their later stages (Lynch and Conery 2000; Moore and Purugganan 2003; Su et al. 2006; Wendel 2000). Comparison of young to old duplicates suggests that the rate of nonsynonymous

substitutions is higher on average in younger duplicates (Jordan et al. 2004; Lynch and Conery 2000; Nembaware et al. 2002). This observation was interpreted as evidence of relaxed purifying selection immediately after duplication that was then followed by increased selective constraints as the duplicates aged. However, because pseudogenization rapidly transforms most young duplicates to singletons, it is not yet clear the degree to which evolution of young duplicates is indicative of the early stages of evolution of those exceptional duplicates that evade pseudogenization for dozens of millions of years.

To understand why so many duplicates persist after WGD, such as those that occurred in the ancestor of jawed vertebrates (Dehal and Boore 2005), teleost fish (Amores et al. 1998), and salmonid fish (Allendorf and Thorgaard 1984), additional information is needed about temporal dynamics in protein evolution and expression in the earliest stages of this type of genomic metamorphosis. In particular, we would like to dissect apart the molecular changes in the protein-coding region that occurred when persistent duplicates were young (an early stage of duplicate gene evolution) from those changes that occurred in the *same duplicates* after they became old (a later stage of duplicate gene evolution). Also of interest is the question of whether and how quickly paralogous expression profiles diverge after WGD. Polyploid clawed frogs (*Xenopus* and *Silurana*) are a useful model for studying early genetic events in vertebrate WGD because two independent instances of tetraploidization occurred fairly recently (Chain and Evans 2006; Evans et al. 2004) and because subsequent speciation events occurred after both of these WGDs (Figure 2.1A).

Previous studies have used this system to compare molecular evolution before and after WGD (Chain and Evans 2006; Hellsten et al. 2007; Hughes and Hughes 1993; Morin et al. 2006). These studies indicate that purifying selection on *X. laevis* paralogs is relaxed compared to single-copy genes in the diploid species *S. tropicalis* (Chain and Evans 2006; Hellsten et al. 2007; Morin et al. 2006), compared to single-copy orthologs in mammals (Hellsten et al. 2007; Hughes and Hughes 1993), and compared to single-copy genes in *X. laevis* (Hellsten et al. 2007). Using different statistical methods, independent tests on different genes found evidence for asymmetric amino acid substitution in 4-6% of expressed paralogs in *X. laevis* (Chain and Evans 2006; Hellsten et al. 2007). We have used this system to explore duplicate gene evolution over different time intervals after WGD (tetraploidization), and to evaluate expression divergence of the resulting paralogs in *X. laevis*.

RESULTS

In *Xenopus* and in *Silurana*, because a tetraploid ancestor speciated, the timing of molecular changes that occurred after allopolyploidization can be dissected apart into two stages: an "early" stage of duplication – after allopolyploidization but before speciation of the tetraploid ancestor – and a "later" stage of duplicate gene evolution – after allopolyploidization and speciation of the tetraploid ancestor (Figure 1). This permits the testing of alternative evolutionary

scenarios of duplicate gene evolution. Moreover, the likelihood of sequence data can be quantified under a model with no change in the rate ratio of nonsynonymous to synonymous substitution (Ka/Ks ratio) before versus after tetraploid speciation, and it can be compared to the likelihood of an alternative model in which there is a different Ka/Ks ratio during these two stages of duplicate gene evolution. This analysis is not the same as a comparison of young to old duplicates, which involves comparing different genes that were duplicated at different times – instead it allows comparison of an early stage of evolution to a later stage of evolution of the *same* duplicates.

Synonymous divergence

We collected and analyzed sequence data from fragments of hundreds of expressed paralogs from multiple species with an aim of teasing apart early from later mutations in the protein coding region of persistent paralogs generated by WGD (Figure 2.1). In *Xenopus*, an analysis of 80,856 concatenated base pairs (bp) of expressed paralogs indicates that synonymous substitutions per synonymous site (Ks) between *X. laevis* paralogs (XLα and XLβ in Figure 2.1B) is 0.2111, and Ks between the alpha paralogs of *X. laevis* and *X. borealis* (XLα and XBα in Figure 2.1B) is 0.1393. This suggests that Ks between paralogs in the "early" stage of duplicate gene evolution is up to 0.0718, depending on the location of node 2 in Figure 2.1B. Most synonymous divergence between paralogs therefore accumulated after tetraploidization in *Xenopus* (Figure 2.5), which occurred roughly 20 to 40 million years ago (Chain and Evans 2006) or maybe more (Evans et al. 2004). *Silurana* allotetraploids are about half as old (Evans et al. 2004).

Rapid and persistent purifying selection after duplicate gene evolution

After allopolyploidization, these paralogs were rapidly (immediately or soon after WGD) subjected to strong purifying selection. The level of purifying selection, while relaxed relative to singletons (Chain and Evans 2006; Morin et al. 2006), did not vary substantially between early and later stages of duplicate gene evolution.

More specifically, a combined analysis of thousands of codons from hundreds of expressed paralogs from *X. laevis*, a *X. borealis* ortholog, and a *S. tropicalis* ortholog, indicates that a more parameterized model of sequence evolution with a higher Ka/Ks ratio during the early stage of duplicate gene evolution than the later stage is not preferred ($P = 1.00$, Table 2.1, Figure 2.1B). In fact, a branch-specific model of evolution indicates that the estimated Ka/Ks ratio in the early stage of duplicate gene evolution is slightly lower than in the later stage (Table 2.1). When these data were partitioned by gene fragment the results were the same – there also was not a significant difference in the Ka/Ks ratio at the early compared to the later stage of duplicate gene evolution (Table 2.1). Additionally, a model in which the Ka/Ks ratio of the early lineage is allowed to be lower than one is significantly better than a model in which this rate ratio is fixed at the neutral expectation of one ($P < 0.00001$, Table 2.2) and this analysis also produced the same result when the data were partitioned by gene fragment (Table 2.2). Similarly, branch-site models reveal

a higher proportion of positively selected sites in the later lineage (0.00893%) than the early lineage (0.00061%; data not shown).

Tests of the individual loci have low power because many are small fragments (see Table 2.3). Nonetheless, analyses of 660 fragments from 350 individual loci echo the results of the analyses of combined multi-locus data. The distribution of Ka/Ks ratios in the early and later stages of duplicate gene evolution is similar (Figure 2.2A) and more fragments have a significantly higher Ka/Ks ratio at a later stage (8 fragments) than at an earlier stage (6 fragments), and this difference is not significant (χ^2 test, P=0.997). Additionally, the number of fragments with a higher Ka/Ks ratio in the early stage than the later stage (significant or not) was lower (156 fragments) than the alternative (262 fragments; P = 1.0; see Table 2.3). That Ks in the early stage of duplicate gene evolution was similar to or lower than in the later stage (Figure 2.1, see Supplementary Information), indicates that sampling bias of synonymous substitutions (Chain and Evans 2006; Wyckoff et al. 2005), if present, would bias our analysis of individual fragments towards detecting a higher Ka/Ks ratio in the early stage, which is not what we observed.

The neutral expectation (Ka/Ks equal to one) is significantly rejected in the early lineage of 62 out of 136 individual loci with more than 200 bp (see Table 2.3), and when this ratio is estimated for the early lineage, only 7% of them have an estimated Ka/Ks ratio above one. Taken together, these results indicate that purifying selection was as strong, if not stronger on these duplicates in the early stage of their evolution compared to the later stage.

Early neofunctionalization could potentially result in no difference between the Ka/Ks ratio in the early and later stages of duplicate gene evolution if genes in the early stage experience either positive selection or purifying selection, whereas genes in the later stage experience either relaxed purifying selection or purifying selection. While we can not rule this possibility out because positive selection and relaxed purifying selection both increase the Ka/Ks ratio, a regression of Ka/Ks to Ks for each fragment in the early and later stage of evolution indicates that (positive + relaxed purifying) selection is less prevalent in the early stage than the later stage (Figure 2.2B).

Radical amino acid substitutions are not more common in early versus later stages of duplicate gene evolution

New functions may be achieved by "radical" substitutions – replacement of one amino acid with another that has very different physical properties (Hughes 1999; Hughes et al. 1990). While this is certainly not a requirement for new function to evolve, we nonetheless explored this possibility using a Bayesian approach to estimate the number and frequencies of elemental substitutions – the 75 amino acid substitutions that can occur via a single nucleotide change – at an early and a later stage of duplicate gene evolution, and also in the diploid lineage (see Supplementary Information). Results indicate that elemental substitutions were not more radical in an early stage (Mantel Z statistic = 2.4119) than in a later stage (P =

1.0000). In fact, radical substitutions were slightly more prevalent in the later lineage (Mantel Z statistic = 2.4680). Elemental substitutions also were not significantly more radical in the entire *X. laevis* paralog α lineage (between node 1 and XLα of Figure 2.1B, Mantel Z statistic = 2.43823) than in the diploid lineage (between ST and node 1 in Figure 2.1B; Mantel Z statistic = 2.3920, P = 0.1396). Similar results are obtained when the radicalness of elemental substitutions is categorized according to alternative criteria (data not shown; Zhang 2000).

Simulations were performed to test whether ancestral bias toward more conservative substitutions in the early stages of duplicate gene evolution could explain these results, but this was not the case. Simulated elemental substitutions from a reconstructed ancestral sequence were not more conservative in the early stage of duplicate gene evolution than the later stage (P = 0.6529). As expected, these simulations, which were not under purifying selection, were significantly more radical than the observed data (P < 0.001).

Caveats

We performed additional analyses to address various concerns about the sequence dataset from *X. laevis*, *X. borealis*, and *S. tropicalis*. One consideration is that differences or changes in population size could affect the Ka/Ks ratio because slightly deleterious nonsynonymous substitutions are more likely to become fixed when the effective population size is small. Based on the geographic distribution and molecular diversity of mitochondrial DNA (Evans et al. 2004), the effective population size of *X. borealis* is smaller than that of *X. laevis*. However, we found that the Ka/Ks ratio of *X. laevis* paralogs during the later stage was slightly higher (0.1555) than the corresponding lineage of *X. borealis* (0.1338). This discrepancy was not significant in a two–sided test (P = 0.1790) or in a one-sided test because we expected the ratio to be larger in *X. borealis* (P = 1.0). To ensure that we were comparing ratios in expressed duplicates in both species, we included in this comparison only those data for which expression of both paralogs of both species was confirmed (37,194 bp). We note that more substitutions of both types occurred in *X. borealis* suggesting that the overall rate of evolution may be slightly higher in this species. A lack of significant difference in the Ka/Ks ratio suggests that the difference in effective population size between *X. laevis* and *X. borealis* had a negligible impact on the Ka/Ks ratios of many of their orthologs.

A second consideration stems from the possibility that a substantial portion of the early lineage of duplicate gene evolution evolved in a diploid (between nodes 1 and 2 in Figure 2.1B) as a result of the putative allopolyploid origin of the ancestor of *Xenopus* tetraploids. Because the Ka/Ks ratio of clawed frog paralogs is slightly higher after genome duplication than before it (Chain and Evans 2006; Morin et al. 2006), the Ka/Ks ratio of this entire branch (between nodes 1 and 3 in Figure 2.1B) could be lower than the Ka/Ks ratio of the portion of this branch that evolved after duplication (between nodes 2 and 3 in Figure 2.1B). To explore this issue, we analyzed expressed sequences from another dataset derived from *S. tropicalis* and two closely related tetraploids (9717 bp). Similar to the analysis of *X. laevis* and *X.*

borealis paralogs, the branch-specific tests of *Silurana* paralogs do not provide evidence for an increased Ka/Ks ratio in an early stage (between nodes 7 and 8 in Figure 2.1C) versus a later stage of duplicate gene evolution (between node 8 and EPα in Figure 2.1C; P = 1.0; Table 2.1), nor an increased frequency of radical amino acid substitutions at an early stage of duplicate gene evolution (Mantel Z statistic = 2.7193) compared to a later stage (Mantel Z statistic = 2.1991, P = 0.0882). Simulations indicate that the early stage of duplicate gene evolution in *Silurana* was not significantly biased towards more conservative substitutions (P = 0.5651), the branch-site test reveals no evidence in the concatenated data for positively selected sites in the early branch (although it does on the later branch; data not shown), and the partitioned branch model analysis reveals the same results as the concatenated branch model (Tables 2.1 and 2.2). Also similar to the analysis of *X. laevis* and *X. borealis* paralogs, the branch-specific tests of *Silurana* paralogs illustrate that functional constraints during the early stage of duplicate gene evolution were significantly below neutral expectations (Table 2.2).

A third consideration is that allotetraploidization of the common ancestor of *Xenopus* tetraploids occurred immediately before the first speciation of this ancestor (in other words that the time between nodes 2 and 3 in Figure 2.1B is very small). If this were the case, then it would be more informative to compare an "intermediate" stage of duplicate gene evolution – a period after the first tetraploid speciation in *Xenopus* but before subsequent tetraploid speciations (i.e. between nodes 3 and 4 in Figure 2.1D) – to a later stage of duplicate gene evolution – after an even more recent tetraploid speciation event (between node 4 and XLα in Figure 2.1D). This issue was addressed with additional sequences (6966 bp) from the tetraploid species *X. gilli* and *X. muelleri* that made possible the further dissection and hypothesis testing of the temporal dynamics of evolution after duplication (Figure 2.1D). Based on their close phylogenetic relationships (Evans 2007; Evans et al. 2005; Evans et al. 2004), we used *X. gilli* when we knew both *X. laevis* paralogs were expressed, and we used *X. muelleri* when we knew both *X. borealis* paralogs were expressed. Similar to the other analyses, this comparison revealed that the Ka/Ks ratio is not significantly higher in the intermediate stage compared to the later stage of duplicate gene evolution (P = 0.16; Table 2.1) and that the frequency of radical amino acid substitutions at the intermediate stage of duplicate gene evolution (Mantel Z statistic = 2.4073) is not significantly higher than at a later stage (Mantel Z statistic = 2.0645, P = 0.0887). Simulations again confirm that the intermediate stage was not significantly biased towards more conservative substitutions (P = 1.0000), the branch-site test reveals no evidence in the concatenated data for positively selected sites in the early branch or the later branch (data not shown), and the partitioned branch model analysis again reveals the same results as the concatenated branch model (Tables 2.1 and 2.2). These additional analyses thus provide strong support that purifying selection acted rapidly – within millions of years – and persistently – over tens of millions of years – after WGD in clawed frogs.

Expression divergence

We used microarray data to compare expression profiles from five developmental stages and adult tissue types (treatments) of hundreds of paralogous pairs generated by WGD. Our analyses included developmental treatments from four distinct developmental stages (egg, tadpole stage 11, tadpole stage 18, and adult). Unlike the egg and tadpole stages, however, the adult stage is represented by data from each type of gonad instead of the entire individual. Because the primordial germ cells appear long after the tadpole stages that we assayed (stage 44; Nieuwkoop and Faber 1956), these data provide a coarse perspective on spatial expression in four distinct tissue types: undifferentiated egg, pooled embryonic tissue (which do not have developed gonads), adult testis, and adult ovary.

We performed a power analysis to explore the possibility that cross-hybridization of non-target paralogs could affect the inference of paralogous expression profiles. We compared results from (a) a low paralog specificity analysis that included all probes on the microarray, including ones that cross hybridize to both paralogs, (b) a medium paralog specificity analysis that excluded probes whose sequences cross-hybridized to both paralogs, and (c) a high specificity analysis that excluded probes having up to three mismatches with a nontarget paralog. Additionally, we used two intensity thresholds, "standard" and "conservative", as a basis for the detection of expression of each paralog in each treatment (Methods).

Qualitative comparisons across this developmental series and these tissue types indicate that the bulk of paralogous expression divergence after WGD in clawed frogs is on a quantitative rather than a temporal dimension (Figures 2.3, 2.4). This would be expected if these paralogs were expressed in a highly specific manner in only one of the developmental stages or tissue types that we analyzed. However, many of these paralogs were expressed in multiple tissue types and multiple developmental stages. Consider for instance the 841 paralogous pairs for which the presence/absence expression profile of each paralog was identical in the medium and high paralog specificity analyses (Figure 2.3). In the medium specificity analysis at the standard threshold, 94% of these paralogous pairs were both expressed in at least two treatments and 75% were both expressed in all five treatments.

When both paralogs are expressed, comparison of their expression profiles can indicate either that (a) both are expressed at the same developmental stages and tissue types (identical spatial and temporal expression), (b) the profile of one paralog is a subset of that of the other one (overlapping spatial and temporal expression), or (c) both paralogs have distinct components to their expression profiles (distinct spatial and temporal expression). In the microarray expression data from *X. laevis*, when expression of both paralogs was detected, almost all pairs had identical or overlapping expression profiles in terms of the developmental stages and tissue types in which expression was detected (Figure 2.3). This was true regardless of how conservatively we scored presence/absence of expression or the specificity of the probes on the microarray. Only 2–7% of these pairs included paralogs that both had a unique expression profile wherein one paralog is expressed at a developmental

stage or a tissue type where the other one is not expressed, and vice versa (Figure 2.3).

In contrast to the overall similarity in the developmental timing and locations of paralogous expression, quantitative aspects of a high percentage of paralogous pairs have diverged substantially (Figure 2.4). In the medium paralog specificity analysis for example, 62% of the paralogous pairs also had a Pearson correlation coefficient that was below 0.866, a value below which were 95% of the correlation coefficients between non-paralogous genes. 27% of the paralogous pairs had a correlation coefficient below 0.5. Similar proportions were found in the high paralog specificity analysis (results not shown). At the end of this extreme, 0.3% of the paralogous pairs (3 pairs) in the medium paralog specificity analysis had a correlation coefficient that was more negative than -0.861, a level below which were only 5% of the correlation coefficients of the non-paralogous expression profiles. These three paralogous pairs are expressed in all treatments according to the standard detection threshold and have the following accession numbers (NM_001092603 and NM_001091285, NM_001091759 and NM_001093475, and NM_001091931 and NM_001094047). Their annotations are rudimentary, but the first pair may be involved with RNA splicing and the third pair has sequence similarity to collagen alpha (1) precursor. The normalized expression level of each pair indicates that in most of these treatments, the expression of one paralog is above the median expression level of that paralog across the five treatments whereas the expression of the other paralog is below it.

DISCUSSION

Neutral evolution of gene duplicates eventually leads to pseudogenization of one copy, and the time for this to occur depends on the size of the mutational target (sequence and length of the gene and the level of degeneracy of *cis*-regulatory elements), the rate and biases of molecular evolution (such as the rates of nucleotide substitution, insertions/deletions, and transposable element mobility), and the effective population size of the species (pseudogenes take longer to fix in larger populations) (Kimura 1980; Takahata and Maruyama 1979; Watterson 1983). Non-neutral evolution, however, can curtail pseudogenization. In polyploid clawed frogs, duplicates generated by WGD are subject to more severe functional constraints than the neutral expectation, even though these constraints are relaxed relative to a singleton gene (this study; Chain and Evans 2006; Hellsten et al. 2007; Hughes and Hughes 1993; Morin et al. 2006). Furthermore, even though the typical half-life of duplicates from a variety of organisms (Lynch and Conery 2000) is much lower than the time since tetraploidization of clawed frogs, it is clear that many paralogous pairs are still expressed in *Xenopus* (Chain and Evans 2006; Hellsten et al. 2007; Hughes and Hughes 1993; Morin et al. 2006), suggesting the action of natural selection to preserve their expression. If these paralogs are retained for enough time, functional constraints presumably would increase to a pre-WGD level. However,

here we demonstrate that these constraints did not substantially fluctuate for dozens of millions of years following genome duplication.

One explanation for this observation is that the early stages of duplicate gene evolution occurred before these genomes became disomic (diploidized), and that this resulted in increased purifying selection on both duplicates in the early stages of their evolution. Indeed, some chromosomes may take longer than others to evolve disomic inheritance after WGD (Allendorf and Danzmann 1997; Ferris and Whitt 1979; Gaut and Doebley 1997) and polysomic inheritance has been reported at one locus in the dodecaploid species *X. ruwenzoriensis* (Sammut et al. 2002). However, we removed from our analysis sequences that exhibited signs of gene conversion or recombination (see Supplementary Information) – events that might indicate polysomic rather than disomic inheritance. Additionally, disomic inheritance can occur instantly or soon after WGD by allopolyploidization (Osborn et al. 2003) and disomic inheritance of alleles occurs immediately in laboratory generated polyploids of *Xenopus* (Müller 1977). These observations argue against functional constraints on these paralogs being buoyed by polysomic inheritance in an early stage after allotetraploidization.

The stasis of functional constraints over these early stages of paralog evolution in clawed frogs contrasts sharply with studies of young and older duplicates generated from WGD in non-vertebrates and from segmental duplication in vertebrates. For example, over a level of synonymous divergence similar to *Xenopus* paralogs, older paralogs of the fungus and plant polyploids *Saccharomyces cerevisae* and *Arabidopsis thaliana* are more constrained than younger ones (Lynch and Conery 2000). Likewise, human paralogs with synonymous divergence between 0.05 and 0.1 have a Ka/Ks ratio of about 0.47 but those with synonymous divergence between 0.1 and 0.5 are more constrained with a Ka/Ks ratio of about 0.37 (Jordan et al. 2004). Although those comparisons involve different sets of genes in each taxon, it is worth noting that functional constraints immediately after WGD are more severe in *Xenopus* paralogs, which have a lower Ka/Ks ratio of 0.105 – 0.158 (Table 2.1). These results suggest that (a) the evolutionary trajectories of duplicates generated by segmental duplication differ from those of paralogs generated by WGD and/or that (b) the early stage of evolution of duplicates that are destined to persist differs substantially from that of most young duplicates (the bulk of which rapidly degenerate to singletons). These results are consistent with the observation that young paralogs that evolve quickly are less likely to be retained in the long run (Brunet et al. 2006; Davis and Petrov 2004; Jordan et al. 2004; Shakhnovich and Koonin 2006). Stoichiometric constraints/genic balance is one plausible explanation for more severe and persistent functional constraints on WGD paralogs in clawed frogs as compared to singletons in other organisms (Freeling and Thomas 2006; Lynch and Conery 2000; Papp et al. 2003a).

Temporal dynamics of molecular evolution of expressed duplicates appear to differ in frogs (this study) and yeast (Scannell and Wolfe 2008). While purifying selection is relaxed after WGD in yeast and in *X. laevis*, nonsynonymous substitutions were more prevalent during an early stage of duplicate gene evolution

than a later stage in yeast (Scannell and Wolfe 2008) but not in *X. laevis* (this study). There are multiple possible explanations for this difference. Because the yeast species examined in (Scannell and Wolfe 2008) have a larger effective population size than the frogs we studied, purifying selection in frogs would have to be stronger in order to substantially curtail the fixation of slightly deleterious nonsynonymous substitutions by genetic drift. Perhaps then, the initial phase of duplicate gene evolution – a period during which purifying selection is relaxed compared to singletons but before post-WGD increases in functional constraints are apparent at a molecular level – is more drawn out in frogs than in yeast as a consequence of their different population sizes. Another possibility is that the selective regime following WGD varies between yeast and frogs as a result of fundamental differences in the nexus of protein-protein interactions, functional specialization, complexity, and/or redundancy. It is also possible that the periods of time after WGD that were compared in each of these studies could differ substantially.

If post-duplication neofunctionalization of protein structure is to promote the persistence of both paralogs, amino acid changing nucleotide substitutions must occur in at least one paralog soon after duplication, and this should be followed by increased purifying selection once new function is acquired (Ohno 1970; Ohno 1973). Molecular signs of neofunctionalization of protein structure may include a higher Ka/Ks ratio in early than in later stages of duplicate gene evolution, a higher frequency of radical amino acid changes in early than in later stages of duplicate gene evolution, and/or significantly different rates of nonsynonymous substitution between paralogs. In clawed frogs, multiple lines of evidence suggest that this mechanism is not a prevalent trigger for the persistence of duplicates generated in the initial millions of years after WGD. First, in the early stage of duplicate gene evolution in *X. laevis* only a handful of these persistent paralogs have a Ka/Ks ratio greater than one (i.e. consistent with positive selection; see Table 2.3) and a higher proportion of sites exhibit evidence of positive selection in the later stage of duplicate gene evolution than in the early stage (data not shown). Of course, the Ka/Ks ratio is a very rough metric of positive selection and new protein function could arise by neutral evolution, even via very few amino acid substitutions (Golding and Dean 1998). However, similar to yeast (Scannell and Wolfe 2008), radical amino acid substitutions are not more prevalent in the early stage of duplicate gene evolution. We also did not observe increased purifying selection in the later stage of duplicate gene evolution that would be expected if neofunctionalization occurred in the early stage after WGD. Similarly in yeast, duplicates with a level of divergence similar to *X. laevis* paralogs ($Ks < 0.25$), subfunctionalization as opposed to neofunctionalization is suggested by a loss of shared interactions (He and Zhang 2005b; Wagner 2001).

These analyses find a much higher incidence of quantitative divergence than the 14% suggested by (Morin et al. 2006), but they are similar to another study that suggests 40-50% quantitative expression divergence (Hellsten et al. 2007). Hellsten et al. (Hellsten et al. 2007) found evidence of spatial expression divergence in four out of six *in situ* hybridizations, whereas we found this type of expression

divergence – where each paralog has a unique component to its expression domain – in only 2–7% of the paralogs (Figure 2.3). This disparity is in part a consequence of lack of resolution in the microarray data that we analyzed relative to *in situ* hybridization performed by (Hellsten et al. 2007). Spatial and temporal expression subfunctionalization may be more common on a finer spatial or temporal scale than we were able to detect with these microarray data.

Unequal expression and low correlation of paralogous expression profiles has also been reported in several allopolyploid plants (Adams 2007; Blanc and Wolfe 2004). Genome duplication in plants is associated with non-additive changes in gene expression, suggesting that expression divergence between paralogs can immediately accompany allopolyploidization (Adams et al. 2003; Albertin et al. 2006; Wang et al. 2006). In synthetic allopolyploid *Arabidopsis*, for example, expression of over 5% of genes in synthetic allopolyploid lines deviated from the midpoint of each parental species (Wang et al. 2006). In the recently formed allohexaploid plant species *Senecio cambrensis*, expression analysis of re-synthesized lines suggests that the impact of hybridization and genome duplication on expression divergence are distinct, and that the latter phenomenon can reduce expression divergence, at least in the early stages of polyploid evolution (Hegarty et al. 2006). Later on, for example in *Arabidopsis thaliana* which experienced WGD between 20 and 60 million years ago, 57% of the resulting duplicates have an expression profile with a correlation coefficient less than 0.52 (Blanc and Wolfe 2004). Likewise in yeast the correlation between paralogous expression profiles is lower than 0.5 in 55% of pairs that have a similar level of synonymous divergence (0.1-0.3) as the *X. laevis* paralogs in this study (Gu et al. 2002). Substantial quantitative expression divergence between paralogs soon after WGD therefore does not appear to be unique to *X. laevis*, and is likely the culmination of divergence over evolutionary time and also divergence that occurred immediately upon allopolyploidization.

Without additional information on expression profiles of orthologous genes, at this point we cannot determine whether the observed spatial and temporal expression divergence arose through expansion (expression neofunctionalization) or degradation (expression subfunctionalization) of each expression profile or both. In yeast, expression neofunctionalization occurs via recruitment of *cis*-regulatory elements, but this appears to take a long time (Papp et al. 2003b). In human paralogs that are more diverged ($Ks > 0.25$) than the ones we studied here, the combined expression domains of segmental duplicates is typically larger than that of singletons, and the magnitude of this difference is positively correlated with synonymous divergence, suggesting expression neofunctionalization (He and Zhang 2005b). Expression divergence is correlated with synonymous and nonsynonymous divergence in yeast duplicates with $Ka \leq 0.3$ or $Ks < 1.5$ (Gu et al. 2002), and this correlation has also been found in humans over similar levels of divergence (Makova and Li 2003). However, we did not find this correlation in *X. laevis* paralogs (see Additional file 4). This difference could derive from distinct genetic fates of duplicates generated by WGD versus segmental duplication on either an

expression or functional level (Kim et al. 2006). Other factors that could play a role in the degree to which paralogous expression profiles diverge over time include tissue-specific developmental constraints (Gu and Su 2007), expression intensity and specificity (Liao and Zhang 2006), and the essentiality of a paralog's gene family (Shakhnovich and Koonin 2006).

CONCLUSIONS

It has been suggested that allopolyploidization rather than autopolyploidization preceded the diversification of jawed vertebrates (Spring 1997). Allopolyploids have the advantage that diploidization might occur instantly or more rapidly than in autopolyploids, thereby preventing complications associated with mis-segregation of chromosomes in a polysomic genome (Osborn et al. 2003). By analogy, duplicate gene evolution in allopolyploid clawed frogs offers insights into how the transcriptome of our ancient ancestors may have been sculpted in the wake of these genomic metamorphoses (Dehal and Boore 2005; Ohno 1970), and also after subsequent WGDs in other vertebrates (Amores et al. 1998; Taylor et al. 2003). To the extent that this analogy applies, the initial dozens of millions of years of vertebrate evolution after WGD were likely characterized by strong and persistent functional constraints at the amino acid level. Despite these functional constraints, however, quantitative expression divergence probably occurred in many duplicates during this period and, as has been suggested (Ferris and Whitt 1979), the magnitude of regulatory and structural change was not correlated (see Additional file 4). We speculate therefore that stoichiometric requirements and quantitative expression subfunctionalization commonly trigger persistence of WGD paralogs in the earliest stages of their existence. Following WGD, it appears that other mechanisms that trigger the retention of duplicate genes, such as neofunctionalization of the coding region or spatial expression subfunctionalization (e. g. Amores et al. 2004; Force et al. 1999; He and Zhang 2005b; Huminiecki and Wolfe 2004; Rastogi and Liberles 2005), tend to operate less frequently, later, or over a longer period of time. Interestingly, analysis of teleost paralogs demonstrates that duplicates continue to be lost over hundreds of millions of years (Amores et al. 2004), indicating that the steadfast functional constraints and substantial expression dynamics soon after vertebrate WGD do not immortalize these duplicates.

METHODS

Molecular data

We compiled sequences of expressed paralogs of *X. laevis* from Genbank and various publications (Chain and Evans 2006; Hellsten et al. 2007; Morin et al. 2006) and aligned them with orthologs from the *S. tropicalis* genome assembly 4.1. 454 pyrosequencing was used to obtain sequences of fragments of expressed paralogs of *X. borealis* from testis cDNA and contigs were assembled from these data using BLAST (Altschul et al. 1997) and ALIGN0 (Myers and Miller 1988)

from the FASTA 2.0 package (Pearson and Lipman 1988), and manual alignment in MacClade (Maddison and Maddison 2000). Manufacturer protocols were followed to isolate RNA using an RNA extraction kit (Qiagen), to prepare cDNA using BD SMART PCR cDNA synthesis kit (Clontech), and to normalize the cDNA using the Trimmer cDNA normalization kit (Evrogen JCS). Additional targeted sequencing of paralogs from *X. laevis*, *X. borealis*, *X. gilli*, *X. muelleri*, *S. epitropicalis*, and *S.* new tetraploid was performed by co-amplifying portions of these paralogs from cDNA from a variety of tissues (blood, heart, brain, testis, liver, muscle). Portions of individual paralogs were then cloned with the TA cloning kit (Invitrogen) and sequenced. These data are deposited in Genbank (see Table 2.3).

Using a combination of targeted amplification, cloning, and sequencing of cDNA, 454 pyrosequencing of cDNA, and database searches, 80,856 bp were collected from 660 fragments of 350 expressed paralogous pairs from the tetraploid *X. laevis*, one expressed paralog from the tetraploid *X. borealis*, and an ortholog from the diploid *S. tropicalis*. An additional 9,717 bp were sequenced from portions of thirteen expressed duplicated loci of the tetraploids *S. epitropicalis* and *S.* new tetraploid, and 6,966 bp were sequenced from portions of nine expressed duplicated loci of the tetraploids *X. muelleri* or *X. gilli*. To minimize analysis of paralogs whose evolutionary history may have been homogenized by gene conversion or recombination, we excluded from our analysis sequences with signs of these phenomena (see Figure 5).

Because data were usually obtained from only one expressed *X. borealis* paralog but two *X. laevis* paralogs, most of our molecular analyses focused on molecular evolution of one *X. laevis* paralog – the "α" paralog (Figure 2.1B). This is because, without evidence of expression of the other *X. borealis* paralog, we do not know whether *X. borealis* paralog α is still an expressed duplicate, and we also cannot determine at what point after duplication nonsyonymous substitutions occurred in the other *X. laevis* paralog – paralog "β" (Figure 2.1B). Phylogenetic methods (maximum parsimony, maximum likelihood) were used to identify to which one of the expressed *X. laevis* paralogs that the *X. borealis* paralog was most closely related.

Models of evolution

To test whether the rate ratio of nonsynonymous to synonymous substitutions per site (hereafter the Ka/Ks ratio) differs at early versus later stages of duplicate gene evolution, the likelihood of alternative models of branch-specific evolution (Figure 2.1) was calculated using PAML version 3.15 (Yang 1997). This analysis was performed on concatenated datasets and with the data partitioned by gene fragment. We also used the branch-site test for positive selection (test 2 in Zhang et al. 2005) to test whether there were more sites under positive selection at an early stage compared to a later stage of duplicate gene evolution. In addition, we tested for significant departure from neutrality by comparing a model in which the Ka/Ks ratio of the early lineage was fixed at one and another Ka/Ks ratio was estimated for all other branches, to a model in which one Ka/Ks ratio was estimated

for the early branch and another ratio was again estimated for all other branches. For each comparison, significance of the more parameterized model was evaluated with a χ^2 test. Note that, as a result of a suspected allotetraploid origin of the ancestor of *X. laevis* and *X. borealis*, an unknown portion of the early lineage probably evolved in a diploid species; the potential impact of this and other caveats was explored with additional comparative data and analyses (Figs. 1C, D).

Expression analyses

We collected expression data from previous studies that used a *X. laevis* microarray prefabricated by Affymetrix (Gurvich et al. 2005; Malone et al. 2006; Sinner et al. 2006). Expression data was analyzed from five developmental stages or tissue types: egg, embryonic stages 11 and 18, adult testis, and adult ovary. Raw intensity data were converted to CEL files using GeneChip Operation System software (GCOS v. 1.4 Affymetrix). The robust multi array average (RMA) algorithm was implemented to quantify gene expression in GeneSpring version GX7.3 (Agilent, Inc) using either the Affymetrix library file or custom CDF files ("probe masks") that were generated following Hammond et al. (2005). The data were then normalized to the median of each gene across all arrays and the 50th percentile of each array. A high intra-treatment correlation (R^2 = 92-98%) was found between the biological replicates for each treatment.

The Affymetrix *X. laevis* microarray consists of "probe sets" that are composed of 16 "probe pairs", each of which includes a 25 base pair oligo that is intended to perfectly match the target sequence. Cross hybridization of paralogs could homogenize their expression profiles if it is bidirectional or could amplify differences between them if it is unidirectional. To explore this possibility, we performed a power analysis in which we used probe masks to evaluate paralog specificity of each probe set – i.e. the degree to which the probes on the microarray match one paralog but not the other. We tested three paralog specificities: "low", "medium", and "high". The low paralog specificity analysis included probes that exactly matched (and cross-hybridize to) both paralogs. The medium paralog specificity analysis excluded probes that exactly matched both paralogs. The high paralog specificity analysis excluded probes that perfectly matched both paralogs and also those that had up to and including three mismatches with the non-target paralog. We required each probe set in our analysis to have a minimum of at least 8 probe pairs (and up to 16) at the highest specificity. These probe masks were developed based on comparisons of the probe sequences to a sequences of expressed paralog pairs from previous publications (Chain and Evans 2006; Hellsten et al. 2007; Morin et al. 2006) that were carried out using BLAST searches (Altschul et al. 1997). We evaluated each of these probe specificities under two thresholds for calling presence/absence of expression ("standard" and "conservative" thresholds; Figure 2.3). For the "standard" threshold, a paralog was scored as expressed if its raw intensity was above a background level of 50. For the "conservative" threshold, a paralog was scored as expressed if its raw intensity was above a background level of 200. We note that these thresholds are somewhat arbitrary because some

probesets may hybridize with lower affinities than others, and therefore reveal lower than background raw intensities even though a transcript is in fact expressed. This approach therefore provides only a rough metric of whether or not a transcript is expressed. The Pearson correlation coefficients provide an alternative extreme, because they are based on the expression intensities in all treatments (even those that have below-background raw intensities). These correlation coefficients therefore must be interpreted with the caveat that higher correlations between tissue profiles could be obtained when neither transcript is expressed in many treatments. To contextualize the paralogous correlations, we also calculated the Pearson correlation coefficients between all non-parlaogous expression profiles as in (Blanc and Wolfe 2004).

ACKNOWLEDGEMENTS

We thank J. P. Bollback, P. S. G. Chain, W. G. Fairbrother, G. B. Golding, J. Hassell, J. P. Huelsenbeck, S. P. Otto, and J. R. Stone for helpful discussions or use of resources, R. Morin and Uffe Hellsten for providing information on expressed *X. laevis* paralogs, P. Michalak, B. Wittner, and A. Zorn for providing expression data, and B. Tracey and M. Zubairi for laboratory assistance. This research was supported by the Canadian Foundation for Innovation, the National Science and Engineering Research Council, the Ontario Research and Development Challenge Fund, and McMaster University.

SUPPLEMENTARY INFORMATION

Tests for recombination and gene conversion.

To explore the possibility that recombination or gene conversion occurred among these paralogs, multiple tests were used because their performance varies with the level of divergence, the extent of recombination, and among site rate heterogeneity (Posada 2002; Posada and Crandall 2001). Tests for recombination include the recombination detection program, geneconv, chimera, bootscan, and siscan, as implemented by the Recombination Detection Program (Gibbs et al. 2000; Martin and Rybicki 2000; Maynard Smith 1992; Padidam et al. 1999; Posada and Crandall 2001; Salminen et al. 1995). A variety of parameter settings were explored for each method as in (Evans et al. 2005), and only paralogs with more than 300 bp were analyzed using these tests.

A site is parsimony-informative if it contains at least two types of nucleotides that each occur in at least two taxa. Thus, when analyzing phylogenetic relationships among four taxa, the only character pattern that is parsimony-informative is one in which two taxa share one nucleotide and the other two share a different nucleotide. Using this principal, we tabulated the number and order of parsimony-informative "non-recombined" character patterns, in which the α paralogs of *X. laevis* and *X. borealis* both had the same nucleotide and the β paralog of *X. laevis* and the ortholog of *S. tropicalis* both had a different nucleotide.

Additionally, we tabulated the number and order of parsimony-informative "recombined" character patterns in which the α paralog of *X. laevis* and *X. borealis* each had a different nucleotide, but where each one was identical to the homologous nucleotide of the β paralog of *X. laevis* or the ortholog of *S. tropicalis*. Loci that had three or more consecutive "recombined" character patterns (which could derive from recombination or gene conversion between alleles of different paralogs), were excluded from our analysis. In one gene (Xmegs), a run of four recombined character patterns turned out to be a combination that included two nonrecombined character patterns when a fifth paralog was considered (*X. borealis* paralog β), so this locus was retained.

Conservative versus radical changes after duplication.

We used a Bayesian approach to estimate the number and frequency of each of the 75 elementary amino acid changes at different time points after genome duplication. This approach employed a simulation procedure to stochastically map mutations on a fixed topology (Nielsen 2002). We attempted to accommodate uncertainty in branch lengths and parameter values by sampling 100 sets from a post-burnin posterior distribution that was generated from Bayesian analysis with a constrained topology using MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001). This sample was used to simulate character evolution conditioning on the observed data and allowing all possible character states for each ancestral node with sampling of these states drawn according to their likelihood (Nielsen 2002). Simulations were performed using SIMMAP version 1.0 (Bollback 2006) and PERL scripts were used to reconstruct and tabulate each of the simulated elemental amino acid changes along each branch. Results were similar to those obtained from maximum likelihood analysis of amino acid substitutions.

A lineage with many radical amino acid substitutions has a low correlation between the frequency of each type of substitution and the magnitude of the biochemical differences between the ancestral and descendant amino acid residues. Mantel tests were used to calculate the correlation between the number of each type of elementary amino acid change and the associated biochemical transition associated with each substitution, based on eight physical properties (Urbina et al. 2006). To test whether this correlation was significantly different in the early stage of duplicate gene evolution than in a later stage, the Mantel Z statistic (Sokal and Rohlf 2003) from the early stage was compared to a distribution of Mantel Z statistics generated from 100,000 bootstrapped datasets derived from *n* draws from the multinomial frequency distribution estimated for the later stage, where *n* is a maximum likelihood estimate of the number of observed elemental substitutions in the early stage.

Simulations were performed to test whether phylogenetic inertia (an ancestral bias towards more or less conservative substitutions) could account for the observed proportion of radical and conservative substitutions at each stage of duplicate gene evolution. A maximum likelihood estimate of the ancestral sequence of nodes 1 and 3 in Figure 2.1A, nodes 2 and 3 in Figure 2.1B, and nodes (23) and 4

in Figure 2.1C, was obtained using PAML. For each branch, 100,000 simulations were performed from these ancestral sequences under the general time reversible model of evolution with a proportion of invariant sites and a gamma distributed rate heterogeneity parameter, using SeqGen version 1.3.2 (Rambaut and Grassly 1997). The posterior sample of 100 sets of parameter values and corresponding branchlengths that were used in the stochastic mapping of mutations in the observed data were also used in these simulations. Simulated elemental substitutions were then inferred by maximum likelihood and maximum parsimony. Additionally, the PSEUDOGENE program was used to obtain a rough estimate expected half lives of these loci under neutral evolution, using values for the rate of point mutations and the rate of insertions and deletions estimated from old world primates, as in Zhang and Webb (2003). A reconstruction of the ancestral sequence of these paralogs was used for the simulations and the half-life was estimated for only those loci for which complete transcripts were available in both *X. laevis* paralogs.

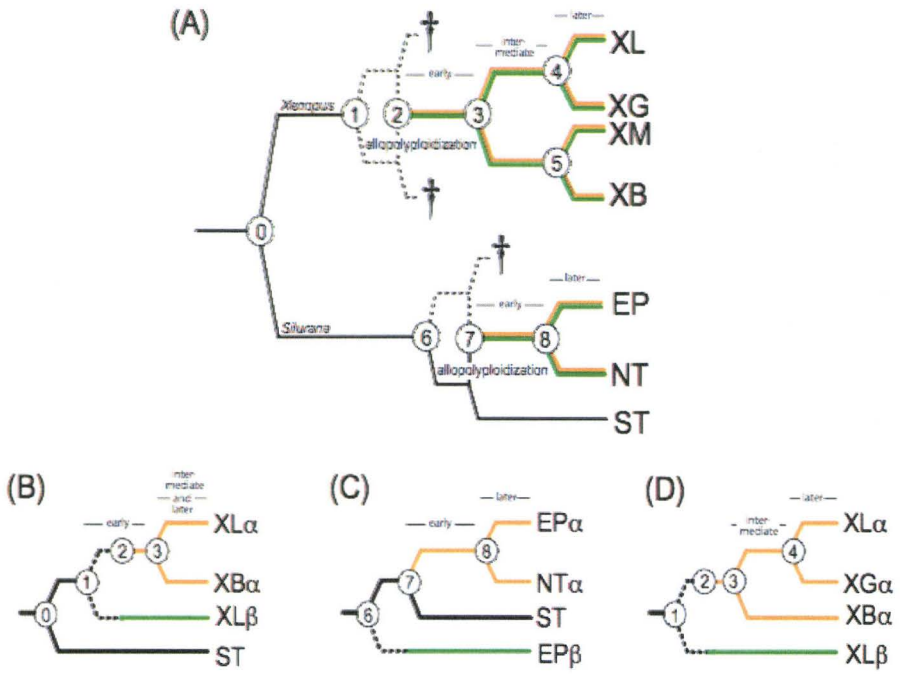


Figure 2.1 - Phylogenetic and genealogical relationships of species and paralogs in this study. Phylogenetic relationships are depicted among species, orthologs, and paralogs of a diploid with 20 chromosomes, *Silurana tropicalis* (ST), two tetraploids with 40 chromosomes, *S. epitropicalis* (EP) and *S. new tetraploid* (NT), and four tetraploids with 36 chromosomes, *Xenopus laevis* (XL), *X. borealis* (XB), *X. gilli* (XG), and *X. muelleri* (XM). (A) Speciation by allopolyploidization has occurred independently in *Xenopus* and in *Silurana* and produced two paralogs in the resulting tetraploid ancestor – α and β – that are indicated as brown and green lineages respectively. After allotetraploidization, some of the diploid lineages went extinct (daggers). As a result of these extinctions, the portion of some paralogous lineages that evolved in a diploid (dashed lines), cannot be dissected apart from the portion that evolved in an allopolyploid. Numbered nodes indicate (0) divergence of the genera *Xenopus* and *Silurana*, (1) divergence of the diploid (2n=18) ancestors of *Xenopus*, (2) allotetraploidization in *Xenopus*, (3) the first speciation event of the tetraploid ancestor of extant *Xenopus*, (4 and 5) more recent speciation events of *Xenopus* tetraploids, (6) divergence of the diploid (2n=20) ancestors of *Silurana*, (7) allotetraploidization in *Silurana*, (8) speciation of a tetraploid *Silurana* without change in genome size. Sequences from individual paralogs were used to construct genealogies to compare (B) an early to a later stage of evolution after WGD in XLα, (C) an early to a later stage of evolution after WGD of EPα and (D) an intermediate to a later stage after WGD in XLα. Depending on the paralog for which data were obtained, sometimes NTα was considered in (C) or XBα was considered in (D).

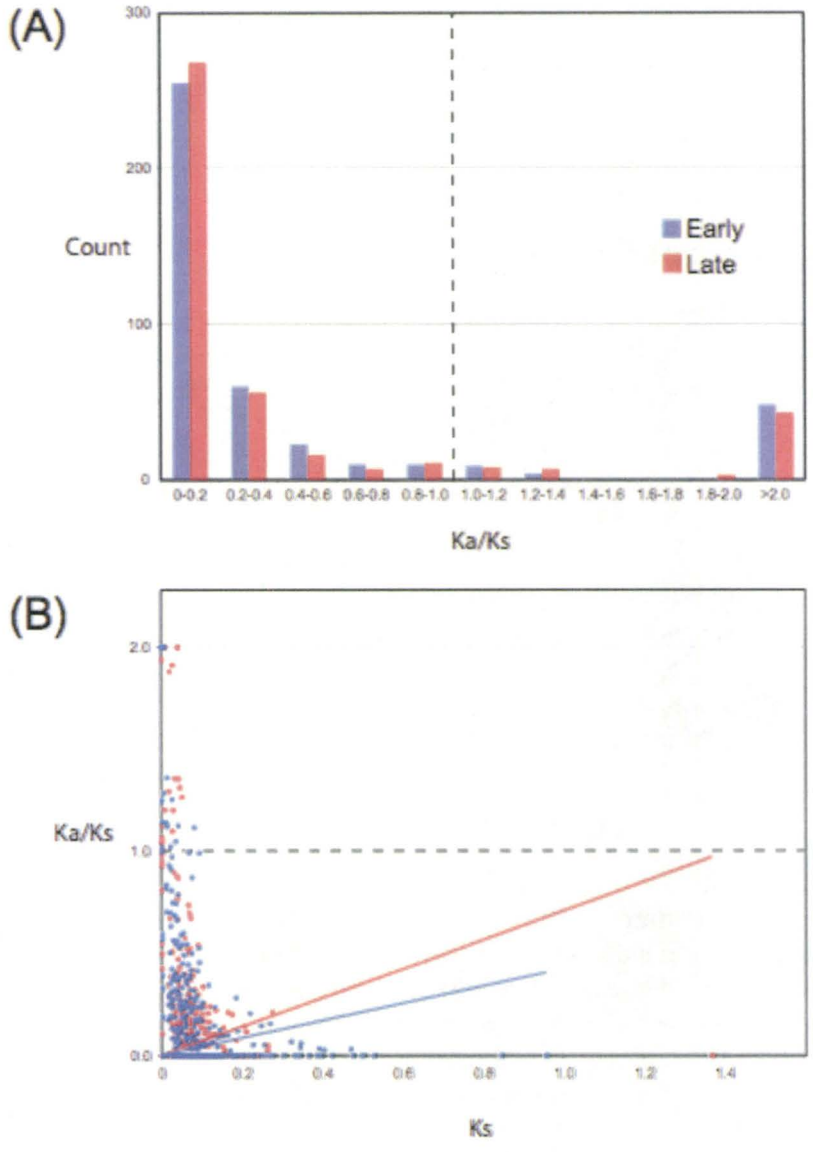


Figure 2.2 - Functional constraints are similar in early and later stages of duplicate gene evolution in *X. laevis* paralogs. (A) Binned Ka/Ks of early (blue) and later (red) stages of duplicate gene evolution. (B) Regression of Ka/Ks versus Ks in the early and later stages indicates that selection (relaxed purifying + positive) is not more common in the early stage of duplicate gene evolution (blue dots) than the later stage (red dots). The Y-intercept of these regression lines was set to zero and Ka/Ks ratios greater 2 (including undefined ratios) were given a value of 2. In (A) and (B), a dashed line indicates the neutral expectation. Fragments with Ka/Ks > 2 are on average half of the size of those with Ka/Ks < 2. Ka/Ks ratios above 2 may therefore be attributable in part to stochastic variance in Ks (Wyckoff et al. 2005).

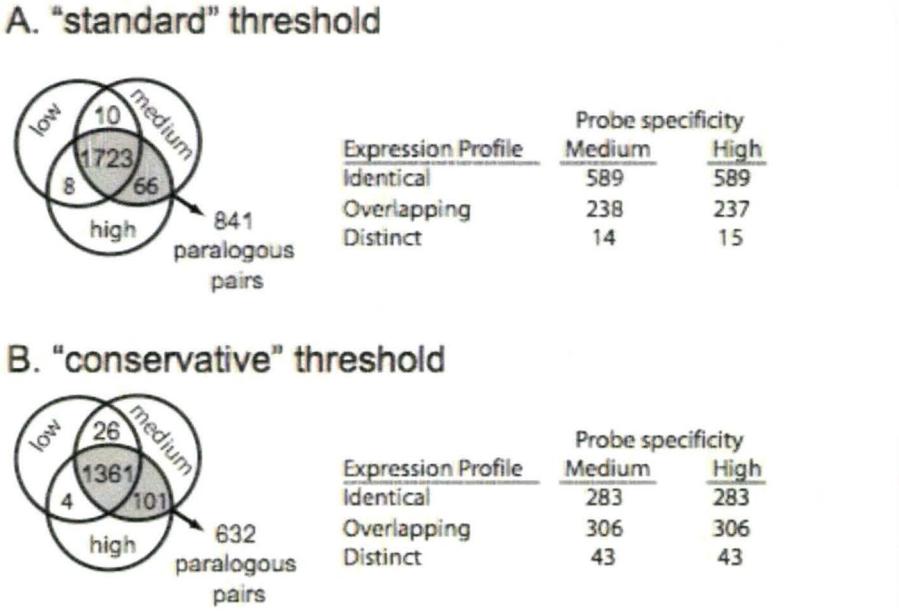


Figure 2.3 – Expression of both paralogs is generally detected in the same treatments, irrespective of the probe specificity (the degree to which each probe matches one but not the other paralog) or the detection threshold (the minimum raw intensity scored as expressed). These data are based on (A) "Standard" and (B) "Conservative" threshold levels for detection of expression and three probe specificities were compared that are labeled low, medium, and high (see Methods). We report paralogous profiles whose presence/absence scores in all five treatments were identical in the medium and high specificity at the standard threshold (shaded in gray on the left of each chart). 1789 and 1462 genes had consistent present/absent expression profiles in the medium and high specificity analyses using the standard and conservative thresholds. These sets of genes included 841 and 632 paralogous pairs, respectively. The tables on the right compare paralogous profiles by tabulating whether they are both present or absent in the same treatments (identical), the expression profile of one overlaps entirely with the other (overlap), or paralogs in which each duplicate has a unique component (distinct).

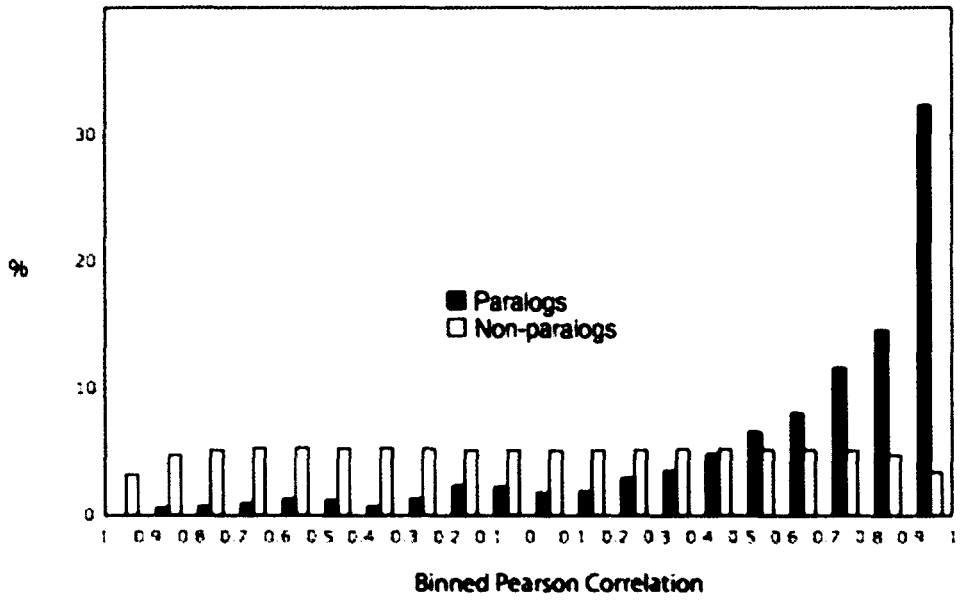


Figure 2.4 - Binned expression profile correlations between 841 pairs of paralogs over five developmental stages or adult tissue types in the medium specificity analysis. The proportion of Pearson correlation coefficients between non-paralogous expression profiles (white bars) and between paralogous expression profiles (black bars). Ninety percent of the non-paralogous expression profiles have a Pearson correlation coefficient that is greater than –0.861 but less than 0.865. The Pearson correlation coefficients of 62% of the paralogous expression profiles are less than 0.865, and 0.3% of them are less than –0.861.

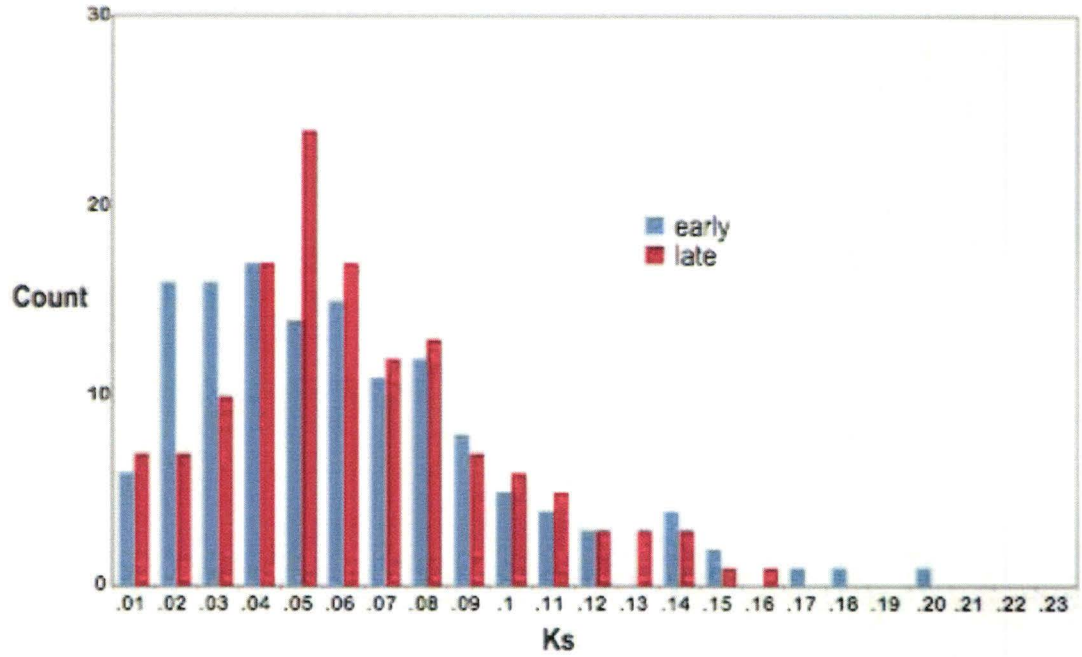


Figure 2.5 - Binned rates of synonymous substitution per site (Ks) of paralog α of gene fragments greater than 200 bp suggest that Ks is lower in the early stage than in the later stage. Ks values were calculated using a free ratio model on the phylogeny depicted in Figure 2.1B in which Ks is estimated independently for each branch. The early stage of evolution (blue bars) corresponds with the paralog α lineage between node 1 and 3 and the later stage of evolution (red bars) corresponds with the XLα lineage between node 3 and XLα.

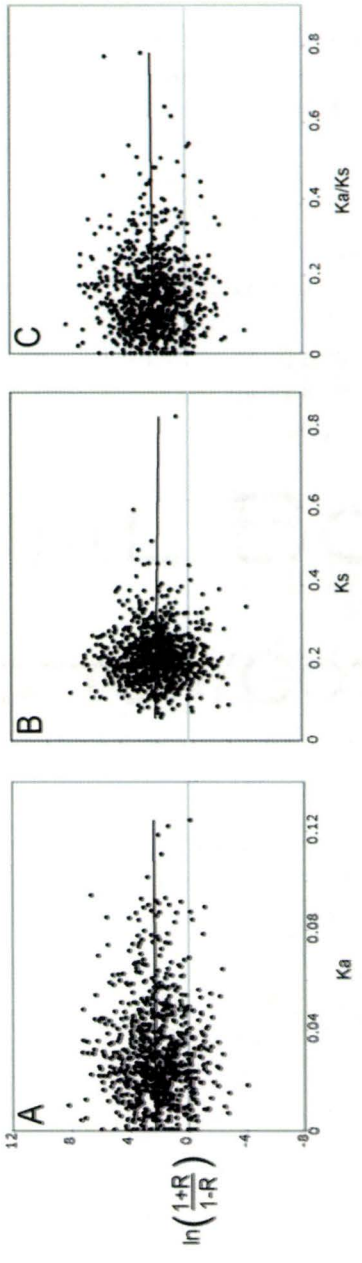


Figure 2.6 - No correlation between expression divergence and (A) Ka, (B) Ks, or (C) Ka/Ks ($R^2 \approx 0.0002$ and $P > 0.50$ for all correlations). Expression divergence is quantified by ln(1+R)/(1-R) where R is the Pearson correlation coefficient between each paralogous expression profile (Gu et al. 2002). In (C) two outliers that have a Ka/Ks ratio over 1 are excluded. There also is not a significant correlation between the Ka/Ks ratio and ln(1+R)/(1-R) (data not shown). Ka/Ks ratios were calculated from complete or large fragments of expressed *X. laevis* paralogs; the average length of these sequences was 1119 bp.

Table 2.1 - Comparison of alternatively parameterized models of evolution in Figure 2.1 indicates no significant difference in the Ka/Ks ratio at an early and a later stage of duplicate gene evolution. Indicated for comparisons depicted in Figure 2.1B, C and D are likelihoods of the null model (early and later Ka/Ks are the same) and the alternative model (early and later Ka/Ks are not the same), the one-sided probability of the Ka/Ks ratio being higher in the early stage, and the Ka/Ks ratios estimated from each of these models. For the first two tests, the Ka/Ks ratio of the diploid lineage was estimated using a different model where a unique Ka/Ks ratio was estimated for each branch (a free ratio model). Also listed are the joint likelihoods of these models from an analysis partitioned by gene fragment. For the partitioned analyses, Ka/Ks ratios for each fragment are either listed in Table 2.3, not listed (NL), or not applicable (NA).

Comparison	# base pairs	-lnL Ho	-lnL Ha	P value	Ka/Ks combined early and late	Ka/Ks ratio early	Ka/Ks ratio late	Ka/Ks diploid
Fig. 1B	80856	-165602.720	-165602.386	0.414	0.164	0.158	0.169	0.126
Fig. 1C	9717	-15699.366	-15697.250	1.000	0.208	0.124	0.346	0.198
Fig. 1D	6966	-13187.865	-13186.872	0.160	0.126	0.187	0.105	NA
Fig. 1B (partitioned)	80856	-160085.863	-159889.926	1.000	NL	Af2	Af2	NL
Fig. 1C (partitioned)	9717	-15400.349	-15393.089	0.888	NL	Af2	Af2	NL
Fig. 1D (partitioned)	6966	-12983.343	-12978.034	0.807	NL	Af2	Af2	NA

Table 2.2 - Comparison of alternatively parameterized models of evolution indicates significant departure from neutrality at an early stage of duplicate gene evolution. Likelihoods of a null model with the Ka/Ks ratio fixed at one at an early stage of duplicate gene evolution and an alternative model with this ratio estimated are indicated. Species acronyms are the same as in Figure 2.1 and abbreviations are the same as in Table 2.1.

Comparison	# bp	lnL Ho	lnL Ha	P value	1 anc Ka/Ks ratio in early lineage in null model	Estimated Ka/Ks ratio in other lineages in null model	Estimated Ka/Ks ratio in early lineage in alternative model	Estimated Ka/Ks ratio in other lineages in alternative model
Fig 1B	808456	-1660377.7641	-1656088.1773	0.0000	1	0.1322	0.158	0.1415
Fig 1C	9717	-15716.3195	-15698.3601	0.0000	1	0.2261	0.2447	0.228
Fig 1D	6966	-13795.9739	-13787.0398	0.0000	1	0.114	0.1352	0.1557
Fig 1B (partitioned)	808456	-1660755.0615	-1660711.0687	0.0000	1	NL	NL	NL
Fig 1C (partitioned)	9717	-15416.4439	-15413.5068	0.0000	1	NL	NL	NL
Fig 1D (partitioned)	6966	-13077.6404	-13076.5679	0.0000	1	NL	NL	NL

Table 2.3 - Information about sequence data including gene acronym, length in base pairs (bp), and Genbank accession numbers, and results of model based analysis of individual fragments. Gene acronyms refer to the name of one *Xenopus* paralog or, if a name was not available, an acronym of a closely related named homolog. *Xenopus borealis* sequences less than 50 bp in length were not submitted to Genbank and are available upon request (AUR). Species and paralog abbreviations are the same as in Figure 2.1. Discontinuous fragments of the same paralog have separate accession numbers. For each fragment, the likelihood of a null (Ho) and alternative (Ha) model of evolution is listed for two tests that correspond with the combined analyses presented in Tables 2.1 and 2.2. If the P value is greater than 0.05 the null model is not rejected. For the first test, in which the alternative model has a different Ka/Ks ratio in the early and later stages of duplicate gene evolution, the estimated Ka/Ks ratios are listed. Note that the null model of no difference between these ratios is not rejected for most fragments.

XLα, XBα, XLβ, and ST; models are tested on the phylogeny depicted in Fig. 1B.

						One way test for significantly higher Ka/Ks ratio in early than later stages of duplicate gene evolution.					One way test for significant departure from neutrality in early stage of duplicate gene evolution.		
Gene	bp	Xlα	XLβ	XB	ST	-lnL Ho	-lnL Ha	P value	Ka/Ks early	Ka/Ks late	-lnL Ho	-lnL Ha	P value
1N569	72	BC072143	BC051601	EU441281 (XBβ)	BC075546	-133.62	-133.62	0.994	>2	0.0001	-133.81	-133.62	0.533
Aadat	114	BC071002	BC059968	EU441282 (XBα)	BC087790	-250.61	-250.61	0.995	0.2455	0.0001	-251.28	-250.61	0.249
Acly	63	BC084776	BC084253	EU441639 (XBα)	BC080908	-110.28	-110.28	1.000	0.0001	0.0001	-110.38	-110.28	0.660
Acly	42	BC084776	BC084253	EU441283 (XBβ)	BC080908	-72.40	-72.40	1.000	0.5021	0.0001	-72.40	-72.40	0.995
Activin	1137	U49914	BC077763	EU441284 (XBα)	-	-2190.76	-2190.70	0.750	0.0389	0.0595	-2194.96	-2190.75	0.004
Adipor1	93	BC044035	BC095928	EU441285 (XBβ)	BC080374	-162.43	-161.27	0.128	0.383	0.0001	-162.33	-162.13	0.530
ADPrh	72	BC072297	BC081147	EU441286 (XBβ)	-	-138.29	-136.29	1.000	0.0001	0.0001	-138.67	-138.29	0.382
Adrp	144	AF184090	BC082351	EU441287 (XBα)	-	-328.68	-328.34	0.413	>2	0.88	-330.24	-329.91	0.419
Adrp	90	AF184090	BC082351	EU441288 (XBβ)	-	-151.20	-151.20	1.000	0.0001	0.0001	-152.31	-151.20	0.136
Adssl1	60	BC093568	BC080025	EU441290 (XBβ)	BC075419	-137.45	-136.42	0.151	>2	0.0001	-137.44	-137.28	0.573
Adssl1	174	BC093568	BC080025	EU441289 (XBα)	BC075419	-351.89	-351.89	1.000	0.0001	0.0001	-353.60	-352.73	0.187
AIP1	63	BC060007	BC077202	EU441291 (XBα)	BC075588	-120.21	-119.69	0.307	0.0001	0.1383	-121.81	-119.73	0.041
Alburnin	93	M18350	M21442	EU441292 (XBα)	-	-201.48	-200.76	0.231	0.0001	>2	-201.68	-200.87	0.204
Alcam	105	BC074313	BC073670	EU441293 (XBα)	CT030562	-207.73	-207.10	0.262	0.0001	0.0407	-210.52	-207.20	0.010
ALDH	1230	AB016717	AB016718	EU441294 (XBβ)	-	-2272.00	-2271.63	0.391	0.0681	0.1572	-2277.50	-2271.86	0.001
Aldh3a2	162	BC088905	BC071106	EU441295; EU441640 (XBα)	BC091032	-319.41	-318.77	0.259	0.8468	0.133	-318.84	-318.83	0.900
Alpha	252	BC073294	BC075196	M32455 (XBα)	-	-619.54	-619.51	0.801	0.129	0.1823	-620.68	-619.75	0.173
An2	72	BC054959	BC080064	EU441296 (XBβ)	BC091067	-142.05	-142.05	1.000	0.0001	0.0001	-142.58	-142.23	0.403
Anapc10	105	BC090191	BC089086	EU441297 (XBα)	BX732858	-154.51	-154.51	1.000	0.0001	0.0001	-155.94	-154.51	0.091
Ankrd45	462	BC087400	BC087290	EU441298; EU441641 (XBα)	BC084990	-1002.51	-1001.36	0.129	1.2706	0.3061	-1001.48	-1001.43	0.758
Anp32B	324	BC073408	DQ096872	EU441299; EU441642; EU441643 (XBα)	BC090607	-552.71	-552.53	0.554	0.1999	0.0736	-553.40	-552.91	0.322
Anp32B	39	BC073408	DQ096872	AUR (XBβ)	BC090607	-74.79	-73.67	0.134	0.0001	>2	-76.33	-75.87	0.338
Anxa1	1011	BC074339	BC053786	EU441300 (XBα)	BC075412	-2224.92	-2224.80	0.624	0.273	0.3804	-2228.92	-2225.14	0.006
Anxa1like	735	BC054187	BC075151	EU441301; EU441644; EU441645 (XBα)	BC064261	-1514.79	-1513.74	0.149	0.5444	0.1222	-1514.09	-1513.93	0.564
Anxa2	486	BC044693	BC042238	EU441302; EU441646; EU441647 (XBα)	BC075523	-916.90	-916.39	0.311	0.0098	0.1768	-921.88	-916.39	0.001
Anxa2	270	BC044693	BC042238	EU441648; EU441649 (XBβ)	BC075523	-577.84	-577.82	0.841	0.1349	0.0965	-578.37	-577.83	0.299
Anxa4	63	BC073582	BC060389	EU441303 (XBα)	BC084910	-105.99	-105.99	1.000	0.0001	0.0001	-106.59	-106.59	1.000
Anxa5	123	BC077642	BC071097	EU441304 (XBβ)	BC082506	-264.95	-264.25	0.236	0.2978	0.0001	-264.91	-264.54	0.390
Anxe7	90	BC078086	BC081070	EU441306 (XBβ)	BC068035	-220.94	-220.93	0.921	0.5068	0.423	-221.03	-220.93	0.666
Anxe7	66	BC078086	BC081070	EU441305 (XBα)	BC068035	-114.79	-114.79	1.000	0.0001	0.0001	-116.95	-115.42	0.080
AP1M2	78	BC077578	BC070627	EU441307 (XBβ)	BC076939	-162.60	-162.60	1.000	0.0001	0.0001	-166.53	-162.60	0.005
Ap2m1	93	BC072057	BC047969	EU441308 (XBα)	BC061374	-162.05	-162.05	1.000	0.0001	0.0001	-164.78	-162.05	0.020
Ap2m1	84	BC072057	BC047969	EU441309 (XBβ)	BC061374	-134.48	-134.48	1.000	0.0001	0.0001	-135.18	-134.48	0.237
Apba2	471	BC084963	BC078109	EU441310; EU441650; EU441651; EU441652; EU441653 (XBβ)	-	-850.21	-850.21	1.000	0.0001	0.0001	-856.02	-851.44	0.002
ApoA1	246	BC053783	BC041498	EU441311; EU441654; AUR (XBβ)	BC077663	-509.48	-508.58	0.179	0.2806	>2	-510.54	-509.73	0.202
Aprin	72	BC086289	BC100220	EU441313 (XBβ)	-	-150.91	-149.10	0.058	0.0001	0.1991	-154.47	-149.10	0.001
Aprin	81	BC086289	BC100220	EU441312 (XBα)	-	-122.91	-122.91	1.000	0.0001	0.0001	-123.11	-122.91	0.530
Ari4	87	BC082392	BC077817	EU441314 (XBα)	BC075432	-143.84	-143.84	1.000	0.0001	0.0001	-144.75	-143.84	0.179
Arpc1	33	BC045043	BC073411	AUR (XBα)	BC076887	-46.61	-45.14	0.087	0.0001	>2	-46.64	-45.14	0.084
ATP5c1	210	BC068867	BC072367	EU441315; EU441655; AUR (XBα)	NM001016407	-364.82	-363.32	0.084	0.0001	0.1556	-373.83	-363.42	0.000
ATP6AP2	87	BC097537	BC056060	EU441316 (XBα)	BC088056	-161.42	-160.56	0.188	0.0001	>2	-164.52	-161.33	0.012
ATP6AP2	192	BC097537	BC056060	EU441317 (XBβ)	BC088056	-364.36	-364.36	1.000	0.0001	0.0001	-369.73	-365.11	0.002
Atp6v1d	57	BC072761	BC077888	EU441318 (XBβ)	BX743285	-137.06	-136.16	0.180	0.0001	1.3499	-138.07	-137.07	0.158
ATPssb	81	AF187862	BC082702	EU441320 (XBβ)	-	-134.71	-134.71	1.000	0.0001	0.0001	-135.65	-134.98	0.249
ATPssb	357	AF187862	BC082702	EU441319; EU441656; EU441657 (XBα)	-	-738.19	-738.01	0.541	0.2619	0.0965	-738.40	-738.04	0.395
Axin2	96	BC082364	AF140243	EU441321 (XBα)	-	-204.24	-203.68	0.289	0.0992	0.0001	-206.24	-204.82	0.092

Table 2.3 (continued 1)

B0at1	93	BC081075	BC081049	EU441322 (XBα)	-	-159.76	-159.76	1.000	0.0001	0.0001	-161.86	-159.76	0.040
B3gnt5	288	BC088967	BC077332	EU441323; EU441658 (XBα)	BC063912	-632.88	-632.82	0.742	0.5243	0.7598	-633.43	-633.11	0.425
Barren	93	BC056095	BC068643	EU441324 (XBα)	-	-169.02	-169.02	1.000	0.0001	0.0001	-172.56	-169.30	0.011
Bat3	69	BC059307	BC060479	EU441325 (XBβ)	CT485713	-128.34	-127.88	0.342	0.3438	0.0001	-128.54	-128.32	0.506
Bcl6	60	BC077915	BC084912	EU441326 (XBβ)	-	-132.27	-132.27	0.996	0.296	0.0001	-132.35	-132.27	0.685
Beta	372	V01433	BC071139	M32456 (XBα)	-	-934.73	-933.46	0.111	0.6168	0.1222	-936.19	-936.13	0.742
BFB	114	D29796	D49373	EU441327 (XBα)	-	-265.68	-265.45	0.497	0.9944	0.3789	-265.54	-265.54	0.988
Bzrf1	66	BC068762	BC084418	EU441328 (XBβ)	BC080924	-124.08	-120.19	0.005	0.0001	>2	-129.52	-124.63	0.002
Bnip1	78	BC077345	BC054986	EU441329 (XBβ)	CX918049	-137.84	-137.84	1.000	0.0001	0.0001	-139.66	-138.21	0.088
Brap29	477	BC076818	BC072342	EU441330 (XBα)	CT030383	-985.86	-985.64	0.502	0.162	0.2999	-988.99	-985.78	0.011
Btg1	123	AB028243	BC041244	EU441331 (XBα)	-	-227.97	-227.54	0.352	>2	1.2009	-228.34	-227.71	0.261
C2orf25	102	BC097712	BC092030	EU441332 (XBα)	NM001016731	-204.80	-204.62	0.559	0.4836	0.1033	-204.69	-204.67	0.829
C6orf111	87	BC081125	BC072160	EU441334 (XBβ)	-	-193.77	-192.41	0.099	0.0001	1.1954	-194.45	-192.73	0.064
C6orf111	63	BC081125	BC072160	EU441333 (XBα)	-	-121.26	-121.26	0.997	0.1494	0.0001	-121.99	-121.26	0.227
C6orf67	408	BC045047	BC063271	EU441335 (XBα)	BC061349	-925.34	-925.05	0.447	0.2449	0.0804	-928.10	-925.69	0.028
C7orf20	240	BC074468	BC070733	EU441336; EU441659 (XBα)	NM001016708	-448.70	-448.14	0.290	0.0001	0.0569	-452.62	-448.44	0.004
Calnexin	702	BC044970	BC041719	EU441337 (XBα)	-	-1321.24	-1320.88	0.401	0.273	1.3423	-1323.59	-1323.20	0.377
Calreticulin	759	BC046699	BC044068	EU441338 (XBβ)	-	-1503.56	-1503.41	0.583	0.0545	0.0918	-1510.90	-1503.68	0.000
CAP1	93	BC041224	BC077282	EU441339 (XBβ)	BC067981	-200.33	-200.33	0.961	0.4034	0.3556	-200.52	-200.44	0.693
Carf	279	BC060029	BC044102	EU441340; EU441660 (XBβ)	CR855636	-639.29	-639.18	0.638	>2	0.5204	-640.32	-640.28	0.782
Carhsp1	105	BC084399	BC044047	EU441341 (XBα)	-	-230.95	-230.95	1.000	0.0001	0.0001	-232.43	-232.43	1.000
Cask	78	BC077777	BC092148	EU441342 (XBα)	-	-169.06	-168.32	0.227	0.0001	>2	-169.79	-169.17	0.267
Caspase10	711	BC068638	BC060356	EU441343; EU441661; EU441662; EU441663; EU441664; EU441665 (XBβ)	BC089646	-1596.18	-1594.60	0.075	0.3799	1.8687	-1598.31	-1597.17	0.130
Cav2	57	BC072274	BC077468	EU441344 (XBα)	BC089699	-104.98	-104.98	0.996	>2	0.0001	-104.99	-104.98	0.888
CB1	57	BC088950	BC041302	EU441345 (XBβ)	BC061430	-120.52	-120.52	1.000	0.0001	0.0001	-122.30	-120.52	0.059
Ccng2	102	BC074293	BC070835	EU441346 (XBα)	CU025046	-214.47	-214.47	0.978	0.2353	0.2475	-215.16	-214.75	0.363
Ccng2	75	BC074293	BC070835	EU441347 (XBβ)	CU025046	-175.71	-175.71	0.995	0.0818	0.0001	-176.85	-175.71	0.131
CD63	105	BC054957	BC077961	EU441348 (XBα)	BC080508	-188.56	-188.56	1.000	0.0001	0.0001	-190.48	-189.21	0.110
Cdc2	552	M60681	BC045078	EU441349 (XBα)	-	-1102.22	-1101.46	0.220	0.0377	0.1652	-1110.23	-1105.52	0.002
CDO1	159	BC071092	BC060414	EU441350; EU441666 (XBβ)	BC061333	-312.52	-312.02	0.320	0.0001	0.0615	-314.69	-312.07	0.022
Centrin	471	U37538	BC054948	EU441667; AUR (XBβ)	-	-803.76	-803.13	0.263	0.0634	0.0001	-805.95	-803.68	0.033
Cfh	225	BC060353	BC046950	EU441351; EU441668 (XBα)	BC063193	-532.12	-531.63	0.322	0.1335	0.5734	-532.77	-531.69	0.142
CFI	189	X59958	BC041753	EU441352; EU441669 (XBα)	-	-459.40	-458.89	0.312	0.2253	0.9728	-460.70	-459.43	0.111
Chchd2	75	BC068623	BC073239	EU441353 (XBβ)	BC080141	-121.10	-121.10	1.000	0.0001	0.0001	-122.04	-121.10	0.169
Chchd3	60	BC072860	BC042358	EU441670 (XBβ)	-	-129.35	-129.35	1.000	>2	>2	-130.39	-129.95	0.348
CHML	198	BC061662	BC078011	EU441671; AUR (XBα)	-	-431.85	-431.84	0.898	0.3962	0.4817	-432.12	-431.85	0.466
Chmp5	528	BC077776	BC041499	EU441354 (XBα)	BC089652	-943.26	-943.06	0.531	0.0001	0.0339	-944.18	-943.19	0.159
CIB	129	BC088946	BC084822	EU441355; AUR (XBα)	BC088779	-240.83	-240.27	0.287	0.1151	0.0001	-242.04	-240.72	0.104
Ckif7	66	BC056111	BC076766	EU441356 (XBβ)	BC090372	-124.10	-124.10	0.997	0.0001	0.0001	-124.10	-124.10	1.000
Clic4	72	BC076836	BC072787	EU441357 (XBβ)	BC080344	-142.89	-142.18	0.233	0.0001	0.2041	-144.90	-142.32	0.023
CNDP2	237	BC075171	BC056069	EU441358; EU441672; EU441673 (XBα)	-	-475.64	-475.40	0.488	0.0857	0.0001	-478.61	-475.84	0.019
Cnih	66	BC080391	BC078462	EU441359 (XBα)	BC080958	-110.30	-110.30	1.000	0.0001	0.0001	-114.24	-110.30	0.005
Cnot10	105	BC077237	BC068748	EU441360 (XBα)	CR761656	-219.02	-217.65	0.098	>2	0.0001	-218.12	-217.92	0.528
Commd5	573	BC075209	BC088938	EU441361 (XBβ)	BX709875	-1147.90	-1147.54	0.394	0.0328	0.1074	-1155.65	-1147.84	0.000
Cpd1	105	BC094475	BC043985	EU441362 (XBα)	CR942787	-179.22	-178.33	0.181	0.0001	0.3153	-182.42	-179.75	0.021
Cprotein	132	BC071084	BC041534	EU441363 (XBα)	BC077013	-279.93	-278.21	0.064	>2	0.1134	-278.69	-278.22	0.330
Ctdp1	378	BC092306	BC082378	EU441365; EU441676; EU441677; EU441678 (XBβ)	-	-827.56	-827.39	0.556	0.1073	0.1943	-834.18	-827.45	0.000
Ctdp1	282	BC092306	BC082378	EU441364; EU441674; EU441675 (XBα)	-	-601.94	-600.29	0.069	0.0001	0.3122	-603.42	-600.30	0.012
Cullin3	213	BC077239	BC073186	EU441367; EU441679 (XBβ)	-	-360.75	-360.75	1.000	0.0001	0.0001	-363.38	-361.28	0.041
Cullin3	135	BC077239	BC073186	EU441366 (XBα)	-	-208.67	-207.80	0.186	0.0001	0.0001	-208.67	-207.80	0.186
CyclinB2	177	BC100180	BC060466	EU441368 (XBβ)	BC080491	-379.27	-377.49	0.059	0.2098	0.0001	-380.70	-379.75	0.167

Table 2.3 (continued 2)

CyclinE	159	L23857	L43513	EU441369; EU441680 (XBβ)	-	-254.61	-254.61	0.995	>2	0.0001	-255.05	-254.61	0.348
CyclinE2	99	BC084929	BC043855	EU441370 (XBα)	CR761678	-240.02	-237.94	0.041	0.0001	>2	-241.38	-238.83	0.024
CyclinE2	69	BC084929	BC043855	EU441371 (XBβ)	CR761678	-145.24	-143.93	0.105	>2	0.0001	-144.52	-144.28	0.487
Cyp2P2	117	BC077934	BC054243	EU441372 (XBβ)	BC089651	-255.41	-253.17	0.034	0.0556	1.5168	-260.95	-256.99	0.005
DAP	201	BC077360	BC070744	EU441373; EU441681 (XBβ)	-	-386.40	-385.22	0.125	>2	0.0001	-385.78	-385.73	0.765
Dc2	138	BC087303	BC073364	EU441682; AUR (XBβ)	AL805744	-265.13	-264.66	0.333	0.3255	0.0001	-264.80	-264.66	0.601
Dc2	48	BC087303	BC073364	EU441374 (XBα)	AL805744	-91.84	-91.84	1.000	0.0001	0.0001	-92.91	-91.84	0.142
Ddah1	99	BC078574	BC081168	EU441375 (XBα)	BC075381	-221.90	-221.79	0.632	0.1833	0.0759	-222.27	-221.84	0.354
Dncic1	60	BC086292	BC070833	EU441376 (XBα)	BC067997	-97.08	-97.08	1.000	0.0001	0.0001	-99.97	-97.08	0.016
Dncl1	102	BC073042	BC057215	EU441683 (XBα)	BC076999	-170.88	-170.88	1.000	0.0001	0.0001	-173.28	-170.88	0.028
Dor2	612	BC081046	BC073700	EU441377; EU441684; EU441685; EU441686; EU441687; EU441688 (XBβ)	CR926338	-1187.21	-1184.52	0.020	1.0885	0.0562	-1184.86	-1184.86	0.930
DP1	222	BC080383	BC082841	EU441691; EU441692; AUR (XBβ)	BC087763	-449.22	-449.06	0.572	0.6635	0.0001	-449.39	-449.22	0.561
DP1	315	BC080383	BC082841	EU441378; EU441689; EU441690 (XBα)	BC087763	-550.15	-550.15	1.000	0.0001	0.0001	-559.06	-550.52	0.000
Dpys13	42	BC082618	BC046836	AUR (XBα)	BC074633	-76.27	-76.27	1.000	0.0001	0.0001	-78.05	-76.27	0.059
Dystroglycan	81	BC046260	BC073500	EU441379 (XBα)	-	-146.70	-146.30	0.371	0.0001	0.0001	-146.70	-146.30	0.371
EDNRA	51	BC044316	U06633	EU441380 (XBβ)	-	-94.09	-94.09	1.000	0.0001	0.0001	-94.60	-94.09	0.313
EEF1D	87	BC072139	BC088696	EU441381 (XBα)	BC088544	-193.23	-193.23	1.000	0.0001	0.0001	-193.43	-193.23	0.526
EF	87	X52976	X52977	EU441382 (XBβ)	-	-176.86	-176.86	0.992	0.0001	0.0001	-178.30	-177.52	0.211
EIF1A	75	BC068786	BC074155	EU441383 (XBα)	BC074588	-115.54	-115.54	1.000	0.0001	0.0001	-118.05	-115.54	0.025
eIF4e	246	BC078129	BC089136	EU441693 (XBα)	NM001015909	-440.30	-440.28	0.843	0.1398	0.2134	-441.08	-440.29	0.209
eIF4e	93	BC078129	BC089136	EU441384 (XBβ)	NM001015909	-149.26	-149.26	1.000	0.0013	0.0001	-149.26	-149.26	1.000
Eml1	603	BC097694	BC088910	EU441386; EU441694 (XBβ)	CR761353	-1418.77	-1418.77	0.990	0.2366	0.2343	-1421.64	-1418.90	0.019
Eml1	102	BC097694	BC088910	EU441385 (XBα)	CR761353	-271.05	-270.35	0.237	0.0434	0.4183	-273.12	-270.41	0.020
Encr1	63	BC073388	BC046953	EU441387 (XBα)	BC080910	-106.63	-106.63	0.993	0.0001	>2	-107.43	-106.63	0.204
ENO	360	BC041279	BC054169	EU441389 (XBβ)	-	-733.57	-733.57	1.000	0.0001	0.0001	-740.81	-735.18	0.001
ENO	516	BC041279	BC054169	EU441388 (XBα)	-	-1103.25	-1097.49	0.001	0.4722	0.0253	-1099.19	-1098.31	0.186
Enpp2	99	BC044675	BC089138	EU441390 (XBα)	CR926368	-213.04	-212.83	0.514	0.3922	0.0965	-212.96	-212.90	0.735
Enpp2	72	BC044675	BC089138	EU441391 (XBβ)	CR926368	-141.65	-141.28	0.393	0.0001	0.1711	-142.86	-141.49	0.098
EZ	270	AF351126	BC097526	EU441392; EU441695 (XBα)	-	-524.54	-524.54	1.000	0.0001	0.0001	-529.50	-526.37	0.012
Fabp4	105	BC061929	BC078499	EU441393 (XBα)	-	-234.60	-233.32	0.110	>2	0.1239	-234.05	-233.75	0.438
Faim2	132	BC074272	BC074388	EU441394 (XBβ)	-	-274.51	-274.45	0.728	0.0001	0.2211	-274.57	-274.51	0.728
Faim2	45	BC074272	BC074388	AUR (XBα)	-	-68.15	-68.15	1.000	>2	>2	-68.47	-68.15	0.424
Fam79	162	BC063732	BC082905	EU441696; EU441697 (XBα)	-	-331.02	-330.63	0.376	0.129	>2	-331.10	-331.02	0.697
Fam79	75	BC063732	BC082905	EU441395; AUR (XBβ)	-	-144.65	-144.65	1.000	0.0001	0.0001	-145.94	-144.65	0.108
Fbxo8	105	BC073016	BC074433	EU441396 (XBβ)	-	-182.08	-182.08	0.997	0.1845	0.0001	-182.70	-182.08	0.266
FGFR	297	M55163	U24491	EU441397 (XBα)	-	-603.01	-602.70	0.432	0.2898	0.0595	-602.81	-602.70	0.634
Fizzy1	150	BC042288	BC076805	EU441398; EU441698 (XBβ)	BC061363	-297.45	-297.06	0.374	0.065	0.0001	-299.67	-297.30	0.030
FLT1	66	BC044264	BC056023	EU441399 (XBβ)	-	-180.94	-180.66	0.454	0.8685	0.0001	-180.95	-180.94	0.911
Formlnbp21	225	BC077383	BC077747	EU441400 (XBα)	CR761869	-506.26	-506.15	0.631	0.3007	0.1609	-507.52	-506.83	0.242
Fp	246	BC047261	BC060446	EU441401; EU441699 (XBβ)	-	-500.15	-500.05	0.658	0.2476	0.1258	-501.30	-500.06	0.115
Fscn1	153	BC077847	BC097600	EU441402; EU441700 (XBβ)	-	-275.90	-275.90	1.000	0.0001	0.0001	-282.76	-278.01	0.002
Fundc1	174	BC076744	BC092015	EU441701; EU441702; AUR (XBβ)	BX744732	-331.94	-328.76	0.012	0.0001	>2	-334.05	-330.39	0.007
Fundc1	96	BC076744	BC092015	EU441403; AUR (XBα)	BX744732	-158.16	-155.79	0.029	0.0001	>2	-161.33	-158.18	0.012
Furin	171	M80471	AY901983	EU441404 (XBβ)	-	-377.93	-377.70	0.495	1.0671	>2	-378.62	-378.61	0.968
FXR1	333	DQ083375	BC043638	EU441405; EU441703; EU441704 (XBβ)	-	-612.15	-612.14	0.874	0.0559	0.0726	-615.57	-612.14	0.009
Gabpa	66	BC077619	BC100222	EU441406 (XBα)	CT025182	-114.87	-114.79	0.681	0.0001	0.0001	-114.87	-114.79	0.681
Gadd45g	81	BC045055	BC078567	EU441407 (XBα)	BC075545	-173.79	-173.79	0.941	0.3537	0.4278	-174.28	-174.14	0.595
Gadd45g	93	BC045055	BC078567	EU441408 (XBβ)	BC075545	-165.50	-165.50	1.000	0.0001	0.0001	-165.55	-165.50	0.756
Galectin	99	AB060970	BC081109	EU441409 (XBβ)	-	-225.66	-223.02	0.022	0.0001	>2	-227.48	-225.78	0.065
Galectin1	216	AB056478	BC053816	EU441410 (XBβ)	-	-452.94	-452.50	0.350	1.3807	0.2382	-452.57	-452.55	0.864

Table 2.3 (continued 3)

Gaq	165	L05540	BC081126	EU441411; EU441705 (XBβ)	-	-303.95	-303.95	1.000	0.0001	0.0001	-308.93	-304.72	0.004
Gas6	81	BC060355	BC076835	EU441706 (XBα)	NM001015965	-204.57	-204.50	0.715	0.4746	0.2197	-204.60	-204.51	0.676
Gbe1	93	BC084621	BC076746	EU441412 (XBβ)	CX882649	-195.27	-195.27	1.000	0.0001	0.0001	-196.43	-195.56	0.185
Gcap1	60	BC080088	BC074337	EU441413 (XBα)	-	-99.88	-99.88	1.000	0.0001	0.0001	-101.29	-99.88	0.093
Gcnt1	111	BC084257	BC048021	EU441414 (XBα)	-	-196.47	-195.56	0.178	0.0001	0.296	-198.39	-195.76	0.022
Geminin	63	AF067856	AF068781	EU441415; AUR (XBα)	NM001039736	-116.28	-116.28	1.000	0.0001	0.0001	-117.52	-116.28	0.116
Geminin	441	AF067856	AF068781	EU441416; EU441707 (XBβ)	NM001039736	-975.78	-973.14	0.022	0.4523	0.0462	-975.50	-974.94	0.289
Gene.20	192	BC084087	BC084136	EU441708 (XBα)	CT485685	-366.43	-366.23	0.532	0.0908	0.2586	-369.34	-367.69	0.069
Gene.743	114	BC070984	BC094404	EU441710; AUR (XBβ)	-	-239.48	-239.30	0.551	0.0001	0.0952	-240.93	-240.01	0.176
Gene.743	99	BC070984	BC094404	EU441709 (XBα)	-	-221.88	-221.58	0.441	>2	0.2279	-221.83	-221.75	0.679
Gene.863	81	BC072938	BC075122	EU441711 (XBα)	NM001016812	-183.15	-183.15	1.000	0.0001	>2	-183.94	-183.94	0.999
Gene.863	348	BC072938	BC075122	EU441712; EU441713; EU441714 (XBβ)	NM001016812	-727.59	-727.47	0.624	0.2464	0.1411	-728.99	-727.48	0.082
Gene.920	81	BC056657	BC075214	EU441715 (XBα)	CR761813	-171.66	-171.66	0.993	0.204	0.2007	-173.03	-172.01	0.154
Gene77	126	BC078463	BC073656	EU441716 (XBα)	NM001016479	-271.50	-268.98	0.025	>2	0.0001	-271.25	-270.79	0.337
GeneXL.310	285	BC081150	BC072032	EU441717; EU441718 (XBβ)	BC081310	-587.59	-587.58	0.926	0.5343	0.4729	-588.02	-587.84	0.541
GeneXL.367	132	BC046664	BC041210	EU441719; EU441720 (XBβ)	BC090115	-284.87	-284.41	0.334	>2	0.6814	-285.36	-284.94	0.360
Gkap42	192	BC073450	BC041228	EU441417; AUR (XBα)	BC087974	-389.42	-388.44	0.162	>2	0.3538	-389.40	-388.61	0.207
Gkap42	516	BC073450	BC041228	EU441418; EU441721; EU441722 (XBβ)	BC087974	-1045.49	-1045.13	0.395	0.7049	0.251	-1045.71	-1045.67	0.780
Glo1	69	BC080129	BC076752	EU441419 (XBα)	BC090582	-125.91	-125.91	1.000	0.0001	0.0001	-125.94	-125.91	0.830
Gltp	168	BC068757	BC082857	EU441420 (XBα)	AL967440	-361.43	-361.43	0.973	0.3287	0.3078	-361.82	-361.69	0.605
Gmfg	204	BC068858	BC078467	EU441421 (XBα)	BC091700	-343.41	-343.41	1.000	0.0001	0.0001	-344.78	-343.60	0.125
Gorasp2	69	BC097604	BC043840	EU441422 (XBα)	BC074543	-166.31	-165.57	0.225	0.0001	>2	-166.71	-166.09	0.264
Gpsn2	105	BC073263	BC056018	EU441423 (XBβ)	BC080144	-201.41	-200.48	0.171	0.0898	>2	-205.40	-203.61	0.059
GSTS1	69	BC053774	BC060462	EU441424 (XBα)	BC077016	-137.88	-137.33	0.294	0.4648	0.0001	-137.88	-137.79	0.665
Hdgf	141	BC072083	BC084058	EU441425; AUR (XBα)	BC090580	-273.21	-272.82	0.380	0.1258	0.0001	-275.05	-272.98	0.042
Hech	294	BC079995	BC046570	EU441426; EU441725 (XBβ)	CT025206	-516.19	-515.29	0.182	0.1211	0.0001	-516.26	-515.64	0.264
Hech	114	BC079995	BC046570	EU441723; EU441724 (XBα)	CT025206	-172.24	-172.24	1.000	0.0001	0.0001	-173.99	-172.24	0.061
Herpud1	69	BC054211	BC046268	EU441427 (XBβ)	BC067925	-91.08	-91.08	1.000	0.0001	0.0001	-92.86	-91.08	0.060
Hgene.137	390	BC082840	BC087359	EU441726; EU441727; EU441728; EU441729 (XBβ)	CX333056	-708.45	-708.21	0.486	0.0416	0.1088	-713.27	-708.34	0.002
Hgene.383	105	BC087518	BC075254	EU441730 (XBβ)	BC082533	-221.84	-220.77	0.144	>2	0.1349	-221.09	-220.83	0.476
Hgene.462	90	BC080012	BC074271	EU441731 (XBα)	CX426943	-181.10	-180.26	0.194	0.5471	0.0001	-181.18	-181.11	0.701
Hgene.478	180	BC094151	BC081015	EU441732; EU441733 (XBα)	BC088770	-424.37	-423.43	0.169	>2	0.7268	-425.75	-424.83	0.175
Hibadh	90	BC084329	BC070849	EU441428 (XBβ)	BC091056	-181.03	-177.48	0.008	0.0001	>2	-182.44	-178.40	0.004
Hip1	96	BC070829	BC077182	EU441429 (XBβ)	-	-189.86	-189.04	0.200	0.236	0.0001	-190.47	-189.38	0.139
Hmg3	519	BC044009	AY652624	EU441430; EU441734 (XBα)	BC075290	-1031.31	-1030.30	0.156	0.2617	0.0615	-1031.92	-1030.92	0.157
Hmgb1	354	BC073449	BC054148	EU441431; EU441735 (XBβ)	BC063332	-713.89	-713.12	0.216	0.0337	0.1806	-718.23	-713.79	0.003
Hmgb1	39	BC073449	BC054148	AUR (XBα)	BC063332	-49.75	-49.75	0.995	0.0001	0.0001	-49.75	-49.75	1.000
HnrpAB	273	BC074212	BC043814	EU441432; EU441736; EU441737 (XBβ)	CT030339	-587.98	-587.94	0.788	0.0469	0.0812	-589.54	-587.95	0.074
Hnrpdl	105	BC068788	BC045124	EU441738 (XBα)	BC082729	-170.93	-170.06	0.186	0.2568	0.0001	-170.92	-170.50	0.361
Hs3st1	105	BC063731	BC075183	EU441433 (XBβ)	BC088545	-223.44	-223.14	0.443	0.0001	0.1801	-226.63	-223.14	0.008
Hsp5	108	BC041200	BC077757	EU441434 (XBα)	-	-209.75	-209.75	1.000	0.0001	0.0001	-221.22	-209.83	0.000
Hsp5	393	BC041200	BC077757	EU441435; EU441739; EU441740; EU441741 (XBβ)	-	-735.02	-734.95	0.707	0.0681	0.0443	-740.38	-734.99	0.001
Hsp8	180	BC046262	BC041201	EU441436; EU441742 (XBβ)	-	-378.46	-377.46	0.158	0.1801	0.0001	-377.97	-377.59	0.388
HSPA9B	438	BC045130	BC045259	EU441437; EU441743 (XBβ)	BC067910	-829.29	-829.29	1.000	0.0001	0.0001	-833.97	-830.39	0.008
Hspd1	201	BC046687	BC072058	EU441744; EU441745; AUR (XBα)	CT025413	-404.32	-402.92	0.094	0.0001	0.1706	-408.73	-402.96	0.001

Table 2.3 (continued 4)

Ifrd1	39	BC057749	BC083040, AF549961	EU441438 (XBβ)	NM001016124	-77.71	-77.01	0.235	0.319	0.0001	-77.90	-77.63	0.462
Impdh2	72	BC044122	BC042315	EU441439 (XBα)	BC080955	-169.98	-169.98	1.000	0.0001	0.0001	-171.58	-169.98	0.074
Ing5	144	BC084763	BC084202	EU441440; EU441746 (XBβ)	BC059768	-235.71	-235.71	1.000	0.0001	0.0001	-238.38	-236.08	0.032
Integrin	126	M20140	M20180	EU441441 (XBβ)	-	-223.50	-223.50	1.000	0.0001	0.0001	-226.01	-223.50	0.025
Irap	255	BC059334	BC084749	EU441443; EU441747 (XBβ)	-	-641.06	-639.62	0.090	0.0001	0.7589	-642.06	-639.62	0.027
Irap	78	BC059334	BC084749	EU441442 (XBα)	-	-164.78	-164.78	1.000	0.0001	0.0001	-168.28	-166.38	0.051
Irf2	141	BC077187	BC087378	EU441444; EU441748 (XBα)	BC080892	-324.67	-324.34	0.416	0.6216	>2	-324.89	-324.83	0.745
Irf2	24	BC077187	BC087378	AUR (XBβ)	BC080892	-36.80	-36.69	0.638	>2	0.0001	-36.53	-36.49	0.796
IRF6	96	BC071111	D86492	EU441445 (XBα)	NM001030322	-217.67	-216.64	0.151	0.0387	>2	-222.10	-217.13	0.002
Itgb1bp3	459	BC084382	BC071072	EU441446; EU441749 (XBβ)	BC080978	-864.89	-864.74	0.584	0.2683	0.5596	-867.69	-867.03	0.249
ITM2A	348	BC072047	BC044320	EU441447; EU441750; EU441751; EU441752 (XBα)	BC074629	-726.04	-720.93	0.001	0.0262	1.1233	-734.70	-722.21	0.000
ITM2A	126	BC072047	BC044320	EU441448 (XBβ)	BC074629	-233.32	-232.86	0.334	0.2329	0.0001	-233.37	-232.93	0.350
Kcnip1	93	BC082465	BC074264	EU441449 (XBα)	-	-153.87	-153.87	1.000	0.0001	0.0001	-154.24	-153.87	0.391
Kcnip1	18	BC082465	BC074264	AUR (XBβ)	-	-33.37	-32.40	0.163	>2	0.0001	-32.63	-33.20	1.000
Kdelr1	222	BC060380	BC080126	EU441450 (XBβ)	BC084170	-437.35	-436.22	0.133	0.0001	0.1685	-440.87	-436.89	0.005
KIAA0103	258	BC072200	BC041255	EU441451; EU441753 (XBβ)	BC088602	-449.40	-449.33	0.702	0.0576	0.0001	-453.91	-449.39	0.003
KIAA1387	207	BC045225	BC047970	EU441452; EU441754 (XBα)	BC076650	-358.47	-358.28	0.538	0.0503	0.0001	-362.86	-358.59	0.003
KIAA1387	63	BC045225	BC047970	EU441453 (XBβ)	BC076650	-101.92	-101.92	1.000	0.0001	0.0001	-105.26	-101.92	0.010
Kif20A	63	BC077224	BC084922	EU441454 (XBβ)	CT485689	-156.81	-156.26	0.294	0.2873	>2	-157.15	-156.57	0.281
Kit	102	BC061947	Z48770	EU441455 (XBα)	-	-222.29	-221.80	0.322	0.319	0.0001	-222.73	-222.36	0.387
Kit	90	BC061947	Z48770	EU441456 (XBβ)	-	-246.14	-246.14	0.990	>2	>2	-249.28	-248.57	0.235
Kip2	24	X94082	BC071083	AUR (XBβ)	-	-47.14	-46.55	0.279	0.0001	0.0001	-48.14	-46.94	0.121
Krs1	99	BC047965	BC046578	EU441457 (XBβ)	BC067987	-195.41	-194.22	0.124	0.0001	0.2922	-199.34	-194.66	0.002
Krt18	66	BC054993	BC072305	EU441458 (XBα)	BC061366	-150.01	-150.01	1.000	0.0001	0.0001	-154.84	-150.01	0.002
Krt18	78	BC054993	BC072305	EU441459 (XBβ)	BC061366	-147.84	-147.03	0.204	>2	0.327	-147.69	-147.22	0.332
L1	954	X05216	X05217	EU441460 (XBα)	-	-1776.74	-1775.47	0.110	0.0788	0.0001	-1779.36	-1775.83	0.008
L30	159	BC073560	BC053758	EU441461 (XBα)	BC077047	-259.01	-259.01	1.000	0.0001	0.0001	-263.08	-259.01	0.004
LaminB	102	AF077838	X06344	EU441462 (XBβ)	-	-170.75	-170.75	1.000	0.0001	0.0001	-171.59	-170.90	0.239
LDH	42	U07179	U07176	AUR (XBα)	-	-69.70	-69.70	1.000	>2	>2	-69.87	-69.70	0.566
LEF1	1017	AF287147	AF287148	EU441463 (XBα)	-	-1881.14	-1881.09	0.747	0.1906	0.1367	-1882.86	-1881.32	0.079
LR	276	BC053822	BC078594	EU441464 (XBα)	-	-538.94	-538.52	0.360	0.0001	0.1615	-541.31	-538.59	0.020
Lupus	192	X68817	X68818	EU441465; EU441755 (XBβ)	-	-444.49	-443.95	0.299	0.0793	0.0001	-446.35	-445.15	0.121
Lyar	72	BC077174	BC073300	EU441756 (XBα)	-	-139.26	-139.24	0.829	0.7127	0.3633	-139.28	-139.28	0.924
Mad21	51	BC068714	BC045227	EU441466 (XBβ)	BX739935	-93.64	-93.64	1.000	0.0001	0.0001	-96.40	-93.64	0.019
Mak16l	72	BC043795	BC078464	EU441467 (XBα)	BC061599	-143.84	-142.41	0.091	0.4181	0.0001	-142.55	-142.41	0.593
MAP	261	BC044036	BC090193	EU441469; EU441757 (XBβ)	-	-459.02	-459.02	1.000	0.0001	0.0001	-459.37	-459.37	1.000
MAP	57	BC044036	BC090193	EU441468 (XBα)	-	-75.69	-75.69	1.000	>2	>2	-75.69	-75.69	1.000
Map1lc3a	72	BC056047	BC043946	EU441470 (XBβ)	BC080488	-162.83	-163.62	0.650	>2	0.0001	-163.62	-162.83	0.209
Map1lc3b	201	BC060359	BC094072	EU441471 (XBβ)	BC064267	-389.79	-389.55	0.483	0.322	0.087	-389.79	-389.55	0.490
MAP2K1IP1	189	BC056049	BC056026	EU441472; EU441758 (XBα)	CR760135	-393.96	-393.85	0.628	0.0325	0.0638	-408.47	-393.87	0.000
Mapkapk2	84	BC084300	BC070986	EU441473 (XBα)	CR761979	-144.43	-144.43	1.000	0.0001	0.0001	-144.71	-144.43	0.449
Mark3	171	AF509737	BC084772	EU441474; EU441759 (XBα)	-	-296.30	-296.30	1.000	0.0001	0.0001	-302.17	-296.30	0.001
Mars	102	BC077353	BC057757	EU441475 (XBβ)	-	-219.17	-217.71	0.087	0.1007	1.3672	-222.02	-219.51	0.025
MatG10	165	BC040971	BC084786	EU441476 (XBα)	-	-291.26	-291.26	1.000	0.0001	0.0001	-294.80	-291.26	0.008
Mbc2	60	BC089293	BC046701	EU441477 (XBβ)	-	-116.93	-116.08	0.193	0.2777	>2	-117.88	-117.46	0.358
Mcm4a	96	BC083031	BC072870	EU441478 (XBβ)	BC074670	-173.96	-173.96	1.000	0.0001	0.0001	-175.94	-173.96	0.047
Mcts1	78	BC070753	BC087444	EU441479 (XBα)	BC077034	-118.86	-118.58	0.454	0.235	0.0001	-119.10	-118.58	0.311
Meis3	39	BC084920	BC044024	AUR (XBβ)	BC075589	-67.30	-67.30	1.000	0.0001	0.0001	-69.33	-67.30	0.044
Metap1	54	BC054204	BC070567	EU441480 (XBβ)	BC088554	-99.53	-99.00	0.303	0.3832	0.0001	-99.74	-99.52	0.513
Mfn1	147	BC084774	BC072898	EU441481 (XBβ)	NM001016189	-348.89	-347.83	0.145	0.2856	0.0001	-349.76	-349.18	0.284
Mgat2	102	BC074371	BC068949	EU441482 (XBα)	BC075542	-209.52	-208.91	0.270	0.0001	0.2508	-211.57	-209.42	0.038

Table 2.3 (continued 5)

Mgat2	117	BC074371	BC068949	EU441483 (XBβ)	BC075542	-217.49	-217.49	1.000	0.0001	0.0001	-219.59	-218.12	0.087
MMP13	72	BC056040	BC046939	EU441484 (XBβ)	BC076908	-180.98	-180.97	0.992	>2	0.0001	-181.16	-180.98	0.538
Morc3	96	BC070772	BC077542	EU441485 (XBα)	-	-184.82	-183.96	0.190	0.2362	>2	-185.47	-185.01	0.335
Myogenin	561	AY046531	AY046532	EU441486 (XBβ)	-	-1039.42	-1038.40	0.153	0.0001	0.2687	-1042.23	-1039.75	0.026
Myoz2	144	BC042308	BC082908	EU441487; EU441760 (XBα)	BC087984	-277.51	-277.09	0.354	0.588	>2	-277.96	-277.80	0.572
Nap1	240	BC043903	DQ020268	EU441488; EU441761; EU441762 (XBα)	BC068215	-435.52	-435.43	0.663	0.1178	0.2374	-436.81	-435.45	0.100
NCAM	177	M25696	M76710	EU441490; EU441763 (XBβ)	-	-368.75	-367.42	0.104	0.0001	0.3346	-370.11	-367.43	0.021
NCAM	138	M25696	M76710	EU441489 (XBα)	-	-254.67	-254.67	1.000	0.0001	0.0001	-257.82	-255.15	0.021
Ndufa10	315	BC084366	BC068897	EU441491; EU441764; EU441765 (XBα)	BC087982	-615.41	-614.55	0.188	>2	0.1316	-614.60	-614.55	0.748
Ndufa10	213	BC084366	BC068897	EU441492; EU441766 (XBβ)	BC087982	-384.03	-384.03	0.998	0.3639	0.0001	-384.43	-384.03	0.368
Nfe2l2	363	BC074462	BC043997	EU441493; EU441767; EU441768 (XBβ)	BC079927	-838.74	-838.42	0.430	0.3886	0.1817	-839.44	-838.59	0.191
NMHCA	204	BC057729	AF055895	EU441494; EU441769 (XBα)	-	-429.96	-429.96	1.000	0.0001	0.0001	-431.66	-429.96	0.065
NMHCA	201	BC057729	AF055895	EU441495; EU441770 (XBβ)	-	-399.88	-399.88	1.000	0.0001	0.0001	-402.82	-399.89	0.016
NO38	780	X05496	X56039	EU441497 (XBβ)	-	-1643.80	-1641.93	0.053	0.9815	0.1737	-1642.04	-1642.03	0.971
NO38	93	X05496	X56039	EU441496 (XBα)	-	-211.90	-211.76	0.601	0.0127	0.0419	-214.08	-211.77	0.032
Npal3	78	BC047987	BC100197	EU441499 (XBβ)	NM001016927	-177.00	-177.00	1.000	>2	>2	-178.06	-177.77	0.448
Npal3	309	BC047987	BC100197	EU441498; EU441771 (XBα)	NM001016927	-641.29	-640.10	0.122	0.0001	0.1666	-644.17	-640.10	0.004
Npc2	252	BC060392	BC068868	EU441500; EU441772 (XBα)	BC080500	-547.73	-547.70	0.805	0.2355	0.1569	-548.62	-548.00	0.267
Ntf2	342	BC054170	BC073371	EU441501 (XBβ)	BC084526	-650.38	-650.30	0.699	0.1342	0.0663	-650.92	-650.33	0.275
Nudc	27	BC082700	BC070681	AUR (XBα)	CT010535	-38.93	-38.93	1.000	0.0001	0.0001	-40.02	-38.93	0.138
Nup88	291	AJ617672	AJ617673	EU441502; EU441773; EU441774 (XBβ)	-	-616.11	-616.07	0.767	0.1272	0.1749	-619.21	-616.07	0.012
OncogeneCets1	96	X52692	BC075161	EU441503 (XBβ)	-	-168.79	-168.79	1.000	0.0001	0.0001	-169.59	-168.79	0.206
P55pik	453	BC081193	BC077814	EU441504; EU441775; EU441776; EU441777; EU441778 (XBβ)	CR760256	-789.06	-788.35	0.233	0.1018	0.3868	-794.09	-790.48	0.007
Pa2g4	36	BC084760	BC073401	AUR (XBβ)	BC080337	-36.19	-36.19	0.996	0.0001	0.0001	-38.24	-36.19	0.043
Pabpc1	78	BC052100	BC072110	EU441505 (XBα)	BC076931	-157.34	-157.34	1.000	0.0001	0.0001	-158.27	-158.11	0.562
PcnaA	192	BC041549	BC057758	EU441506; EU441779 (XBα)	BC080365	-301.33	-301.33	1.000	0.0001	0.0001	-305.34	-301.33	0.005
Pdhe1a	165	BC080995	BC077220	EU441507; EU441780 (XBβ)	-	-288.81	-288.41	0.369	0.1398	0.0001	-289.69	-288.52	0.126
Phf10	318	BC077418	BC082869	EU441508; EU441781; EU441782 (XBβ)	-	-580.09	-578.58	0.082	0.0001	0.257	-583.88	-578.59	0.001
Pik4ca	66	BC077604	BC076796	EU441509 (XBα)	-	-104.83	-104.83	1.000	0.0001	0.0001	-105.75	-104.83	0.175
Pkm2	21	BC044007	BC079921	AUR (XBα)	CR760921	-29.38	-29.38	1.000	0.0001	0.0001	-30.76	-29.38	0.097
Pmp2	351	BC056855	BC072965	EU441510; EU441783 (XBα)	BC090106	-723.11	-721.90	0.120	0.9125	0.0677	-722.64	-722.64	0.986
Pmpcb	303	BC088718	BC072067	EU441511; EU441784; EU441785 (XBβ)	CR942396	-586.02	-585.89	0.612	0.0437	0.0001	-591.41	-586.08	0.001
Pomc	636	X03843	X03844	EU441512 (XBα)	-	-1354.09	-1354.06	0.794	0.1202	0.1693	-1356.17	-1355.21	0.165
Pphln1	60	BC084344	BC075150	EU441513 (XBβ)	BC061358	-130.20	-129.83	0.394	>2	0.6748	-130.17	-129.85	0.424
Ppib	102	BC084369	BC054168	EU441514 (XBβ)	CN103294	-203.46	-203.46	1.000	>2	>2	-204.40	-204.09	0.431
Ppp1cc	357	BC054188	BC090213	EU441515; EU441786; EU441787; EU441788 (XBα)	BC067911	-578.08	-578.08	1.000	0.0001	0.0001	-582.25	-578.20	0.004
Ppp1cc	162	BC054188	BC090213	EU441516; EU441789; AUR (XBβ)	BC067911	-289.47	-289.47	1.000	0.0001	0.0001	-289.70	-289.47	0.504
Ppp1r3c	93	BC092028	BC068825	EU441517 (XBβ)	-	-146.35	-146.35	1.000	0.0001	0.0001	-147.88	-146.35	0.080
Ppp2B	150	BC043624	BC073612	EU441518; EU441790 (XBβ)	-	-275.55	-275.55	1.000	0.0001	0.0001	-280.42	-275.65	0.002
Ppp4r2	42	BC077952	BC088694	AUR (XBβ)	-	-72.74	-72.74	1.000	0.0001	0.0001	-74.69	-73.31	0.097
Prdx1	372	BC092102	BC072833	EU441519; EU441791 (XBβ)	BC084184	-675.88	-675.19	0.240	0.0001	0.0567	-684.54	-676.14	0.000
Prdx4	201	BC087512	BC073532	EU441520 (XBα)	BC076692	-364.11	-364.11	1.000	0.0001	0.0001	-367.25	-364.79	0.027
Prdx4	270	BC087512	BC073532	EU441521; EU441792; EU441793 (XBβ)	BC076692	-467.99	-467.99	1.000	0.0001	0.0001	-478.29	-468.11	0.000
Prl	159	BC092151	BC075216	EU441522 (XBα)	CX852073	-364.86	-363.27	0.074	0.0001	0.7106	-366.97	-365.08	0.052

Table 2.3 (continued 6)

Prothymosin	231	BC054174	BC045213	EU441523 (XBα)	-	-393.79	-393.70	0.666	0.0791	0.3136	-394.51	-394.22	0.442
Protocadherin	741	BC047955	BC072235	EU441524; EU441794; EU441795; EU441796; EU441797 (XBβ)	-	-1647.59	-1647.21	0.389	0.1962	0.4018	-1650.57	-1648.30	0.033
Psap	627	BC054988	BC080997	EU441525; EU441798; EU441799 (XBα)	-	-1453.76	-1453.76	0.942	0.2352	0.2497	-1456.40	-1453.90	0.026
Psma3	33	BC041518	BC058201	EU441526 (XBα)	BC087567	-30.81	-30.81	1.000	0.0001	0.0001	-30.81	-30.81	0.998
Psma3	150	BC041518	BC058201	EU441527 (XBβ)	BC087567	-219.71	-218.87	0.194	0.0001	0.418	-223.18	-221.47	0.064
Psma7	72	BC074225	BC084072	EU441800 (XBβ)	BC061282	-128.71	-128.71	0.994	0.0001	0.0001	-131.98	-128.71	0.011
Psmb2	240	BC072908	BC070836	EU441528; EU441801 (XBβ)	BC084185	-457.21	-456.83	0.382	0.072	0.0001	-458.60	-456.92	0.067
Psmd12	195	BC079690	BC070583	EU441529; EU441802 (XBα)	NM001016001	-432.83	-432.64	0.536	>2	0.2556	-433.34	-433.27	0.713
Psmd14	303	BC073436	BC045094	EU441530; EU441803; EU441804 (XBβ)	BC091596	-521.83	-521.83	1.000	0.0001	0.0001	-526.34	-522.38	0.005
Psmd9	132	BC041532	BC074472	EU441531 (XBβ)	BX723168	-255.48	-253.67	0.057	>2	0.0001	-254.41	-253.90	0.310
Psme1	216	BC082461	BC073340	EU441532; EU441805; AUR (XBα)	BC088020	-459.78	-459.78	0.973	0.2102	0.2229	-460.92	-460.55	0.390
Psme1	267	BC082461	BC073340	EU441533; EU441806; EU441807; AUR (XBβ)	BC088020	-550.71	-546.17	0.003	>2	0.0001	-547.11	-546.87	0.485
PSME3	93	BC068852	BC082675	EU441534 (XBα)	BC090116	-161.61	-159.81	0.058	>2	0.0001	-160.28	-159.81	0.332
PtdInsTP	108	BC072371	BC077831	EU441535 (XBα)	-	-200.21	-200.21	1.000	0.0001	0.0001	-201.82	-200.21	0.072
Ptp4a1	234	AB127963	BC077539	EU441536 (XBβ)	BC080902	-393.87	-393.14	0.226	0.2876	0.0001	-393.67	-393.36	0.433
Pyp	54	BC070619	BC097793	EU441537 (XBβ)	NM001017130	-110.27	-110.27	0.996	0.0707	0.0001	-111.65	-110.27	0.096
Rab11b	93	BC082421	BC041250	EU441538 (XBβ)	BC084173	-178.58	-178.58	1.000	0.0001	0.0001	-179.24	-178.79	0.343
Rab27a	132	BC082420	BC073442	EU441539 (XBα)	BC080481	-203.81	-203.81	1.000	0.0001	0.0001	-205.04	-203.81	0.117
Rabin3	117	BC077516	BC054310	EU441540; AUR (XBβ)	BC061285	-218.39	-216.68	0.065	>2	0.0001	-217.72	-217.30	0.362
Rabl2b	213	BC077550	BC074476	EU441541; EU441808 (XBα)	-	-396.41	-396.11	0.437	0.0691	0.2423	-399.23	-396.57	0.021
Racgap1	237	BC070771	BC046676	EU441542; EU441809; EU441810 (XBα)	BC067994	-522.72	-522.57	0.588	0.7296	0.2998	-522.89	-522.86	0.797
Racgap1	105	BC070771	BC046676	EU441543; AUR (XBβ)	BC067994	-193.75	-193.75	0.997	0.3398	0.0001	-193.97	-193.75	0.508
RAG1	1134	AY874341	AY874315	EF535912 (XBβ)	-	-2332.46	-2332.44	0.872	0.1388	0.1571	-2338.19	-2333.06	0.001
Ranbp1	321	Y09128	BC054182	EU441544; EU441811; EU441812 (XBα)	BC061426	-534.17	-534.08	0.680	0.2551	0.1128	-534.63	-534.35	0.453
RAP1A	177	BC073286	BC078112	EU441545 (XBβ)	-	-287.48	-287.48	1.000	0.0001	0.0001	-290.16	-287.48	0.021
Rap1Gds1	75	BC079775	BC087423	EU441546 (XBβ)	-	-146.30	-146.30	1.000	0.0001	0.0001	-147.55	-146.30	0.114
Rassf6	129	BC043762	BC084799	EU441813; AUR (XBα)	BC075422	-262.81	-261.94	0.188	0.3043	0.0001	-262.89	-262.87	0.841
Rassf6	210	BC043762	BC084799	EU441547; EU441814 (XBβ)	BC075422	-368.31	-367.75	0.291	0.0001	0.0836	-369.62	-367.79	0.055
Rbm12b	102	BC043858	BC060345	EU441548 (XBβ)	CR584843	-230.06	-229.30	0.220	>2	0.1015	-229.55	-229.40	0.575
Rcbtb1	93	BC068656	BC068834	EU441549 (XBα)	NM001017352	-226.51	-226.46	0.750	0.0521	0.0001	-227.07	-226.47	0.270
Rcn2	354	BC077885	BC072037	EU441550; EU441815; EU441816; EU441817 (XBα)	CR848220	-792.22	-792.22	0.924	0.5665	0.5169	-793.95	-793.69	0.470
Rfc5	87	BC072889	BC044712	EU441551 (XBβ)	BC084510	-146.22	-146.22	1.000	0.0001	0.0001	-147.33	-146.22	0.136
RGs2	222	BC084852	BC108887	EU441552; EU441818 (XBα)	NM001030451	-486.70	-486.22	0.330	0.9321	0.1582	-486.24	-486.24	0.976
RIoK2	78	BC077472	BC084165	EU441554 (XBβ)	-	-156.75	-155.07	0.066	>2	0.0001	-155.70	-155.55	0.578
RIoK2	225	BC077472	BC084165	EU441553; EU441819; EU441820 (XBα)	-	-439.94	-438.07	0.053	0.0001	0.2227	-445.35	-438.85	0.000
RPL34	63	BC099259	BC078541	EU441821 (XBα)	CT025306	-108.36	-108.36	1.000	0.0001	0.0001	-109.59	-108.36	0.116
Rpl5	204	BC042258	BC041227	EU441555; EU441822 (XBα)	BC059751	-421.68	-420.11	0.076	0.061	1.6989	-424.18	-421.59	0.023
Rpl5	87	BC042258	BC041227	EU441556; AUR (XBβ)	BC059751	-144.67	-144.43	0.482	>2	0.0001	-144.76	-144.57	0.536
RPN1	69	BC045212	BC044106	EU441823 (XBα)	BC064882	-120.20	-119.69	0.313	0.2403	0.0001	-120.10	-119.69	0.365
RpS12	111	BC044028	BC056655	EU441824 (XBα)	BC080154	-200.96	-200.96	1.000	0.0001	0.0001	-203.87	-201.06	0.018
Rps28	93	BC099294	BC078605	EU441557 (XBβ)	NM001016950	-145.05	-145.05	1.000	0.0001	0.0001	-146.23	-145.05	0.125
Rps8.1	87	BC041282	BC041307	EU441558 (XBα)	BC077028	-157.16	-157.16	1.000	0.0001	0.0001	-158.51	-157.28	0.117
Rrbp1	201	BC059298	BC073565	EU441559; EU441825 (XBα)	BC074706	-452.04	-451.88	0.566	0.2608	0.5035	-453.45	-452.16	0.108
Rrbp1	210	BC059298	BC073565	EU441560; EU441826; EU441827 (XBβ)	BC074706	-457.24	-456.76	0.324	0.998	0.1919	-456.76	-456.76	1.000
Rwdd1	327	BC080086	BC073326, BC044708	EU441561; EU441828; EU441829; EU441830 (XBα)	-	-688.14	-688.13	0.902	0.2474	0.288	-689.45	-688.26	0.122

Table 2.3 (continued 7)

Rwdd1	102	BC080086	BC073326, BC044708	EU441562; AUR (XBβ)	-	-207.62	-207.43	0.536	0.0001	0.0902	-207.89	-207.76	0.609
Saps3	387	BC070802	BC070770	EU441564; EU441833; EU441834; AUR (XBβ)	CX430880	-772.73	-772.00	0.228	0.4378	0.0927	-772.35	-772.07	0.455
Saps3	309	BC070802	BC070770	EU441563; EU441831; EU441832 (XBα)	CX430880	-576.55	-576.15	0.373	0.1566	0.5075	-580.32	-578.52	0.058
Scap2	201	BC084386	BC084787	EU441565; EU441835 (XBα)	BC089739	-418.55	-413.42	0.001	>2	0.0001	-416.65	-415.98	0.248
Scarb2	96	BC084242	BC045028	EU441566 (XBα)	CR760704	-215.26	-214.50	0.217	>2	0.146	-215.26	-215.03	0.498
Scarb2	153	BC084242	BC045028	EU441567; EU441836 (XBβ)	CR760704	-320.45	-320.45	1.000	0.0001	0.1143	-320.52	-320.52	0.999
Scc2	159	AY661732	AY661733	EU441568; EU441837 (XBβ)	-	-320.62	-320.10	0.307	0.0001	0.3714	-321.32	-320.11	0.119
Scrp1	90	BC075175	BC084248	EU441569 (XBα)	BC064859	-161.17	-160.31	0.191	0.7885	0.0001	-160.90	-160.88	0.851
Sdcbp	63	BC081003	BC043629	EU441838 (XBβ)	BC076670	-96.75	-96.75	1.000	0.0001	0.0001	-99.47	-96.75	0.020
Sec61g	90	BC078558	BC097730	EU441570 (XBα)	-	-135.19	-135.19	1.000	0.0001	0.0001	-136.82	-135.19	0.072
Sec61g	66	BC078558	BC097730	EU441571 (XBβ)	-	-101.22	-101.22	1.000	0.0001	0.0001	-101.96	-101.22	0.222
Sec63	585	BC072820	BC110927	EU441572; EU441839; EU441840; EU441841 (XBα)	-	-1045.47	-1045.00	0.333	0.0586	0.2131	-1049.25	-1045.56	0.007
SenP8	195	BC059353	BC089291	EU441573; EU441842 (XBα)	CT030060	-425.90	-424.76	0.130	>2	0.1056	-425.44	-425.19	0.480
SenP8	267	BC059353	BC089291	EU441574; EU441843 (XBβ)	CT030060	-529.84	-529.77	0.707	0.1117	0.1988	-531.38	-529.78	0.073
Septin11	342	BC077941	BC073250	EU441575; EU441845; EU441846; EU441847 (XBβ)	-	-598.52	-598.52	0.971	0.0737	0.0781	-601.88	-598.52	0.010
Septin11	84	BC077941	BC073250	EU441844; AUR (XBα)	-	-138.46	-138.46	1.000	0.0001	0.0001	-142.39	-138.46	0.005
SerpinI1	561	BC074366	AM050698	EU441576; EU441848; EU441849; EU441850 (XBα)	-	-1038.08	-1038.08	0.970	0.1725	0.1833	-1039.02	-1038.50	0.305
SerpinI1	381	BC074366	AM050698	EU441577; EU441851; EU441852; EU441853 (XBβ)	-	-731.22	-729.92	0.108	0.0967	0.0001	-735.86	-731.77	0.004
SGK	96	BC073077	BC074305	EU441854 (XBα)	-	-224.27	-224.27	1.000	0.0001	0.0001	-224.30	-224.27	0.807
SH3p4	57	BC097775	BC088909	EU441578 (XBβ)	-	-78.55	-78.55	1.000	0.0001	0.0001	-79.62	-78.55	0.142
Shcbp1	243	BC082676	BC073190	EU441579; EU441855 (XBα)	CR761438	-509.66	-509.28	0.382	0.2391	1.2713	-510.15	-509.74	0.363
Siae	84	BC059983	BC084103	EU441580 (XBβ)	-	-148.29	-148.29	1.000	0.0001	0.0001	-148.40	-148.29	0.652
Siah	246	BC046706	BC074445	EU441581; EU441856 (XBα)	-	-502.00	-501.76	0.483	0.6227	0.207	-501.81	-501.76	0.753
Ski	81	X68683	U89999	EU441582 (XBα)	-	-205.53	-205.36	0.560	0.0001	0.2456	-205.57	-205.37	0.527
Slc30a1	60	BC100176	BC046675	EU441583 (XBα)	BC080447	-149.27	-148.25	0.154	>2	0.2353	-149.14	-148.53	0.270
Slc30a6	504	BC072974	BC077520	EU441584; EU441857; EU441858; EU441859; EU441860 (XBβ)	BC088526	-985.97	-985.61	0.392	0.0908	0.2636	-990.22	-985.76	0.003
Snap29	69	BC099298	BC041208	EU441585 (XBβ)	BC061308	-134.29	-133.41	0.185	0.5286	>2	-136.46	-136.24	0.509
SNT	42	AF390895	BC046943	AUR (XBβ)	-	-94.21	-92.62	0.075	1.1677	0.0001	-94.56	-94.56	0.904
SP22	483	AF394958	DQ406635	EU441586 (XBα)	-	-1009.17	-1009.10	0.701	0.0687	0.0369	-1011.55	-1010.00	0.078
Spats2	816	BC077749	BC057748	EU441587 (XBα)	BC063897	-1733.89	-1732.92	0.164	1.0732	0.2811	-1734.70	-1734.70	0.966
Spc6a	186	BC087304	BC085062	EU441588 (XBα)	CX509398	-375.75	-374.70	0.148	0.3535	0.0273	-376.13	-375.96	0.559
SPNR	885	BC078118	BC089298	EU441589 (XBα)	-	-1881.92	-1881.20	0.230	0.7141	0.2038	-1881.25	-1881.20	0.755
Ssr1	90	BC077979	BC056853	EU441590 (XBβ)	BC068212	-140.43	-140.43	1.000	0.0001	0.0001	-141.74	-140.43	0.105
Stat3	51	BC044717	AB017701	EU441861 (XBα)	-	-112.18	-112.18	0.996	1.3374	0.0001	-112.20	-112.18	0.825
STI1	96	BC046709	BC078016	EU441591 (XBα)	-	-158.02	-157.21	0.202	0.0001	0.3353	-159.65	-157.22	0.028
Stmn1	417	BC073451	BC054159	EU441862 (XBβ)	BC080482	-784.27	-784.25	0.849	0.1375	0.0967	-785.06	-784.29	0.214
Sumo	123	Z97073	BC090210	EU441863; EU441864; AUR (XBβ)	-	-210.96	-210.96	1.000	0.0001	0.0001	-213.52	-211.26	0.033
Survivin2	111	AB197249	BC089271	EU441592; EU441865 (XBα)	CR848111	-224.78	-224.78	1.000	0.0001	0.0001	-227.38	-224.98	0.028
Suv420	78	BC086459	BC073331	EU441593 (XBα)	CR760868	-160.71	-160.71	1.000	>2	>2	-161.29	-160.82	0.333
Suv420	183	BC086459	BC073331	EU441594 (XBβ)	CR760868	-347.71	-346.97	0.224	>2	0.22	-347.99	-347.92	0.715
Syap1	78	BC043829	BC071003	EU441596 (XBβ)	BC064154	-138.67	-138.17	0.318	0.0001	0.3986	-139.35	-138.17	0.125
Syap1	75	BC043829	BC071003	EU441595 (XBα)	BC064154	-163.69	-163.69	0.994	0.1608	0.0001	-165.49	-163.69	0.058
Tac1	102	BC092125	BC092325	EU441597 (XBα)	-	-189.80	-189.80	1.000	0.0001	0.0001	-190.65	-189.80	0.192
Tagln	132	BC061650	BC084848	EU441598 (XBβ)	BC091023	-232.29	-232.29	1.000	0.0001	0.0001	-233.59	-232.88	0.234
TaoK1	192	BC068781	BC043764	EU441599; EU441866 (XBα)	-	-378.57	-378.57	0.993	0.3544	0.0001	-378.76	-378.57	0.543

Table 2.3 (continued 8)

TB1	81	BC072218	BC082843	EU441600 (XBβ)	NM001016130	-158.07	-157.64	0.352	0.1909	0.0001	-158.56	-158.00	0.288
TCP1	216	BC044673	BC068901	EU441601; EU441867 (XBα)	-	-380.25	-380.10	0.583	0.0001	0.0995	-380.77	-380.44	0.419
TEF	54	BC054981	BC082861	EU441602 (XBβ)	NM001017319	-114.89	-114.40	0.326	>2	0.3799	-114.98	-114.69	0.447
Tegt	63	BC047131	BC079707	EU441603 (XBβ)	-	-121.64	-121.18	0.340	0.518	0.0001	-121.59	-121.49	0.659
Tekt3	75	BC088703	BC056029	EU441604 (XBα)	CR760902	-162.10	-162.10	1.000	>2	>2	-163.43	-163.06	0.389
TfElys	99	BC086281	BC076775	EU441605 (XBβ)	-	-275.25	-274.78	0.335	>2	0.355	-275.29	-275.06	0.501
TFIIAg	210	BC088977	BC072888	EU441868 (XBα)	BC077041	-354.71	-354.71	1.000	0.0001	0.0001	-358.54	-354.71	0.006
TFIIAg	51	BC088977	BC072888	EU441606 (XBβ)	BC077041	-76.24	-76.24	1.000	0.0001	0.0001	-77.40	-76.24	0.128
TFIIS	33	BC070555	BC077765	AUR (XBβ)	BC059769	-53.05	-51.49	0.077	0.0001	>2	-54.56	-53.32	0.116
Tgm2	123	BC072304	BC056053	EU441607 (XBα)	-	-239.63	-239.58	0.753	1.2734	0.5116	-240.81	-240.77	0.786
Thap1	96	BC077429	BC084824	EU441608; AUR (XBα)	BC075499	-173.42	-173.29	0.599	0.5344	1.332	-174.34	-174.23	0.642
Thap1	465	BC077429	BC084824	EU441609; EU441869; EU441870 (XBβ)	BC075499	-860.10	-859.83	0.465	0.1793	0.0619	-860.93	-859.83	0.139
TIAR	123	AJ416631	BC045086	EU441610 (XBβ)	-	-191.79	-190.94	0.192	0.3967	0.0001	-191.18	-190.94	0.488
Tig1	114	BC059975	BC073285	EU441611; AUR (XBα)	-	-206.92	-206.92	0.995	0.3691	0.0001	-207.01	-206.92	0.674
Timm10	48	BC097518	BC072896	EU441612 (XBα)	BC064213	-109.86	-109.86	1.000	0.0001	0.0001	-112.64	-110.08	0.024
Tmem59	294	BC073604	BC084373	EU441613; EU441871; EU441872 (XBβ)	-	-629.22	-628.00	0.119	0.052	0.3817	-633.00	-628.13	0.002
Tmplt	117	BC077325	BC090205	EU441614; EU441873 (XBα)	BC082959	-224.63	-223.54	0.139	1.093	0.0001	-224.03	-224.02	0.944
Tnpo2	90	BC084097	BC084978	EU441615 (XBα)	-	-194.26	-194.26	1.000	0.0001	0.0001	-195.69	-194.26	0.091
Tprkb	519	BC077299	BC045235	EU441616 (XBα)	BC064176	-951.66	-951.26	0.371	0.0534	0.2309	-953.74	-951.26	0.026
Trap	75	BC041301	BC053768	EU441617 (XBα)	BC077002	-135.50	-135.25	0.480	0.0001	>2	-135.50	-135.25	0.477
Trap	66	BC041301	BC053768	EU441618 (XBβ)	BC077002	-138.61	-138.44	0.557	0.1129	0.3454	-140.20	-138.46	0.062
TshA	360	L07619	BC072372	EU441619 (XBβ)	-	-670.32	-668.76	0.078	0.0001	0.3197	-672.82	-668.96	0.005
Tubb2c	84	BC043974	BC054297	EU441620 (XBα)	-	-120.88	-120.88	1.000	0.0001	0.0001	-126.18	-120.88	0.001
Txndc10	87	BC092019	BC073549	EU441621 (XBβ)	-	-155.79	-154.94	0.192	0.265	0.0001	-156.18	-155.89	0.447
U1	84	AF441126	X93491	EU441622 (XBα)	-	-186.44	-186.23	0.513	0.1708	0.5387	-188.25	-187.24	0.156
Ube2c	186	BC075141	BC088818	EU441623; EU441874 (XBβ)	AL679209	-349.49	-349.49	1.000	0.0001	0.0001	-351.05	-349.79	0.113
Ube2d1	96	BC076728	BC084849	EU441624 (XBβ)	CN114651	-137.90	-137.90	1.000	0.0001	0.0001	-139.55	-137.90	0.069
Ube2e	69	BC077923	BC077801	EU441625 (XBα)	-	-134.42	-133.85	0.288	0.0001	0.3003	-135.53	-134.31	0.118
UDPgcg	96	AY112732	BC084966	EU441626 (XBα)	-	-146.84	-146.84	1.000	0.0001	0.0001	-147.27	-146.84	0.355
Ufd1l	222	BC072284	BC061930	EU441627; EU441875 (XBα)	BC076680	-435.42	-433.88	0.080	0.31	0.0001	-435.90	-434.96	0.170
Unc50	102	BC084865	BC071145	EU441628 (XBα)	-	-177.37	-177.37	0.996	0.0001	>2	-178.13	-177.37	0.217
vATPase	93	BC063729	BC060343	EU441629 (XBβ)	BC075390	-159.77	-159.77	1.000	0.0001	0.0001	-163.10	-159.77	0.010
Vps29	429	BC097520	BC073281	EU441630; EU441876 (XBα)	BC077001	-739.10	-739.10	1.000	0.0001	0.0001	-745.49	-739.37	0.000
Vps29	63	BC097520	BC073281	EU441631 (XBβ)	BC077001	-87.97	-87.97	1.000	0.0001	>2	-87.97	-87.97	1.000
WARS	426	BC046713	BC068695	EU441632; EU441877; EU441878; EU441879 (XBα)	-	-923.35	-922.61	0.223	0.118	0.5878	-926.28	-923.98	0.032
XAC	78	U26270	U26269	EU441633 (XBβ)	-	-156.28	-154.48	0.058	>2	0.0001	-155.13	-154.78	0.403
Xap5	72	BC041272	BC098980	EU441634 (XBβ)	CR942792	-121.41	-121.41	1.000	0.0001	0.0001	-124.28	-121.41	0.017
Xmegs	588	XXXX	AB066589	EU441635 (XBα)	-	-1484.73	-1484.72	0.926	0.405	0.38	-1486.35	-1484.82	0.080
Xtp5	57	BC088719	BC076757	EU441636 (XBα)	-	-136.03	-135.56	0.337	>2	0.2862	-135.80	-135.58	0.509
Ywhab	84	BC041526	BC084055	EU441637 (XBβ)	BC084514	-164.83	-164.83	1.000	0.0001	0.0001	-165.87	-164.83	0.149
Znf593	357	BC084361	BC073548	EU441638; EU441880 (XBβ)	NM001016550	-699.99	-699.97	0.853	0.1031	0.145	-701.12	-700.07	0.147

EPα, NTα, EPβ or NTβ, and ST; models are tested on the phylogeny depicted in Fig. 1C.

Gene	Length (bp)	EPα	NTα	EPβ or NTβ	ST	Ho	Ha	P-value	Ka/Ks early	Ka/Ks late	-lnL Ho	-lnL Ha	P value
Alpha	252	EU446223	EU446224	EU446225 (EPβ)	-	-356.24	-356.24	1.00	0.97	0.00	-356.24	-356.24	1.000
BOIP	864	EU446226	EU446227	EU446228 (EPβ)	-	-1443.71	-1443.44	0.60	0.18	0.53	-1444.41	-1443.60	0.202
Calnexin	1314	EU446229	EU446230	EU446231 (NTβ)	-	-2036.77	-2036.22	0.46	0.00	0.11	-2052.65	-2049.30	0.010
Calreticulin	759	EU446232	EU446233	EU446234 (NTβ)	-	-1176.60	-1174.40	0.14	0.00	0.83	-1179.73	-1176.41	0.010
Cdc2	552	EU446235	EU446236	EU446237 (EPβ)	-	-833.26	-833.26	1.00	0.00	0.00	-834.98	-833.93	0.148
DAZ	741	EU446238	EU446239	EU446240 (EPβ)	-	-1259.84	-1259.84	1.00	0.28	0.00	-1260.54	-1259.84	0.237
Hey	342	EU446241	EU446242	EU446243 (EPβ)	-	-514.51	-514.51	1.00	0.00	0.00	-517.17	-514.82	0.030
p450	714	EU446244	EU446245	EU446246 (NTβ)	-	-1075.26	-1074.94	0.57	0.17	0.00	-1076.74	-1075.66	0.140
Pomc	636	EU446247	EU446248	EU446249 (EPβ)	-	-997.59	-997.59	1.00	>2	0.01	-997.83	-997.59	0.491
RAG1	1134	AY874308	AY874313	AY874312 (NTβ)	-	-1828.94	-1828.63	0.58	0.13	0.35	-1832.58	-1829.33	0.011
RAG2	1005	EF535959	EF535960	EF535963 (NTβ)	-	-1584.87	-1584.23	0.43	0.08	0.39	-1588.08	-1584.23	0.006
Spats2	816	EU446250	EU446251	EU446252 (EPβ)	-	-1305.32	-1303.77	0.21	0.12	>2	-1307.39	-1305.54	0.054
Xmegs	588	EU446253	EU446254	EU446255 (EPβ)	-	-987.43	-986.01	0.23	0.26	>2	-988.10	-987.03	0.143

XLα and XBα, XGα or XMα, XLβ; models are tested on the phylogeny depicted in Fig. 1D.

Gene	Length (bp)	XLα	XBα	XGα or XMα	XLβ	Ho	Ha	P-value	Ka/Ks early	Ka/Ks late	-lnL Ho	-lnL Ha	P value
Activin	795	U49914	EU441284	EU446256 (XMα)	BC077763	-1349.58	-1349.53	0.826	0.019	0.000	-1366.63	-1349.65	0.000
Alpha	252	BC073294	M32455	EU446257 (XMα)	BC075196	-473.88	-473.88	0.997	>2	>2	-475.58	-475.51	0.699
Beta	372	V01433	M32456	EU446258 (XMα)	BC071139	-764.14	-763.94	0.658	0.209	>2	-765.90	-764.06	0.055
Cdc2	552	M60681	EU441349	EU446259 (XMα)	BC045078	-1022.58	-1022.12	0.336	0.047	>2	-1031.35	-1022.52	0.000
L1	675	X05216	EU441460	EU446260 (XMα)	X05217	-1178.69	-1178.69	1.000	0.000	>2	-1191.80	-1178.91	0.000
LR	276	BC053822	EU441464	EU446261 (XMα)	BC078594	-491.85	-490.72	0.133	0.477	>2	-490.96	-490.96	0.932
Pomc	636	X03843	EU441512	EU446262 (XMα)	X03844	-1198.60	-1195.57	0.014	0.000	>2	-1206.57	-1196.07	0.000
RAG1	2820	EF535887; EF535914	EF535886; EF535912	EF535888; EF535915 (XGα)	L19324	-5194.18	-5194.18	0.937	0.124	>2	-5229.56	-5222.37	0.000
Xmegs	588	EU446264	EU441635	EU446263 (XGα)	AB0666589	-1309.85	-1309.42	0.352	0.249	>2	-1319.28	-1316.51	0.019

CHAPTER 3

Gene expression patterns and evolutionary rates influence functional persistence of paralogs generated by whole genome duplication in clawed frogs (*Xenopus*)

PREFACE

To better evaluate the impacts of proposed duplicate gene mechanisms that act on expression, we deployed one of the largest datasets of same-age duplicate genes ever used to look at their molecular evolution and expression characteristics before and after polyploidization. A diploid outgroup served as a surrogate for ancestral function and expression, and allowed us to estimate the relevance of regulatory changes in the retention of duplicate genes. We also analyzed data from genes that are found as single copies in the polyploid species, permitting us to test whether different types of genes are preferentially retained as duplicates after whole genome duplication.

ABSTRACT

Gene duplication can lead to altered function and expression, reduction of pleiotropy, and is a fundamental agent of biological innovation. With an aim to better understand genetic mechanisms that promote persistence of duplicate genes (paralogs), we compared molecular evolution and expression of thousands of pairs of paralogs and singleton genes (where a paralog once was present but became non-functional) in the tetraploid clawed frog *Xenopus laevis*. To estimate the rate of evolution and gene expression patterns before duplication, we also examined orthologs of these genes from a diploid species *Xenopus (Silurana) tropicalis*. These data indicate that genes expressed in more types of tissues, that is with "broader" expression patterns, are more likely to be retained as duplicate copies. After controlling for breadth of expression, genes with higher levels of expression are more likely to be retained as duplicates. Many duplicate copies undergo expression divergence in that they are expressed at different levels and in different tissues from each other and from their singleton ortholog, often times at lower expression levels and in less tissue types after duplication. At the sequence level, genes that evolve slowly are more likely to persist as functional duplicates. Of these variables, expression breadth has the greatest influence on the odds of duplicate gene persistence and on the rate of evolution, suggesting a key role of regulatory mechanisms for duplicate gene persistence. Consistent with the regulatory subfunctionalization mechanism, genes that are broadly expressed are more likely to display complementary loss of expression. In addition, many duplicate genes have gained new expression domains relative to each other and to their ortholog, which is consistent with the regulatory neofunctionalization mechanism. These results provide support for pervasive spatial subfunctionalization and quantitative

expression divergence in addition to the acquisition of novel expression patterns, highlighting the central role of these mechanisms in transcriptome remodeling after whole genome duplication. We speculate that these same processes may have been significant during early evolution of other duplicated genomes, such as that of the ancient ancestor of jawed vertebrates.

INTRODUCTION

Gene duplication is an important catalyst for biological innovation by allowing new functions to evolve (neofunctionalization) (Ohno 1970), by "division of labor" via complementary degradation of function (subfunctionalization) (Force et al. 1999), and by redundancy that, for example, changes phenotypes affected by genes with dosage-dependent regulation (Galitski et al. 1999; Osborn et al. 2003). Neofunctionalization (NF) and subfunctionalization (SF) occur in protein function and in regulation of gene expression, and when both mechanisms operate after duplication, the resulting paralogs can be categorized as having undergone subneofunctionalization (SNF) (He and Zhang 2005b). Mutations that cause NF, SF, or SNF promote functional persistence of both duplicates because the changes they cause are advantageous, or because they make both paralogs non-redundant, meaning that loss of either one is deleterious (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000; Lynch et al. 2001; Stoltzfus 1999), as opposed to the alternative where one copy becomes non-functional (a pseudogene). Furthermore, functional divergence of duplicates can reduce pleiotropy and, in doing so, catalyzes gene specialization. Therefore, the molecular and transcriptional fate of duplicate genes can have important repercussions on the evolutionary trajectory of entire genomes.

Expression and functional divergence is commonly observed between duplicate genes generated by large- and small-scale duplications (Adams 2007; Adams et al. 2003; Adams and Wendel 2005; Blanc and Wolfe 2004; Byrne and Wolfe 2007; Flagel et al. 2008; Gu et al. 2005; Gu et al. 2002; Ha et al. 2009; Jiang et al. 2007; Li et al. 2005; Yim et al. 2009). Consistent with regulatory neofunctionalization (RNF), some yeast and cotton duplicates have gained new expression domains when compared to a singleton ortholog or to their diploid progenitor parents (Chaudhary et al. 2009; Tirosh and Barkai 2007). SF occurs when complementary parts of the ancestral functions of a gene are lost and thus partitioned across both duplicate copies (Force et al. 1999; Lynch and Force 2000). After regulatory subfunctionalization (RSF), ancestral gene expression levels across tissues and developmental stages are distributed such that each duplicate daughter contributes only a portion of the ancestral expression, in a complementary fashion, thus requiring the retention of both copies. This type of reciprocal silencing can occur due to mutation but can also occur epigenetically (Rodin and Riggs 2003). Duplicate genes can additionally persist as redundant copies or back-ups to essential functions, for gene dosage or for dosage balance (Aury et al. 2006; Chapman et al. 2006; Kondrashov and Koonin 2004; Papp et al. 2003; Qian and Zhang 2008; Veitia

2004). Once duplicate gene retention has been triggered by such mechanisms, which might act concurrently during their evolution, subsequent modification and diversification can help establish their genetic fate within the genome.

The probability of persistence of a duplicate gene is positively correlated with the size of the duplication event – whole genome duplication (WGD) generates duplicate genes (paralogs) that tend to last as functional pairs for longer periods of time and at a higher frequency than single gene duplications (Blomme et al. 2006; Lynch and Conery 2000). This might stem from the duplication of entire networks of genes that interact with one another; the duplicate gene retention mechanisms that operate after whole genome duplication may be different from, or more effective than, those that operate after smaller scale duplication (Casneuf et al. 2006; Chain et al. 2008; Davis and Petrov 2005; Guan et al. 2006). There are many other genetic factors that influence the fate of duplicate genes such as their functions and interactions, and these can have various effects on selective pressures within, and between, organisms (Casneuf et al. 2006; Conant and Wagner 2002; Hakes et al. 2007; He and Zhang 2006; Rodin et al. 2005; Woody et al. 2008). For example, genes arising from large-scale duplications show lower levels of expression divergence in plants (Casneuf et al. 2006; Kim et al. 2006), but the opposite is found in fungi duplicates (Tirosh and Barkai 2007). Complex genes in eukaryotes are also more likely to be retained after duplication, likely via NF and SF (He and Zhang 2005a). Functionally redundant genes encoding multi-domain proteins rather than mono-domain proteins are predicted to persist longer due to the large effect of deleterious mutations (Gibson and Spring 1998). Highly expressed genes are preferentially retained after WGD, for example in *Saccharomyces cerevisiae* (Seoighe and Wolfe 1999) and in *Parameciun tetraurelia* (Aury et al. 2006). Slowly-evolving genes are also preferentially retained after WGD in *Saccharomyces cerevisiae, Caenorhabditis elegans* and teleost fishes (Brunet et al. 2006; Davis and Petrov 2004), and a study of the pseudo-tetraploid frog *Xenopus laevis* reported that duplicate genes with expression subfunctionalization tended to be slowly-evolving before duplication (Sémon and Wolfe 2008). Of interest then, is the extent to which (i) NF, SF, and SNF drive the functional longevity of duplicate genes and (ii) whether these mechanisms operate primarily on gene regulation or protein function. Additionally, genetic features such as the level and breadth of expression or the rate of evolution may predispose genes to one or another mechanism for duplicate gene persistence, raising a third question of (iii) the degree to which different genetic characteristics (e.g., level of expression, breadth of expression, rate of evolution) influence the functional persistence of both parlalogs after WGD.

Xenopus as a model to study duplicate genes after WGD

Xenopus laevis (XL) is an African clawed frog whose genome was duplicated between 21 and 41 million years ago after diverging from the ancestor of the diploid species *Silurana tropicalis* (ST) (Chain and Evans 2006; Evans et al. 2005). Previous work on the molecular evolution and expression of duplicate genes in *Xenopus* has provided some insight into duplicate gene retention. One finding is

that there is strong purifying selection on protein sequence evolution albeit relaxed compared to a singleton ortholog (Chain and Evans 2006; Chain et al. 2008; Hellsten et al. 2007; Hughes and Hughes 1993; Morin et al. 2006). A second finding is that a high proportion of paralogous expression patterns have diverged considerably (Chain et al. 2008; Hellsten et al. 2007; Morin et al. 2006; Sémon and Wolfe 2008). And a third finding is that slowly-evolving genes undergo expression subfunctionalization more frequently than rapidly-evolving genes, and therefore are more likely to be retained as a functional pair after WGD (Sémon and Wolfe 2008). Additional analyses of expression in XL using ST to infer ancestral expression can help further address questions relating to the interactions between gene expression and sequence evolution before WGD and the prevalence and characteristics of mechanisms that act on genes to retain them after WGD.

In this study, we used African clawed frogs (*Xenopus* and *Silurana*) to examine duplicate gene expression and sequence evolution after WGD. We first asked to what degree genetic attributes such as the level of expression, breadth of expression, and rate of molecular evolution influence the odds of functional persistence of duplicate genes in XL. Secondly, we asked what mechanisms operate on expression regulation and on protein sequence evolution that promote duplicate gene persistence, such as neofunctionalization, subfunctionalization and subneofunctionalization. To accomplish this, we analyzed the molecular evolution of over 3,000 'triads' of sequences consisting of two expressed duplicates in the tetraploid species XL and one singleton ortholog in the diploid species ST. For comparison, we also analyzed over 4,000 'dyads' consisting of putative XL singletons and their corresponding singleton ST ortholog. To explore regulatory evolution of XL genes, we examined expression patterns of duplicates and singletons inferred from non-normalized expressed sequence tag (EST) databases from 18 adult tissue types and developmental stages. Together these data include DNA sequences from over 1000 more duplicate genes and almost twice as many expression treatments than have been considered by previous studies of duplicate genes in clawed frogs.

RESULTS AND DISCUSSION

I. Factors impacting the odds of duplicate gene persistence

To what degree do different genetic variables influence the probability of duplicate gene persistence? Certain characteristics of molecular evolution and gene expression patterns prior to duplication, for example, might predispose a gene to persisting as a functional duplicate pair after WGD instead of having one duplicate become a pseudogene. We used data from the diploid species ST as an estimate of the "pre-duplication" gene characteristics to test whether (i) expression level, (ii) expression breadth, and (iii) rate of molecular evolution before WGD impact the odds of functional persistence of both paralogs in XL after WGD.

a. Expression level: Highly expressed genes are preferentially retained

We tested the possibility that highly expressed genes are more likely to be retained after duplication, reasoning that either these genes are more essential or higher expression provides more opportunity for quantitative expression subfunctionalization. As a surrogate for the "pre-duplication" expression of XL duplicate genes and XL singleton genes, we used the current expression of their corresponding orthologs in the diploid species ST. Consistent with our expectation, we found retained XL duplicates had significantly higher average expression levels than XL singletons in ST adult tissues ($P < 0.0001$, Mann-Whitney-Wilcoxon test; Figure 3.1a). These same trends were found when larval stages were incorporated, and also at more conservative expression thresholds (Methods). These findings corroborate results from a previous study (Sémon and Wolfe 2008), although the measure of expression level differs in that we calculate average expression level based on only those tissues in which a gene is expressed rather than averaging across all tissues, thus lessening the impacts of expression breadth (Methods).

However, there remains a strong positive correlation between average expression level and expression breadth ($P < 0.0001$), so we sought to control for this other factor when assessing the impact of expression level on the propensities of duplicate gene persistence. To this end, we developed a permutation test (permutation test 1, Methods) that allowed us to ask whether genes with low levels of expression are less likely to be retained as duplicates or (the related question) whether genes with high levels of expression are more likely to be retained as duplicates, while controlling for variation in breadth of expression and the tissue type where expression is detected. When we performed each of these tests, the first was significant ($P = 0.0223$; Figure 3.1b) while the second was just above a significant threshold of 0.05 ($P = 0.0540$; Figure 3.1c). With a more conservative criterion for determining expression levels (Methods), neither test is significant ($P = 0.2276$ and 0.2828, respectively), which could be caused by the more conservative dataset being smaller. Overall, these results suggest a positive correlation between adult tissue expression level before WGD and duplicate persistence after WGD but highlight an important role of correlated parameters including expression level and expression breadth. In other words, because these variables appear to be interrelated, they have a collective impact on the odds of functional persistence of duplicate genes in *Xenopus laevis*.

b. Expression breadth: Broadly expressed genes are preferentially retained

Another explanation for the persistence of duplicate genes is that broader expression, expression across tissues and development, is either more essential or provides more opportunities for spatiotemporal RSF. Consistent with our expectation, we found that genes expressed in more adult tissues ($P < 0.0001$, Mann-Whitney-Wilcoxon test) and in more developmental stages ($P < 0.0001$, Mann-Whitney-Wilcoxon test) preferentially persist as duplicates rather than singletons after WGD (Figure 3.2a and 3.2b). A similar trend was found when we measure breadth as the amount of EST libraries (both adult tissues and developmental stages)

in which a gene is found, and at a more conservative expression threshold to characterize spatial and temporal expression breadth (data not shown). Similar results are also found when evaluating the tissue-specificity statistic (Methods), which considers the expression level between tissues ($P < 0.0001$, Mann-Whitney-Wilcoxon test), suggesting that genes that are less tissue-specific are more likely to persist as duplicates. There are also significantly more ST orthologs of XL singletons (534) than of duplicates (246) that have single-tissue expression ($P < 0.0001$, G-test): genes with expression patterns under which spatiotemporal RSF is not possible. The strong positive correlation between average expression level and breadth raises the possibility that relationships involving expression breadth may also be related to expression levels (and vice versa, as mentioned above). That we see breadth correlated with duplicate gene persistence at more conservative thresholds suggests that breadth remains an indicator of duplicate persistence when we consider only those tissues in which genes are highly expressed. We also performed permutation tests that control for levels of expression (permutation test 2, Methods), and we still found that singleton orthologs of genes retained in duplicate are expressed in many tissues and developmental stages ($P = 0.0026$), whereas those of singletons have narrow expression ($P < 0.0001$). Together, these findings show that gene expression level and breadth are strong and inter-related indicators of which genes persist as duplicate genes rather than singletons after WGD, and that expression breadth remains a significant parameter after controlling for levels of gene expression.

c. Rate of molecular evolution: Slowly-evolving genes are preferentially retained

It has been suggested that slowly-evolving genes may be more likely retained in duplicate copy following duplication because they accumulate fewer substitutions over time than faster evolving genes, thus retaining their interchangeability for longer periods providing more opportunity for RSF to occur (Sémon and Wolfe 2008). Using this larger dataset, we tested whether rates of molecular evolution can be used to predict if a gene is retained in duplicate after WGD. A key feature of this analysis is that the rate of molecular evolution of singletons and duplicates in XL was inferred from the comparison of two diploid species (ST and another diploid frog from the same family – either *Pipa* or *Hymenochirus*). This analysis therefore provides an estimate of rates of evolution that is independent of WGD in XL. We calculated pre-duplication rates of nonsynonymous (dN) and synonymous (dS) substitutions between these species, which are probably the best approximations of rates of molecular evolution to date because we used the closest known frog species related to *Xenopus* (Methods).

Consistent with the results of Sémon and Wolfe (2008), who estimated pre-duplication rates between more distantly related species, genes retained as duplicates in XL evolved more slowly before WGD than those that became singletons (average $dN = 0.0084$ vs 0.0116 and average $dS = 0.1188$ vs 0.1355 respectively; $P < 0.0001$ for dN and dS, Mann-Whitney-Wilcoxon test) and were also under stronger

functional constraints (average dN/dS = 0.1075 vs 0.1530 respectively, $P < 0.0001$, Mann-Whitney-Wilcoxon test; Figure 3.3). One concern is that our results could be biased because the rate of molecular evolution is usually negatively correlated with expression level and breadth (Supplementary Information). For this reason, we performed permutation tests that attempt to isolate the effects of expression level and breadth (permutation tests 3 and 4, Methods) when evaluating the effects of dN on duplicate gene persistence. But in support of this general trend, these permutations indicate that genes that have been retained in duplicate are descendents of slower evolving genes when we control for expression levels ($P = 0.0080$) and expression breadth ($P = 0.0093$).

d. Relative impacts of each variable: Breadth of expression has the largest impact on the odds of duplicate gene persistence

Gene expression level, expression breadth and protein sequence rates of evolution are often coupled but might be affected by different selective forces (Liao and Zhang 2006). Without disentangling expression level and breadth from rate of sequence evolution, it is difficult to pinpoint whether genes are retained in duplicate because of their high expression, because of the opportunities for expression pattern partitioning, because they are slowly-evolving allowing more time for mechanisms of retention to act, or a combination of these factors. We therefore performed a logistic regression to partition the impact of all of these variables on the odds of duplicate gene retention in *Xenopus*. Our results show that, although all measured variables are predictors of duplicate gene persistence, breadth of expression has the strongest influence on the fate of genes after WGD. When all three factors are taken into account, expression breadth ($P < 2x10^{-16}$) and expression level ($P = 4.51x10^{-5}$) are positively correlated with duplicate gene persistence and dN ($P = 6.85x10^{-5}$) is negatively correlated (Figure 3.4). The coefficients of the standardized values give an approximate strength of each variable, whereby breadth (0.54390) has more than double the effect of both intensity (0.21584) and dN (-0.18471)

II. Mechanisms of duplicate gene persistence after WGD

What mechanisms act on expression and protein sequence to promote paralog persistence? Superfluous genetic copies are expected to be purged from the genome within a few million years as a result of mutation (Lynch and Conery 2000). We have identified 3640 pairs of duplicate genes in XL that have survived about 40 million years (Chain et al. 2008; Evans et al. 2005), suggesting that mechanisms involving natural selection have acted to retain the function of these duplicate genes after WGD. Certain mechanisms of duplicate gene retention such as NF, SF, and SNF depend on divergence of duplicate genes, either through amino acid substitutions or changes in regulation of expression. We explored the degree to which expression divergence between paralogs in XL is consistent with alternative expectations of regulatory NF, SF, and SNF.

Non-normalized EST data were used to evaluate the divergence between duplicate gene expression patterns across 14 adult tissues and 4 larval stages.

Because low expression levels can introduce noise and biases, we repeated our analyses using different thresholds of expression, restricting our dataset to genes that were expressed above expression levels based on a statistical test developed by (Audic and Claverie 1997). Expression pattern divergence was measured using both the Spearman's rank correlation coefficient rho (ρ), and the Pearson's product-moment correlation coefficient (R). We report both correlations because properties of each statistic accommodate different aspects of the data, although the results from each are strongly positively correlated. Whereas Spearman's correlation is more robust against biases in the EST data by calculating a relative ranking of gene expression across tissues, the Pearson correlation is more sensitive to the proportional differences in expression levels between duplicate genes. Additional efforts to estimate the prevalence of RNF and RSF were conducted by comparing duplicate gene expression patterns with those of a singleton ortholog from a diploid outgroup which acts as a proxy for ancestral expression. This allowed us to categorize expression patterns into different classes based on explicit expectations under various models of duplicate gene retention.

a. Many paralogous expression patterns are substantially diverged

Immediately following duplication of an entire gene including its regulatory elements, a reasonable null expectation is that duplicate genes would have nearly identical expression patterns and a correlation coefficient near one (Gu et al. 2002). In a species such as XL which is probably an allopolyploid (Evans 2007; Evans 2008; Kobel 1996), this may not actually be true because expression divergence could have occurred in each of the diploid ancestors prior to WGD. In any case, when paralogous expression diverges in level or breadth, the correlation between them decreases. The correlation coefficient between duplicate gene expression patterns was measured and compared to a null distribution of correlation coefficients between 50,000 random pairs of genes from our duplicate dataset (Figure 3.5). Between random pairs, 95% had correlation coefficients below a Spearman's ρ of 0.612 and below a Pearson's R of 0.665. Between duplicate genes, we found 78.5% and 78.2% that also had correlation coefficients below these levels, respectively. Many duplicate genes have negative correlations, 33.2% with $\rho < 0$, and 39.3% with $R < 0$. Similar trends are found when evaluating expression at more conservative thresholds. The high proportion of duplicate gene expression divergence is similar to those found in yeast (Gu et al. 2002), rice (Yim et al. 2009), and *Arabidopsis* (Blanc and Wolfe 2004). These correlations suggest that the majority of duplicate genes in XL have diverged in expression from one another to the extent that their expression patterns are as dissimilar as two random genes. This substantial and pervasive expression divergence is higher than previous studies of *Xenopus* based on microarray data and EST data (Chain et al. 2008; Hellsten et al. 2007; Sémon and Wolfe 2008), probably because of the greater number of tissues used here.

b. Multiple mechanisms act to promote functional persistence of paralogous pairs

In addition to assessing paralogous expression pattern divergence using correlations, discrete expression profiles may be useful indicators of the evolutionary history of genes and of the type of selection that might have acted to keep both copies functional in the genome. We classified XL duplicate genes into 6 profile classes based on gene expression distribution patterns and that of their singleton ortholog in ST (Figure 3.6). Comparison to ST is an essential component in distinguishing between different mechanisms of duplicate gene retention because the diploid outgroup is used to estimate the ancestral profile of expression, allowing us to differentiate between loss and gain of expression after duplication.

We infer expression redundancy (i.e. overlapping paralogous expression profiles) to contribute to the retention of some duplicate genes, although we cannot rule out the contributions of protein SF or NF, or expression gain and loss in other tissues not sampled. These particular genes can have completely overlapping expression, or partial overlap with either loss or gain of expression domains (Figure 3.6a-c). An expression profile is consistent with regulatory subfunctionalization (RSF) when both duplicate genes have lost expression in a complementary fashion, or "reciprocal silencing", such that combined they have an expression profile similar to that of their ortholog (Figure 3.6d). Another alternative, regulatory neofunctionalization (RNF) necessitates duplicate genes to have asymmetric patterns of expression: one paralog with an equal expression profile with the ortholog, and the other paralog with a novel expression tissue and loss of ancestral expression in some capacity (Figure 3.6e). Some genes were found to have complementary expression profiles in addition to novel expression, a pattern consistent with the regulatory subneofunctionalization (RSNF) model where there is evidence for both SF and NF, and these were classified separately (Figure 3.6f).

We quantified the proportion of genes that have these six particular paralogous expression profiles based on whether we find ESTs of a given gene in a library. Data from 2650 triads, in which ESTs were found for both paralogs and their ortholog, revealed only 15 genes (0.6%) that had completely overlapping expression, while 328 (12.4%) had overlapping with loss of expression and 599 (22.6%) had overlapping with gain of expression, relative to each other and to a singleton ortholog. In addition, 212 genes (8.0%) had profiles consistent with RSF, 681 genes (25.7%) with RNF, and 815 genes (30.8%) with RSNF. The high proportion of genes in the last three categories suggests that over half of the duplicate genes have diverged in expression in ways that are consistent with the retention of duplicate genes through the loss or gain of expression domains. About a third of the genes show both unique gain and complementary loss of expression across adult tissues and larval stages. These genes are consistent with RSNF, a situation in which RSF could initially retain both duplicate genes, acting as a transition state to longer-term RNF (He and Zhang 2005b; Rastogi and Liberles 2005). Moreover, most duplicate genes are expressed in tissues where their ST ortholog is not detected. The proportion of genes assigned to each class changed

somewhat when the presence and absence of gene expression was scored based on more stringent thresholds, but many genes still display expression divergence consistent with particular mechanisms of retention (Supplementary Information). While expression redundancy may be widely associated with the retention of some duplicate genes (Aury et al. 2006; Dean et al. 2008), these results suggest that RSF, RNF and RSNF have important roles in the preservation of duplicate genes after WGD. Moreover, our estimates of RSF are similar to previous estimates of expression asymmetry in *Xenopus* using different methods (Hellsten et al. 2007; Sémon and Wolfe 2008) and considerably higher (~5-35%) if we consider that a portion of the genes categorized under RSNF were initially retained by RSF alone. Furthermore, subtler quantitative RSF in genes categorized as overlapping, in which both copies have similar tissues of expression albeit different levels of expression, can contribute to duplicate gene retention. In contrast with RSF, we found many more genes that have gained new expression domains, consistent with widespread RNF and RSNF.

These results must be considered in the context of uncertainties that remain with respect to our estimates of the relative contributions of mechanisms of duplicate gene retention that involve gene expression. Of those genes categorized under RSNF, for example, we cannot exclude the possibilities that (i) gain of expression domain occurred prior to loss, in which case RNF would have initially acted to retain these genes; (ii) RSF alone retained these genes, and gain of expression domain occurred as a byproduct of divergence; or (iii) some of these profiles are incorrect because of changes in the expression profile of the ST ortholog. Additionally, the categorization of certain genes in particular classes could be due to lack of sample coverage if we did not detect expression based on ESTs because of low expression levels in either species. The identification of RNF could also be misdiagnosed if there was loss of expression in the diploid species, made possible if expression or function is partly redundant due to another close gene family member (Woody et al. 2008), or even an unrelated gene (Nowak et al. 1997). Furthermore, differences between XL duplicate gene profiles and those of ST could be a result of global expression divergence between species following WGD (Supplementary Information).

c. Genes expressed highly and broadly before WGD have several potential fates after WGD

One explanation for the finding that highly and broadly expressed genes tend to persist as functional duplicates is that their persistence is catalyzed by quantitative or spatiotemporal RSF. However, if these types of genes are more essential, under stronger pleiotropic constraints, or if they interact in more variable or complex environments with important functions (Duret and Mouchiroud 2000; Hastings 1996; Pál et al. 2001), we might expect the level of expression to be inversely correlated with paralogous expression divergence. Our results are consistent with this prediction.

After WGD, we find that XL duplicate genes have higher expression levels and greater expression breadth than XL singletons ($P < 0.0001$, Mann-Whitney-Wilcoxon test). Among the surviving paralogs, distinct patterns of gene expression prior to WGD can affect the expression patterns of duplicate genes after WGD. For example, genes with spatiotemporal profiles of RSF and RSNF have significantly broader expression in ST than any of the other classes of genes ($P < 0.0001$, Mann-Whitney-Wilcoxon test) where broader expressed genes are more likely to be categorized under RSF than RSNF ($P = 0.0356$, Mann-Whitney-Wilcoxon test; Figure 3.7). Because RSF requires the ancestral gene to be expressed in at least 2 different tissues or developmental stages, it is more likely to occur in genes with higher breadth. A logistic regression which considers both variables supports this, whereby the odds of duplicate genes being classified under RSF does not significantly increase with higher ancestral expression level ($P = 0.115$), but does with higher ancestral breadth ($P < 2x10^{-16}$).

In addition to finding broadly expressed genes undergoing RSF, genes with higher correlations between their expression profiles after WGD (based on Spearman's rho and Pearson R) have higher average expression levels before duplication ($P < 0.0001$, ANOVA). This indicates that many highly expressed genes retain similar expression after duplication, as evinced by those genes that have completely retained overlapping expression (Figure 3.7). Due to the nature of the expression profile correlations between paralogs, there is an inverse relationship between these correlations and breadth of expression in ST ($P < 0.0005$, ANOVA). However, after controlling for tissues of expression using permutations, duplicate genes with conserved expression profiles (rho and R) still have higher average expression levels before WGD compared to genes with diverged expression profiles ($P = 0.0081$ and $P = 0.0460$ respectively) and compared to singleton genes ($P = 0.0435$ and $P = 0.0337$ respectively).

One group of genes with high expression was categorized under "overlapping with lost expression". It is possible that such genes are being retained via quantitative RSF, in which varying levels of expression in complementary tissues necessitates both genes to be expressed. In contrast, genes that have gained unique expression, either those under "overlapping with gain of expression" or RNF, have lower expression and moderate breadth in ST, and the few duplicate genes that have completely overlapping expression are highly tissue-specific (Figure 3.7). Genes that have become singletons after WGD have characteristically low expression levels and rather narrow expression as well. Taken together, our results suggest that highly and broadly expressed genes before duplication are preferentially retained as duplicates after duplication; some of these genes with both high and broad expression diverge in ways consistent with RSF and RSNF, while those with high expression but lower breadth retain similar expression consistent with gene dosage and stoichiometric constraints. Other genes with lower expression levels and moderate breadth tend to gain novel spatiotemporal expression characteristics or become singletons.

Broadly expressed genes are likely to experience regulatory subfunctionalization

Genes expressed at higher levels and in more tissues evolve slowly in numerous species (Drummond et al. 2005; Duret and Mouchiroud 2000; Gu and Su 2007; Hastings 1996; Jordan et al. 2005; Kim and Yi 2006; Kuma et al. 1995; Liao and Zhang 2006; MacEachern et al. 2006; Pál et al. 2001; Subramanian and Kumar 2004; Yang et al. 2005; Zhang and Li 2004). We also found evidence consistent with these expectations in clawed frogs (Supplementary Information). Whereas some studies have found that gene expression levels are primarily responsible for rates of coding sequence evolution (Drummond et al. 2005; Liao and Zhang 2006; Subramanian and Kumar 2004), Spearman correlations in this study suggest that in clawed frogs, the breadth of expression has a stronger relationship with dN than does the level of expression (ST: $P < 0.0001$, rho = -0.1482 vs $P = 0.2278$, rho = 0.0215 respectively; XL: $P < 0.0001$, rho = -0.1356 vs $P = 0.3523$, rho= 0.0119 respectively).

The impact of breadth of expression on duplicate gene persistence is further illustrated by testing whether there exists a negative relationship between rates of molecular evolution and duplicate gene expression divergence, as has been found in other studies (Gu et al. 2002; Makova and Li 2003; Pál et al. 2001). Genes with conserved expression have high expression but narrow expression, and we did not find significant relationships between rates of molecular evolution and duplicate gene expression correlations (Supplementary Information). In contrast, genes that have been subfunctionalized were both highly and broadly expressed before WGD. Logistic regressions suggest that dN influences the odds of a gene being subfunctionalized after controlling for expression ($P = 0.0465$) but this effect disappears after controlling for expression breadth ($P = 0.659$). Therefore, genes that have undergone RSF are preferentially slow evolving as was previously found in (Sémon and Wolfe 2008), but this appears to be strongly linked to breadth of expression. Moreover, logistic regression indicates that the odds of duplicate genes being subfunctionalized are significantly affected by expression breadth ($P = 4 \times 10^{-11}$) but not by expression level or rates of molecular evolution ($P = 0.82034$ and $P = 0.67023$ respectively). Our results therefore suggest that expression breadth rather than expression level or rate of evolution plays an important role in determining which duplicate genes are affected by regulatory subfunctionalization.

CONCLUSIONS

Whole genome duplication (WGD), such as the ones that occurred in multiple vertebrate lineages, creates opportunities for genetic modifications with fitness consequences that are distinct from those associated with non-redundant (singleton) genes. Subsequent rearrangements and duplicate gene divergence creates the potential for rapid and pervasive transcriptome remodeling. Here we compared expression and evolution of genes that have been retained in duplicate in *Xenopus laevis* following WGD to genes in this species that have become singletons

after WGD. We found that genes that are expressed at higher levels, in more tissues, and that are slower evolving are preferentially retained as duplicates after WGD, with expression breadth before WGD being the largest determinant of duplicate gene retention. Numerous functional and persistent paralogous pairs have substantially diverged regulation, and have expression profiles consistent with expectations of regulatory subfunctionalization, neofunctionalization and subneofunctionalization. The descendants of genes expressed at a high level and in many tissue types had more similar expression profiles or were subfunctionalized, whereas genes that were neofunctionalized came from lower and more narrowly expressed genes. Our findings suggest that there are likely stronger dosage constraints on duplicate genes that are expressed at higher levels and in more tissue types that retain similar expression, whereas other genes partition their expression by reducing expression levels and breadth. In general, genes that are expressed at lower levels and in fewer tissues might be more likely to revert to singleton status due to lack of opportunities for expression partitioning or due to reduced selective constraints and faster rates of evolution, or a combination of these and other factors.

METHODS

Analyses of Molecular Evolution

This study examines molecular evolution and expression of duplicate and singleton genes in the tetraploid species *Xenopus laevis* (XL) and singleton genes in the diploid species *Silurana tropicalis* (ST). Thus, our database is comprised of gene triads, which include a pair of XL paralogs and the corresponding ST ortholog, and gene dyads, which include an XL singleton (in which the other paralog is probably a pseudogene) and the corresponding ST singleton.

Nucleotide sequences from 8162 sets of gene triads and dyads were gathered from (1) published lists of XL paralogs and their orthologous singleton in ST (Chain and Evans 2006; Hellsten et al. 2007; Hughes and Hughes 1993; Morin et al. 2006), as well as (2) nucleotide sequence data acquired from public databases (NCBI, TIGR and JGI) following similar protocols as in these previous studies. Because online databases contain redundant sequences, clustering of the XL and ST databases was performed with TGICL (Pertea et al. 2003) grouping sequences having ≥98% identity over a length of at least 300bp. BLAST was then used to find paralogs and orthologs using a reciprocal best-hit approach, whereby the BLAST hits of each XL gene returns one another, the top BLAST hit of both XL genes is the same singleton ST gene, and the two top hits of the reciprocal BLAST from ST returns the two XL genes.

To reduce the chances of retaining sequences we mislabeled as paralogs, the genetic distances between XL paralogs and their singleton ortholog was calculated using dnadist from the PHYLIP package, using a Jukes-Cantor correction (Felsenstein 1993). Our selection criteria were based on an examination of 99 *confirmed* paralog pairs derived from whole genome duplication, based on phylogenetic results from additional polyploid frogs that speciated from XL after the

genome duplication event (Chain et al. 2008). The minimum paralogous distance between confirmed XL paralogs was 0.02455, and paralogs were at least one third divergent from each other as they were from ST, with the minimum ratio of (XL paralog difference)/(average XL – ST) being 0.34654. These distances, in addition to a minimum sequence length of 201 nucleotides to avoid analyzing short sequences, were used as a minimum threshold for our triads. The average triad sequence length (overlapping coding regions) is 983bp. We then used the minimum and maximum distance between an XL paralog and an ST singleton as thresholds for dyads. As a final measure to minimize mislabeled, redundant, or alternatively spliced sequences in our gene lists, a BLAST search was performed using all of these genes against each other. We eliminated sequences that had overlapping regions and we kept the longest of the alternative transcripts of a gene.

The 3640 triads were obtained using a BLAST search of the Entrez Nucleotide XL database against itself and against the TIGR XL database, grouping top hits. Putative paralogs were sequences that had a minimum percent identity of 80% to avoid grouping paralogs that stem from an older duplication than the polyploidization event. Putative orthologs were obtained using reciprocal BLAST with the Entrez Nucleotide ST database and had the following criteria: (1) each putative XL paralog returned the same ST hit, and (2) ST sequences returned both putative paralogs as top hits. This approach reduces the chances of grouping paralogs that stem from an older duplication event than the whole genome duplication.

The 4522 dyads were identified using a BLAST search of the XL and ST databases. Putative XL singletons were sequences with the following criteria: (1) each had a reciprocal top hit with ST, and (2) each had a unique ST hit when compared with other sequences that were more than 200bp and 1% divergent from itself (to avoid allelic sequences). In other words, because other XL sequences that are diverged might be paralogs, they were required to have a different top ST hit (whereas paralogs have the same top ST hit). Singleton status of 17 XL genes was confirmed using PCR-based assays.

Singleton orthologs in ST were blasted against the ST database to confirm that they are not duplicate genes that arose after the divergence of the most recent common ancestor of ST and XL. Once grouped in putative triads or dyads, sequences were aligned using Muscle (Edgar 2004), and Perl scripts were used to predict the beginning and end of coding regions by looking for the longest open reading frame in either direction, although we note that a common start codon was not found in up to 1064 triads and 447 dyads. Non-overlapping portions of the coding regions were omitted (5' and 3' regions, indels, degenerate bases and missing sequences were removed), and sequences were subsequently manually inspected to check for errors. PAML (Yang 1997) was used to evaluate non-synonymous and synonymous divergence in each gene. Sequences from outgroup taxa (*Hymenochirus curtipes* and *Pipa carvalhoi*) were acquired from 454 pyrosequencing and were used to root the *Xenopus* clade. Pyrosequence data was

assembled using gsAssembler and gsMapper (454/Roche) and was aligned to our list of triads and dyads using Perl scripts, BLAST and MUSCLE.

Analysis of Expression

Non-normalized EST (Expressed Sequence Tag) libraries were obtained from NCBI (from a total of 677,784 ESTs from XL and 1,271,375 ESTs from ST), and were classified into 14 adult tissues and 4 embryological stages: brain, bone, eye, heart, kidney, liver, lung, thymus, pancreas, skin, spleen, fat body, ovary, testis, egg, nastrula stage, neurula stage and embryo stage 62. Both species have at least 3900 ESTs in each tissue or stage. ESTs were trimmed to remove vector sequences, and repeats were masked using RepeatMasker (Smit et al. 2004). A BLAST search was performed between these EST sequences and our list of 8162 genes, for which a single EST was assigned to represent a unique duplicate copy or singleton. The minimum distance between paralogs in our list of triads was 2.5%; therefore we chose to assign an EST to a particular sequence if its top BLAST hit had less than 2.4% bp mismatches over a length of at least 75bp, and its second top BLAST hit was above this threshold; this should thus accommodate for allelic differences or sequence errors in ESTs (leading to mismatches between the EST and the allele in our triads) and for the length of ESTs (which might only overlap with parts of the remaining coding regions of our triad sequences). Alternative clustering methods have previously been used to group ESTs with distinct paralogs (Sémon and Wolfe 2008), but we found that even with very stringent criteria, paralogous ESTs were often incorrectly grouped together, possibly due to the gaps present in our triad sequences (see above methods).

Expression level of each gene, in each EST library and in each species was estimated by the number of matching ESTs within a library, standardized by the total number of ESTs in that library. This results in a proportion of ESTs within each library, instead of raw EST counts, thus lessening the bias caused by the variation in library sizes when comparing between libraries. The average expression level of a gene was calculated by taking the mean expression level across only those adult tissues and larval developmental stages in which it is expressed, lessening the impact of expression breadth. This differs from the expression level calculated in a previous study, which uses an average of expression levels across all tissues, whether the gene is expressed in those tissues or not (Sémon and Wolfe 2008). When we evaluated the effects of expression level, we used log-transformed values. Our use of XL and ST sequences of equal length and of solely overlapping regions is predicted to reduce bias in detecting ESTs, for example only detecting ESTs in one paralog because of longer available (published) sequence. 2650 gene triads and 4317 gene dyads had at least 1 EST for which we could incorporate in our expression analyses. To reduce possible bias in digital profiling of expression, we used different threshold levels to determine expression in a Bayesian test for significant expression (Audic and Claverie 1997). The low threshold used a minimum of 1 EST to indicate expression, the medium threshold used a minimum of 4 ESTs, and a high threshold used a minimum of 7 ESTs. All analyses were repeated

using additional normalized EST libraries (that included an additional embryological stage), and overall results were similar (results not shown).

We refer to the number of tissues or developmental stages where a gene is expressed as the expression breadth. In addition to breadth, we also calculated a parameter of tissue specificity (Yanai et al. 2005), which refers to how even or uneven expression of a gene is among the tissues where it is expressed. This index measures the tissue specificity of a gene based on expression breadth and expression level in each tissue. A gene expressed in only one tissue is given a value of 1, and a gene expressed equally across all tissues has a value of 0, as specified by the formula:

$$(\sum_{i=1}^{N} (1-x_i)) / (N-1)$$

where N is the total number of tissues examined, and x is:

(expression in tissue i / highest tissue expression)

Results from our analyses using this parameter were very similar to our analyses using breadth (see Supplementary Information).

Permutation tests and logistic regressions

One goal of this study is to test for the impact of expression and rate of molecular evolution on the probability of duplicate gene retention and of subfunctionalization. But because these variables are not independent, we designed permutation procedures to hold constant the effects of some variables while evaluating the impact of another. These permutation tests grouped genes into categories before calculating average variable values across all categories. When assessing the impact of expression level on duplicate gene persistence, permutation test 1 controlled for patterns of tissue expression on expression level by categorizing genes by expression profile (in terms of tissues of expression). This bins singleton and duplicate genes into expression profile categories, enabling the calculation of a test statistic based on the mean expression level of randomly sampled genes within each profile category (with sample size equal to the number of duplicate genes within each profile), and averaging these means across all profile categories. By running 10,000 permutations, we created a distribution of expected expression level values (by randomly selecting from pools of singleton and duplicate genes) against which we then compared our observed average expression levels of duplicate genes. Permutation test 2 controlled for expression level when assessing the impact of expression breadth on duplicate gene persistence. This was performed in much of the same manner as permutation test 1, but by categorizing genes based on their average expression level and calculating a test statistic from expression breadth values. Permutation tests 3 and 4 were also identical in the way that genes were categorized (either by expression level or breadth), but the test statistics were calculated from dN and dS values.

In addition, we performed logistic regressions with R (R Development Core Team (2005) *R: A language and environment for statistical computing*; http://www.R-project.org) to evaluate the contributions of these variables on the odds of persisting as a duplicate gene (or becoming subfunctionalized) after WGD. After standardizing the variables that we tested (by the standard deviation), the coefficients given by the logistic regression are an approximation of the relative influence that each variable has on the outcome (either the odds of persisting as a duplicate gene or the odds of becoming subfunctionalized).

ACKNOWLEDGEMENTS

We thank Jonathan Dushoff for great advice on analytic approaches. We also thank Uffe Hellstein, Ryan Morin, Erika Lindquist, Daniela Gerhard and Bruce Blumberg for providing information on sequence and EST data, and Carlo Artieri, Brian Golding, Wilfried Haerty, Richard Morton and Jonathon Stone for helpful suggestions.

SUPPLEMENTARY INFORMATION

Do expression characteristics correlate with each other?

In this study, we examine expression level and expression breadth in clawed frogs to see whether genes with these characteristics have greater odds of being retained in duplicate after WGD in clawed frogs. However, there exists a strong positive correlation between these expression metrics in several species (Lercher et al. 2002; Subramanian and Kumar 2004; Vinogradov 2004). In both clawed frog species we studied, expression level positively correlates with breadth of expression in singleton genes (ST: Spearman's rho = 0.4781, $P < 0.0001$; XL: Spearman's rho = 0.2892, $P < 0.0001$), as well as in duplicate genes (XL: Spearman's rho = 0.3789, $P < 0.0001$). In other words, genes that are expressed at high levels are usually expressed in many tissues. We also find that average expression level correlates negatively with tissue-specificity in singleton genes (ST: Spearman's rho = -0.2767, $P < 0.0001$; XL: Spearman's rho = -1.648, $P < 0.0001$), as well as in duplicate genes (XL: Spearman's rho = -0.2237, $P < 0.0001$), a relationship not found in humans and mice (Liao and Zhang 2006). This suggests that genes that are more tissue-specific are expressed at lower levels. As expected, expression breadth and tissue specificity are strongly negatively correlated in singletons (ST: Spearman's rho = -0.8582, $P < 0.0001$; XL: Spearman's rho = -0.9343, $P < 0.0001$) and in duplicates (XL: Spearman's rho = -0.9229, $P < 0.0001$). Because tissue specificity values incorporate aspects of expression level and are strongly correlated with expression level, we chose to conduct our analyses using breadth of expression rather than tissue specificity.

Do expression characteristics correlate with sequence evolution?

Highly and broadly expressed genes have lower rates of protein substitution in many organisms (Drummond et al. 2005; Duret and Mouchiroud 2000; Gu and Su 2007; Hastings 1996; Jordan et al. 2005; Kim and Yi 2006; Kuma et al. 1995; Liao and Zhang 2006; MacEachern et al. 2006; Pál et al. 2001; Subramanian and Kumar 2004; Yang et al. 2005; Zhang and Li 2004). Using outgroups to estimate the molecular evolution of genes in diploids before WGD (Methods), we tested for these relationships and find consistent results in clawed frogs. There is a weak negative correlation between expression level and dN (Spearman's rho = -0.0441, P = 0.0135) and dS (Spearman's rho = -0.0541, P = 0.0024), but only when expression level is averaged over all tissues and larval stages sampled, including those in which the gene of interest is not expressed. The effect of this is that narrowly expressed genes will have low overall expression level because of their absence in many tissues. For this reason, we calculate expression level across only those tissues and larval stages in which a gene is expressed to try to control for breadth (Methods). After doing so, the significant relationship between expression level and molecular rates of evolution disappears (dN: Spearman's rho = 0.0215, P = 0.2278; dS: Spearman's rho = -0.0305, P = 0.0877). However, there remains a negative correlation between expression breadth and dN (Spearman's rho = -0.1482, P < 0.0001), dS (Spearman's rho = -0.0718, P < 0.0001), and also dN/dS (Spearman's rho = -0.1041, P < 0.0001).

We asked whether duplicate copies that have diverged in expression level or breadth evolve at different rates after WGD. Of the pairs of duplicates that are expressed at different levels, the copy that is expressed at a higher level evolves more slowly (using a permutation test, lower dN: P = 0.0348 and lower dS: P = 0.0179). Of the pairs of duplicates that are expressed at different breadths, the copy that is expressed in more tissues and developmental stages evolves more slowly (using a permutation test, lower dN: P < 0.0006 and lower dS: P < 0.0039).

a. Many paralogous expression patterns are substantially diverged

Gene expression can considerably change soon after duplication (Adams et al. 2003; Gu et al. 2005; Gu et al. 2002; Makova and Li 2003). We evaluated paralogous expression divergence in *Xenopus laevis* using Spearman's rho (ρ) and Pearson's R. Between duplicate genes, the median correlation coefficient ρ was 0.27632 and R was 0.13797, compared with previous findings of 0.64 based on 11 tissues (Sémon and Wolfe 2008) and 0.77 based on 5 tissues and larval stages (Chain et al. 2008). These differences suggest that the more tissues and developmental stages that are incorporated in the analysis, the more divergence can be detected. However, we find more similar results to those previous findings when we use higher thresholds for determining expression (medium threshold: median ρ = 0.605, n=451; high threshold: median ρ = 0.728, n=225). To examine the relationships between expression level and breadth before WGD and the extent to which duplicate genes have diverged in expression patterns, we designated expression profiles as divergent or conserved, from a cutoff value based on the

correlation coefficient between the expression profiles of 50,000 random pairs of genes in our duplicate dataset. This gave us a null distribution of expression profile correlations against which we could test whether duplicate genes have diverged in expression as much as two randomly chosen genes. Based on a 95% cutoff of ρ and R, 2101 and 2092 genes had diverged paralogous expression profiles versus 574 and 583 genes with conserved expression profiles, respectively. We did not find a notable relationship between duplicate gene expression divergence and protein sequence divergence, suggesting that these are decoupled (Wagner 2000).

b. Multiple mechanisms act to promote functional persistence of paralogous pairs

We quantified the proportion of genes that have profiles consistent with six particular classes based on EST presence and absence data (Figure 3.7). However, lack of sampling coverage might bias the results by not detecting some genes that are present at lower levels (absence of ESTs). To reduce the impact of profile differences between genes with low expression, we evaluated paralogous expression profiles using higher expression thresholds. The results give somewhat different proportions of genes categorized in each class, but consistently show widespread expression divergence. For example, using a minimum of 4 ESTs (instead of 1) as the criteria for gene expression within a tissue, we find 1.6% overlapping genes, 35.5% overlapping with loss of expression, 22.4% overlapping with gain of expression, 12.8% under RSF, 16.7% under RNF, and 11.0% under RSNF.

The widespread expression changes between XL and ST can also be found among singleton genes: XL singletons have gained new expression domains at similar proportions (78.9%) to pairs of paralogs (79.1%). Similarly, 73.6% of XL singletons have also lost expression domains in addition to gaining new expression domains, compared to 65.5% of paralog pairs. Therefore, even those genes that are found in single copy have diverged from their ortholog in ST. Because RSF requires complementary loss of expression and RNF requires only one paralog to gain and lose expression while the other retains the ancestral expression, we cannot test these particular profiles with singletons. Although we have detected that singleton genes in XL have gained and lost expression domains at similar proportions to paralogous pairs when compared to ST orthologs, we cannot test whether these differences occurred after speciation but before WGD, after WGD while their duplicate copy was still expressed, or after pseudogenization of one copy, or even whether orthologous genes in both species still perform similar functions or are essential (Liao and Zhang 2008; Strähle and Rastegar 2008).

We also examined the evolution of highly and broadly expressed genes *after* WGD, to see if genes retain their characteristics as duplicate genes in XL. There is a significantly positive relationship between highly expressed XL duplicate genes and paralogous expression correlations ρ (Spearman's rho = 0.1940, $P < 0.0001$) and R (Spearman's rho = 0.2267, $P < 0.0001$), in agreement with previous reports that duplicate genes with conserved expression patterns or conserved expression levels tend to have higher expression (Morin et al. 2006; Tirosh and Barkai 2007). In

contrast, duplicate genes that show spatiotemporal patterns of RSF have significantly broader expression ($P < 0.0001$, Mann-Whitney-Wilcoxon test) than all other duplicate genes, but significantly narrower expression ($P < 0.0001$, Mann-Whitney-Wilcoxon test) when compared to genes that underwent RSNF. This is consistent with the RSF model, in which duplicate genes partition their expression such that they now both have reduced expression breadth after WGD.

In contrast to RSF, spatiotemporal RNF might have more opportunities to arise in genes with lower breadth of expression. But because RNF necessitates loss in addition to gain of expression domain (Figure 3.6e), RNF might be identified in genes with medium breadth of expression. Before WGD, we find that genes categorized under RNF have broader expression than genes that have overlapping expression ($P < 0.0001$, Mann-Whitney-Wilcoxon test), yet significantly narrower than genes under RSNF ($P < 0.0001$, Mann-Whitney-Wilcoxon test; Figure 3.7). These genes also have lower average expression levels than any other category of genes that have lost expression (Overlapping with loss, RSF and RSNF; $P < 0.0001$). Relative to genes that have become singletons, all classes of duplicate genes have higher levels of expression, and except for genes that are overlapping, duplicates also have greater breadth of expression.

We also examined how genes that have low expression in ST evolved following WGD, as duplicates in XL. Duplicate genes that show spatiotemporal RNF still have significantly broader expression than genes with overlapping expression ($P < 0.0001$, Mann-Whitney-Wilcoxon test), and are still lower expressed and narrower than genes under RSNF ($P < 0.0001$, Mann-Whitney-Wilcoxon test) after WGD (Figure 3.7). This is not inconsistent with the RNF model, in which the sum of the breadth of both gene copies increases, but their average breadth might remain the same or diminish, depending on the number of expression domains (tissues) that have been gained and lost. The genes in general have also retained their low levels of expression. Under RNF, one duplicate copy loses and gains a novel expression domain, which might be accompanied by mutations in the coding region, thus one of the duplicate copies might demonstrate high average dN if it is evolving quickly. In yeast, for example, the paralogous copy that has acquired novel expression evolves faster than the conserved copy and codes for dispensable proteins as opposed to essential ones (Tirosh and Barkai 2007). We did not detect faster evolution in paralogous copies that have acquired novel expression domains, not before nor after correcting for breadth of expression ($p=0.7678$, Mann-Whitney-Wilcoxon test and $P = 0.499$, using a permutation test, respectively). Like previous studies, we did not find a correlation between asymmetric rates of sequence evolution and expression correlations after WGD (Sémon and Wolfe 2008; Tirosh and Barkai 2007). These findings might not be surprising because NF and SF can obscure relationships that may pre-exist between expression and protein rates of evolution (Drummond et al. 2005).

Effect of breadth on gene profile categorization

We tested whether the breadth of a gene affected our odds of categorizing genes into classes such as Overlapping, RSF, RNF and RSNF, due to sampling rather than biological reasons. We selected all genes that were expressed in at least 4 tissues in ST, and randomly sampled 4 tissues in which expression was detected in either one of the XL paralogs or the ST ortholog. If expression breadth before WGD biases our detection of RSF, we should not see an increase in proportion of genes subfunctionalized with greater breadth of expression (because we only select 4 tissues regardless of breadth). However, we found that more genes with complementary loss of expression in broader expressed genes, even after sampling only a constant subsample of tissues. We repeated this analysis by selecting genes expressed in at least 6 tissues (up to 15 tissues), and find similar results.

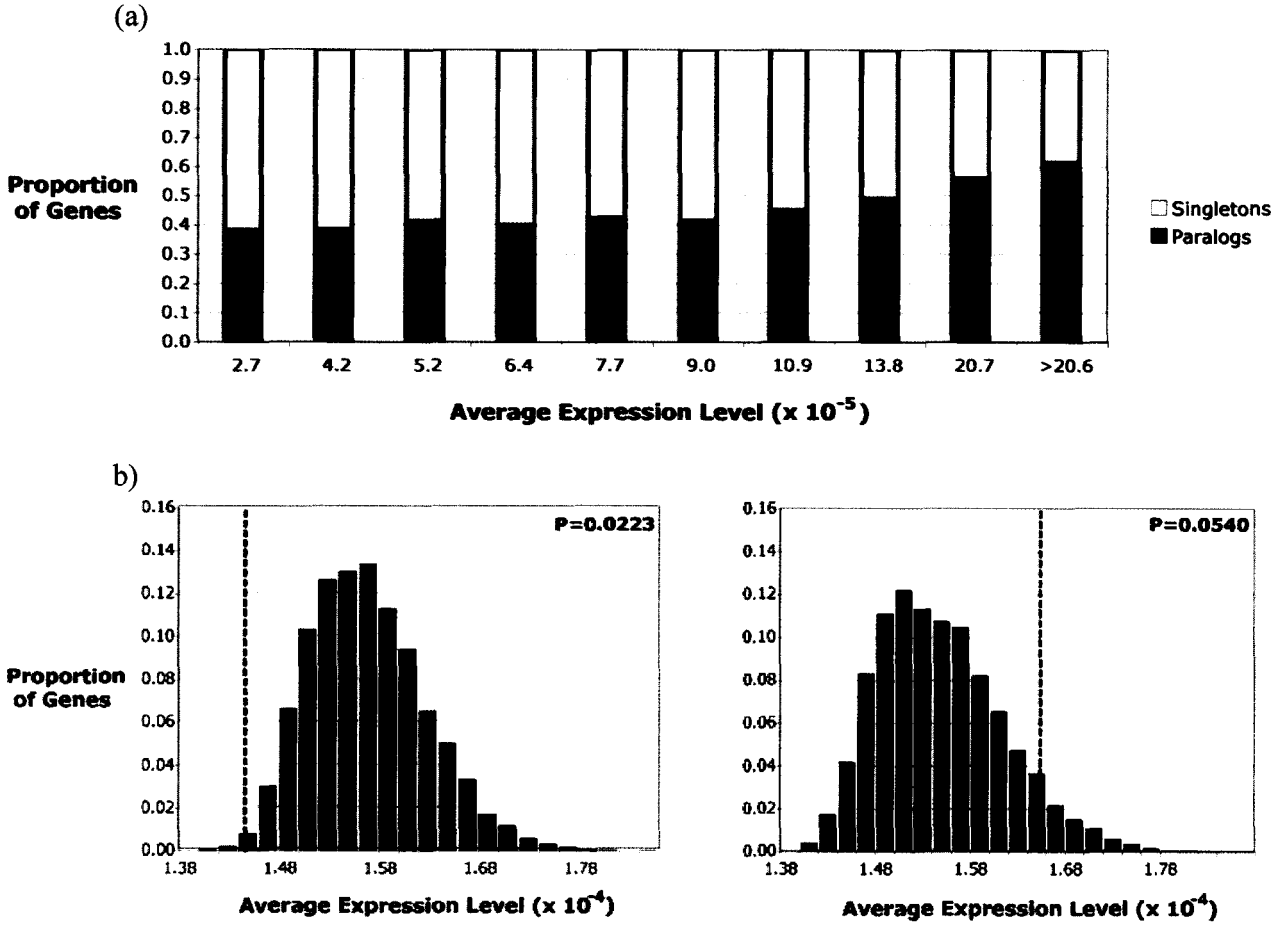


Figure 3.1. Before WGD (in ST), genes that are highly expressed tend to be retained as duplicate genes (paralogs), whereas genes that have low expression tend to be singletons after WGD. (a) Distribution of genes found in duplicate copy (black) versus single copy (white) in XL based on average expression level of their ortholog in ST (equal-sized bins on the x-axis). As the average expression level (proportion of ESTs) of ST genes across all 14 adult tissues and 4 larval stages increases, the proportion of paralogs in XL increases (y-axis). (b) The average expression level across all genes in ST (dashed line) of those genes whose orthologs are singletons in XL and (c) paralogs in XL, relative to the distribution of average expression levels acquired from 10,000 re-sampling permutations. Each permutation consists of the mean average expression level across the same number of genes as the observed data, (b) singletons or (c) paralogs, sampled from random genes (singletons and paralogs) with the same tissue expression patterns as the observed data.

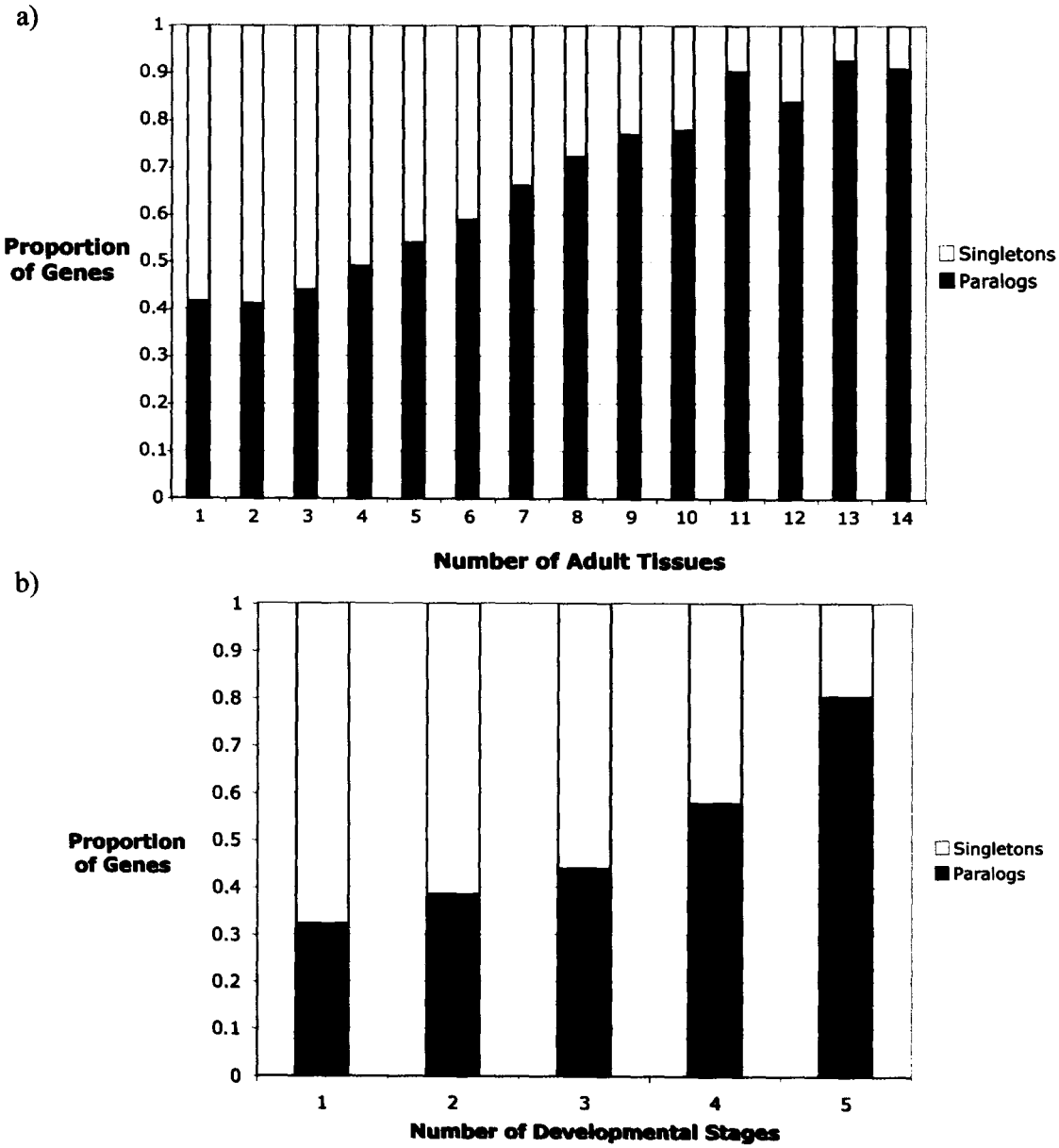


Figure 3.2. Before WGD (in ST), genes that are broadly expressed tend to be retained as duplicate genes (paralogs), whereas genes that have narrow expression tend to be singletons after WGD. Genes found in duplicate copy (black) versus single copy (white) in XL based on (a) spatial expression breadth (across adult tissues) and (b) temporal expression breadth (across developmental stages) of the ortholog in ST (unequal-sized bins of number of tissues or developmental stages on the x-axis). As the breadth of expression of ST genes increases, the proportion of paralogs in XL increases (y-axis).

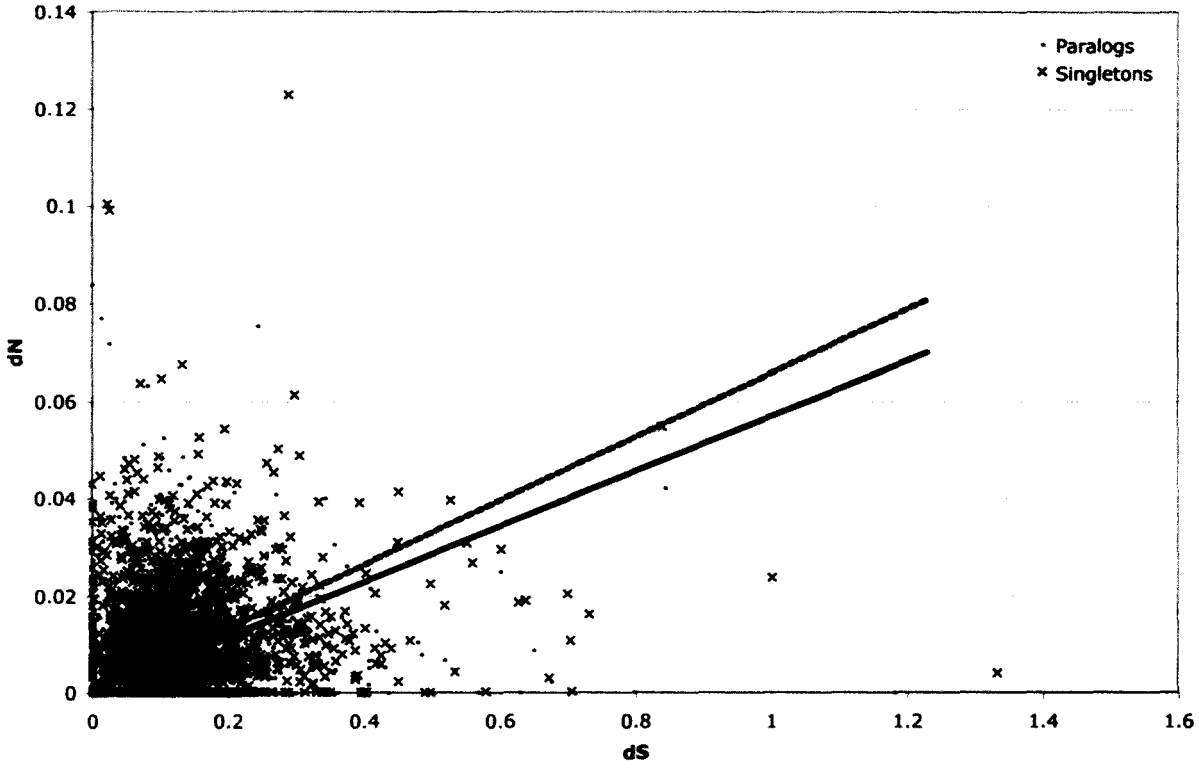


Figure 3.3. Before WGD (in ST), genes that are under stronger functional constraints and are evolving more slowly tend to be retained as duplicate genes (paralogs), whereas genes under more relaxed functional constraints and evolving more quickly tend to be singletons after WGD. (a) Distribution of genes found in duplicate copy (dots) versus single copy (crosses) in XL based on sequence rate of evolution of their ortholog in ST. The x-axis shows the synonymous rate of substitution (dS) and the y-axis shows the nonsynonymous rate of substitution (dN). The regression lines for genes that became singletons (dashed) and paralogs (solid) were forced to intersect at the origin (0,0). The rate of nonsynonymous substitutions per synonymous substitutions is greater in genes that become singletons compared to those that become paralogs.

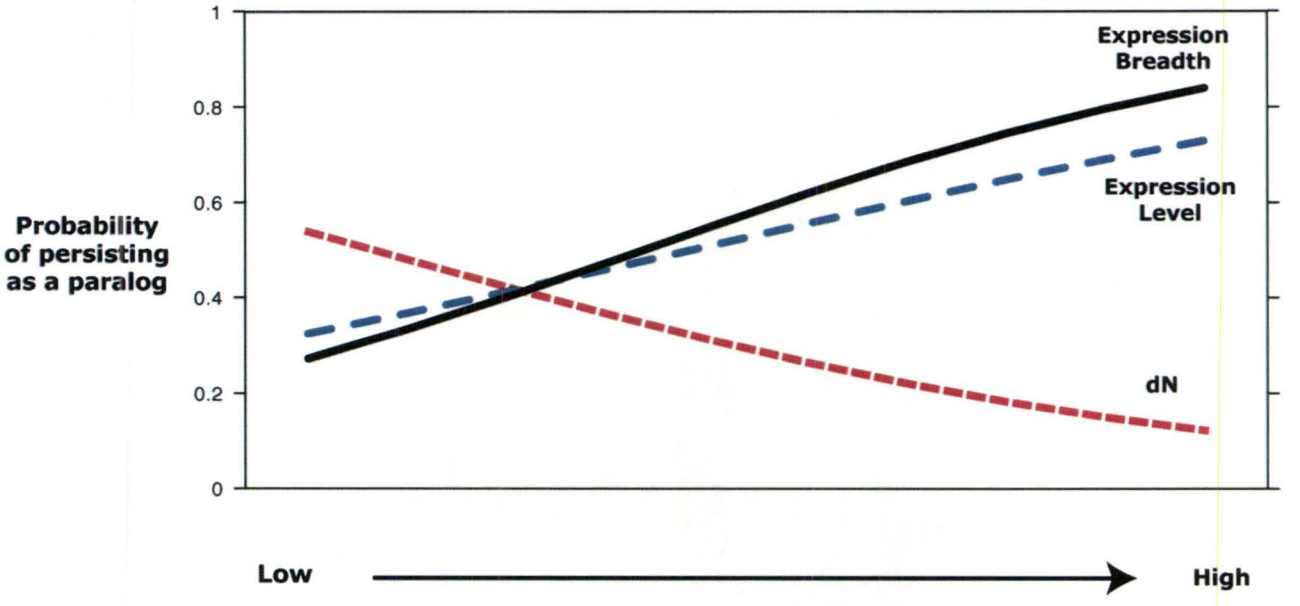


Figure 3.4. Before WGD (in ST), genes that are expressed at higher levels, in more tissues and developmental stages, and are evolving more slowly have greater odds of persisting as duplicate genes (paralogs) after WGD. The effects of expression level (long dashes), expression breadth (solid), and dN (short dashes) as explained by a logisitic regression. When incorporating standardized values of all variables into a logistic regression, the probability of a gene persisting as a paralog is best predicted by the equation $1/(1+e^{-z})$, where $z = -0.06899 + 0.21584$(expression level)) $+ 0.54390$(expression breadth) $- 0.18471$(dN). In other words, there is preferential retention of duplicates in genes that are higher expressed ($P = 4.51 \times 10^{-5}$), broader expressed ($P < 2 \times 10^{-16}$), and slower evolving ($P = 6.85 \times 10^{-5}$) before WGD. Given that the coefficients in the logistic regression formula represent approximate relative influence of each variable, expression breadth is the best predictor of whether a gene persists as a duplicate or a singleton.

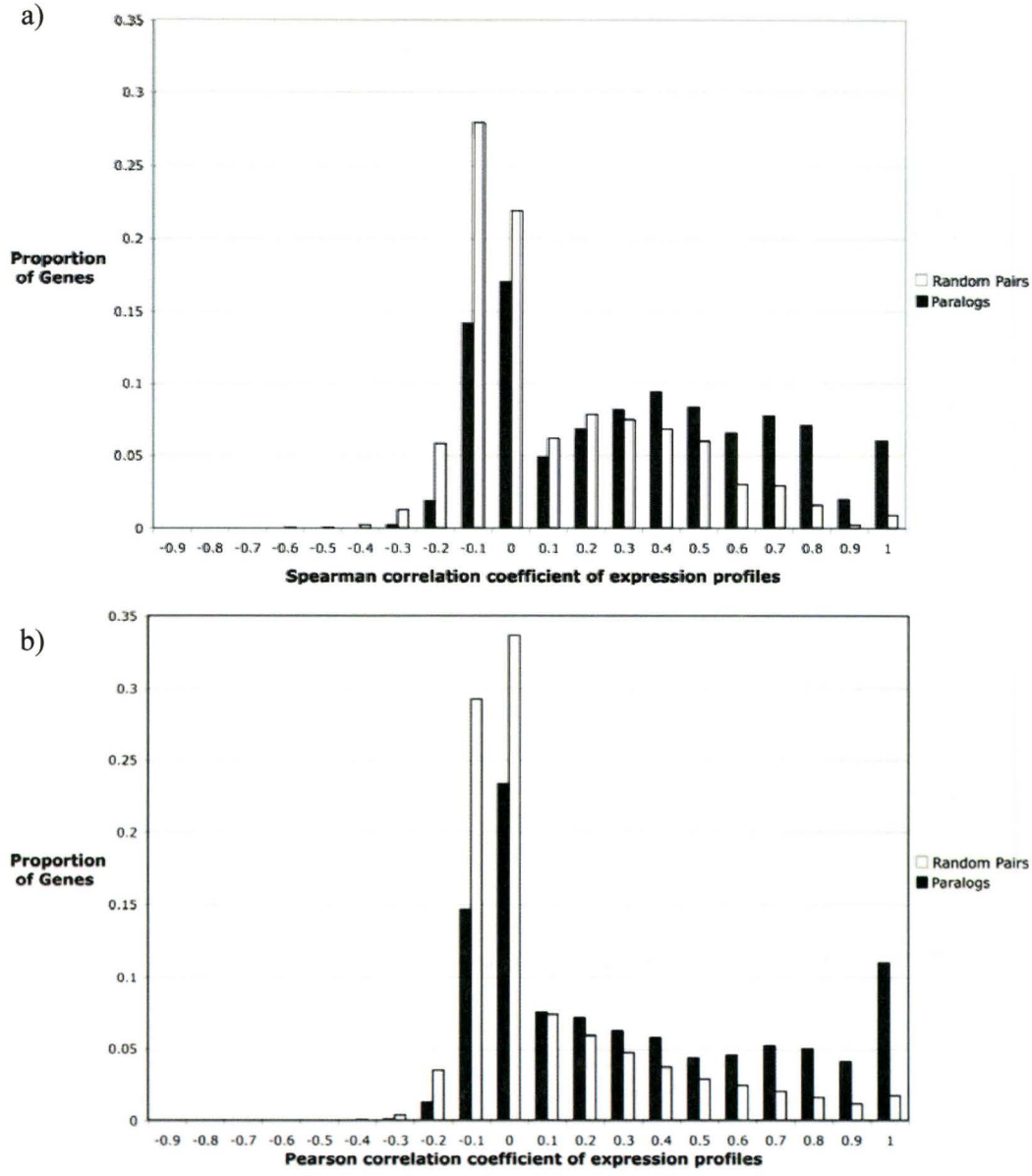


Figure 3.5. Distribution of (a) Spearman's rank correlation coefficient ρ and (b) Pearson's product-moment correlation coefficient R between gene expression profiles across 14 adult tissues and 4 larval stages. Correlation coefficient bins (x-axis) between 2,675 duplicate genes (black) and 50,000 random pairs of genes (white). The distribution of coefficients between 50,000 random pairs represents a null expectation of correlations. The 95th percentile of random pair genes corresponds to (a) ρ=0.612 and (b) R=0.665, under which (a) 78.5% and (b) 78.2% of the duplicate genes also fall below; approximately three quarters of duplicate genes have diverged in expression as much as random pairs.

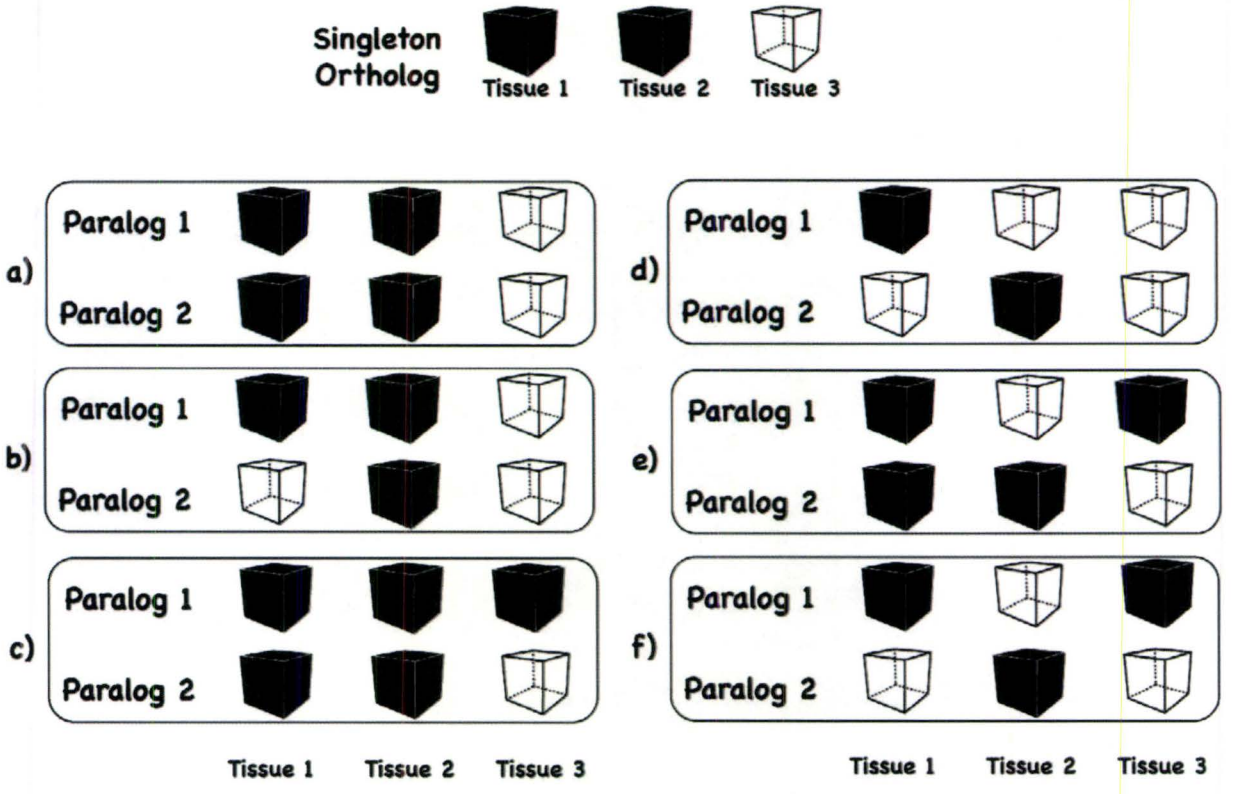


Figure 3.6. Duplicate genes were categorized into six classes based on their expression profiles and the expression profile of their ortholog. Black boxes represent expression within a tissue (an EST was found), white boxes represent no expression within a tissue (no ESTs were found). Examples of profiles consistent with (a) Overlapping profiles (b) Overlapping with loss of expression, (c) Overlapping with gain of expression, (d) RSF (regulatory subfunctionalization), (e) RNF (regulatory neofunctionalization), and (f) RSNF (regulatory subneofunctionalization). (a) Duplicate genes and their ortholog had completely overlapping expression profiles, without any loss or new expression compared to the ortholog. (b) Duplicate genes had overlapping profiles with loss of expression or (c) gain of expression, relative to the ortholog. (d) Duplicate genes had complementary loss of expression in at least a pair of tissues that were also expressed in the ortholog, without any new expression compared to the ortholog. (e) One duplicate copy had overlapping expression compared to the ortholog, while the other copy had lost expression in at least 1 tissue (in which the other copy had expression), in addition to having gained expression in at least 1 tissue compared to the ortholog. (f) Complementary expression loss as in (d), but with at least 1 of the duplicate copies gaining new expression in at least 1 tissue compared to the ortholog. We are unable to identify whether paralogs categorized in (f) had first lost expression domains (SF) or gained a novel expression domain (NF).

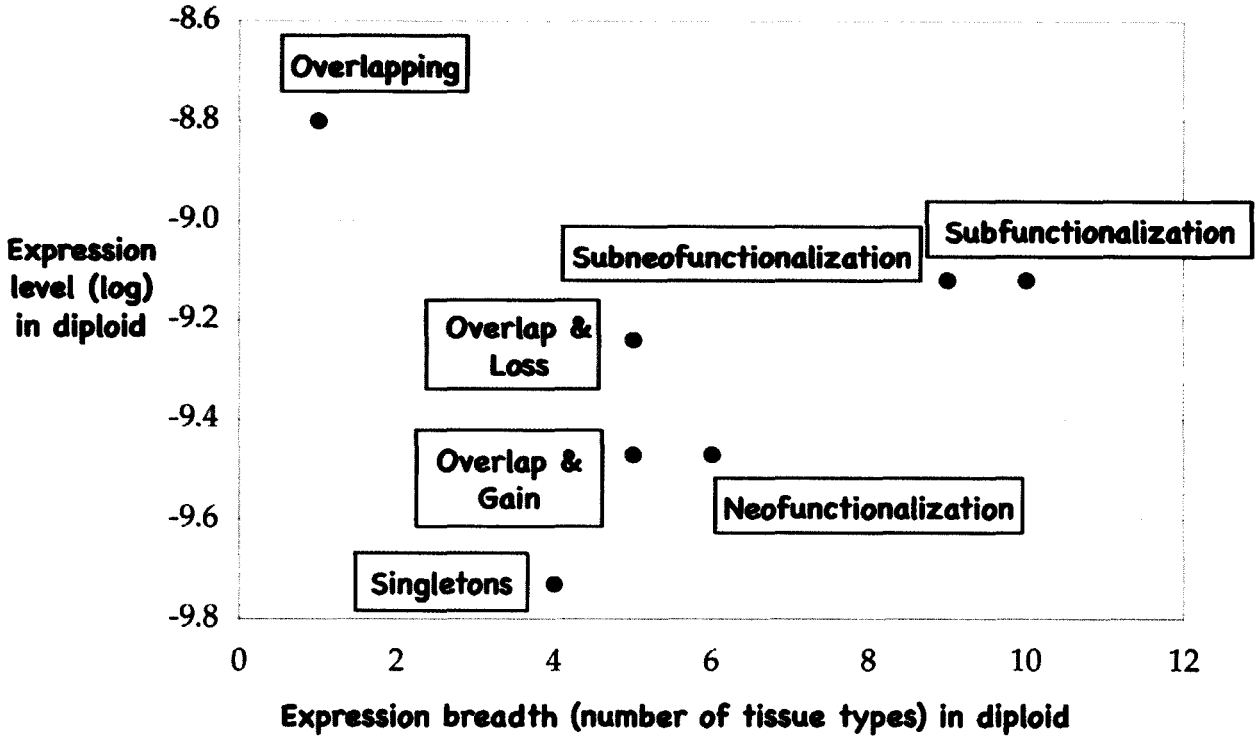


Figure 3.7. Median expression characteristics (log expression level and expression breadth) of orthologs of different classes of post-duplication genes can be distinct. Genes that are highly expressed and found in many tissues often partition their expression (subfunctionalization and subneofunctionalization). Genes that have medium breadth but are not highly expressed often gain new tissues of expression (neofunctionalization and overlap with gain), and those that are highly expressed tend to lose expression in one copy (overlap with loss). A few genes with low breadth were found with completely redundant expression profiles (overlapping). Genes that are found in singleton after WGD may have medium breadth of expression, but normally have low expression level.

CHAPTER 4

Single-species microarrays and comparative transcriptomics

Chain F.J.J., D. Ilieva, and B.J. Evans (2008). *PLoS ONE* 3(9): e3279.

PREFACE

The previous chapters focused on the evolution and expression of duplicate genes in a polyploid genome. In an effort to explore duplicate gene expression differences between polyploid species and their impacts on hybrid misexpression, we evaluated gene expression between three polyploid species and their hybrids using microarrays. Upon further analysis of the data, we found technical problems comparing transcriptomes of divergent genomes using a microarray designed for only one of the species.

ABSTRACT

Prefabricated expression microarrays are currently available for only a few species but methods have been proposed to extend their application to comparisons between divergent genomes. Here we demonstrate that the hybridization intensity of genomic DNA is a poor basis on which to select unbiased probes on Affymetrix expression arrays for studies of comparative transcriptomics, and that doing so produces spurious results. We used the Affymetrix *Xenopus laevis* microarray to evaluate expression divergence between *X. laevis*, *X. borealis*, and their F1 hybrids. When data are analyzed with probes that interrogate only sequences with confirmed identity in both species, we find results that differ substantially from analyses that use genomic DNA hybridizations to select probes. Our findings have implications for the experimental design of comparative expression studies that use single-species microarrays and for our understanding of divergent expression in hybrid clawed frogs. These findings also highlight important limitations of single-species microarrays for studies of comparative transcriptomics of polyploid species.

INTRODUCTION

Microarrays designed for one species have been used to explore expression divergence between species (Becher et al. 2004; Cáceres et al. 2003; Enard et al. 2002; Khaitovich et al. 2004; Meiklejohn et al. 2003; Michalak and Noor 2003; Nuzhdin et al. 2004; Ranz et al. 2003; Uddin et al. 2004; Weber et al. 2004). These studies deploy different types of microarrays on species with varying levels of divergence, and these experimental variables influence the potential for technical bias. In particular, the designs of experiments that deploy two-color versus one-color microarrays differ, and therefore can be differently subject to technical bias when these arrays are used to compare expression between species. Microarrays

with short oligonucleotide probes might be more profoundly impacted by a single base pair mismatch than ones with longer oligonucleotides. Additionally, studies of species that are substantially diverged have more sequence differences and other possible sources of variation (alternative splicing, repetitive elements, duplications, etc.) that increase the chance of technical bias. Differences in technical procedures between laboratories and genetic differences among populations or individuals can also contribute to variation in expression divergence.

In theory, if the "target" species for which the array was designed and a "non-target" species are closely related, some probes on the array should be able to interrogate expression of genes in both species without bias if the sequences that are interrogated by the probes are still the same in both species (Chismar et al. 2002; Grigoryev et al. 2005; Ji et al. 2004). Some studies have attempted to identify and eliminate probes with biased response to the transcriptome of the target and non-target species. One tactic is to select probes on the basis of genomic DNA (gDNA) hybridizations of the target and a non-target species to the microarray chip (Hammond et al. 2005; Hammond et al. 2006; Ranz et al. 2003). If the same amount of gDNA is used in the hybridization, probes that match conserved regions should hybridize with similar intensity to gDNA in both species. Recently, for example, the *Xenopus laevis* Affymetrix microarray chip was used to explore expression divergence between different species of clawed frogs and their hybrids (Malone et al. 2007; Malone et al. 2006; Malone and Michalak 2008a). Comparisons were made between testis and ovary expression profiles of the target species, *X. laevis* (XL), a non-target species, *X. muelleri* (XM), and F1 hybrids from a cross between a XL female and a XM male (hereafter H_{XLXM}). In these studies, hybridizations of gDNA of XM and XL were performed on the XL microarray, and probes whose XM/XL genomic hybridization intensity ratio (gDNA ratio) was not between 0.99 and 1.01 or between 0.99 and 1.10 were excluded from the analysis (Malone et al. 2007; Malone and Michalak 2008a). When expression profiles of testes or ovaries of XL and XM were compared to the same tissue in their hybrids, widespread dominance in expression was reported in hybrids wherein the expression profile of H_{XLXM} tended to be more similar to XL than to the non-target parental species XM (Malone et al. 2007; Malone and Michalak 2008a). About 28 times more genes were significantly divergently expressed in testes in a comparison between XM and H_{XLXM} than between XL and H_{XLXM} (Malone et al. 2007) and about 4.5 times more genes were significantly divergently expressed in ovary in a comparison between XM and H_{XLXM} than between XL and H_{XLXM} (Malone and Michalak 2008a).

With a goal of further exploring these results, we analyzed new expression data from testis tissue of XL, *X. borealis* (XB), and F1 hybrids between XL x XB (XL female and XB male, hereafter H_{XLXB}). XB and XM are equivalently diverged from XL (Evans 2007; Evans et al. 2005; Evans et al. 2004) so our new data provide a phylogenetically meaningful comparison. All of these species are "pseudotetraploid" in that they are historically tetraploid but their genomes have diploidized (bivalents form at meiosis; each chromosome has only one homologous chromosome). XL and (XB+XM) diverged from a common tetraploid ancestor

roughly 21 – 41 million years ago, and XB and XM diverged from a common ancestor roughly 14 – 25 million years ago (Chain and Evans 2006; Evans 2007; Evans et al. 2005; Evans et al. 2004). In the analysis of these new expression data, we included only those probes that interrogate sequences that are identical in XL and in XB based on 454 pyrosequencing of XB cDNA. For comparative purposes, we also performed genomic DNA hybridizations on XL, XB, and XM, and analyzed the new data and also data from other studies (Malone et al. 2007; Malone et al. 2006; Malone and Michalak 2008a; Malone and Michalak 2008b; Sinner et al. 2006) using microarray probes selected using the gDNA hybridization approach of (Malone et al. 2007; Malone and Michalak 2008a; Malone and Michalak 2008b).

RESULTS

Affymetrix *Xenopus laevis* microarray, probemasks, and tissue comparisons.

This study examines expression data collected from a prefabricated *Xenopus laevis* microarray – the Affymetrix GeneChip® *Xenopus laevis* Genome Array. This microarray interrogates over 14400 transcripts. A transcript is interrogated with a set of 16 probes, which is called a "probeset". Each probe within a probeset is an oligonucleotide 25 base pairs in length that hybridizes to a unique portion of an XL transcript. For each species or hybrid in this study, three biological replicates (different individuals) were performed per tissue. Hereafter we refer to the replicated expression data from a single tissue from one species or one type of hybrid (either H_{XLXB} or H_{XLXM}) as a "treatment".

Probemasks are lists of genes that are defined *a priori* to be excluded from analysis (before microarray normalization is performed). In this study, we analyzed data using two types of probemasks. The first type of probemask excluded all probes except those that interrogated sequences that we confirmed were identical in XL and XB, as in (Bar-Or et al. 2006; Khaitovich et al. 2004). We used BLAST (Altschul et al. 1997) to identify probes on the Affymetrix GeneChip® *Xenopus laevis* Genome Array that perfectly match sequences in XB that we obtained using 454 pyrosequencing of normalized XB testis cDNA. Normalization of XB testis cDNA (which is a procedure different from and unrelated to normalization of microarray data) was performed prior to 454 pyrosequencing in order to increase representation of genes with low expression; procedures for cDNA normalization and 454 pyrosequencing are described elsewhere (Chain et al. 2008). The resulting probemask included 5268 probes in a total of 2143 probesets, for an average of 2.458 probes per probeset. Hereafter we refer to this probemask as the "XB+XL perfect match probemask". According to a permutation test in which the same number of probes are assigned to probesets randomly one thousand times, this average number of probes per probeset is significantly higher than random expectations ($P < 0.001$; the mean number of probes per probeset of the permutations was 1.169 and the 95% confidence interval was 1.158 – 1.180). This is consistent with the notion that some genes are conserved across multiple regions that are interrogated by unique probes on the microarray, resulting in significantly more

probes per probeset than random expectations. Despite this biologically relevant pattern, we note that the overall low number of probes per probeset is likely to be associated with more variation in expression intensities than is typical of Affymetrix probesets with 16 probes. Furthermore, because the perfect match probes identified in XB are based on 454 pyrosequencing of normalized testis cDNA, this analysis might be biased in favor of genes that are expressed in testis of this non-target species. Additionally, because we retain only those probes that are identical in XL and XB, this analysis probably is also biased towards slowly evolving genes – or at least genes that have slowly-evolving regions that are interrogated by probes on the microarray.

The second type of probemask was generated based on the non-target to target hybridization ratio of genomic DNA (the gDNA ratio) of XL, XB, and XM as in (Malone et al. 2007; Malone and Michalak 2008a). These probemasks include only those probes with a non-target/target gDNA ratio between 0.99 and 1.1, and hereafter we refer to them as the "XB/XL gDNA probemask" and the "XM/XL gDNA probemask", respectively. The XB/XL gDNA probemask included a total of 1792 probes in 1672 probesets, for an average of 1.072 probes per probeset. This average is similar but still significantly higher (P=0.003) than random expectations according to a permutation test, which had an average of 1.055 (95% confidence interval 1.045 – 1.067). This average is significantly lower than the average of the XB+XL perfect match probemask ($P < 0.001$, permutation test). Only 2.5% of the probes (45 out of 1792 probes) that were retained by the XB/XL gDNA probemask are also retained by the XB+XL perfect match probemask. Less than 1% of the probes (45 out of 5268 probes) that were retained by the XB+XL perfect match probemask were also retained by the XB/XL gDNA probemask.

The XM/XL gDNA probemask included a total of 12888 probes in 8721 probesets and an average of 1.478 probes per probeset. This average is also similar but still significantly higher than the corresponding average of the random permutations of 1.448 ($P < 0.001$; 95% confidence interval 1.437 – 1.460). For comparison, the probemask of Malone *et al.* (2007) included 11485 probesets with an average of less than 2 probes per probeset.

Using both types of probemask (the XB+XL perfect match probemask and the XB/XL gDNA probemask), we evaluated interspecific expression divergence in testis between H_{XLXB} and each parental species (XL or XB) and in brain between XL and XB. We also used both of these probemasks to evaluate intraspecific expression divergence between various XL tissues (egg, tadpole stage 11, ovary, testis, and brain). Additionally, we used the XM/XL gDNA probemask to evaluate expression divergence in testis and ovary between H_{XLXM} and each parental species (XL or XM) and we used this same probemask to evaluate intraspecific expression divergence between the aforementioned XL tissues. We were not able to perform interspecific analyses between XL and XM with a perfect match probemask because sequence data from XM was not obtained.

Dominant expression in hybrids?

When we analyzed testis expression data from XL, XB, and H_{XLXB} using the XB+XL perfect match probemask, expression divergence between XL and H_{XLXB} is slightly less than between XB and H_{XLXB} but similar in terms of the number of significantly divergently expressed genes. Out of 2143 probesets included in this analysis, 182 genes are significantly upregulated in XL testis compared to H_{XLXB} testis whereas 210 genes are significantly upregulated in H_{XLXB} testis compared to XL testis. 280 genes are significantly upregulated in XB testis compared to H_{XLXB} testis whereas 245 genes are significantly upregulated in H_{XLXB} testis compared to XB testis. The number of significantly upregulated genes in each parental species compared to H_{XLXB} is significantly higher in the comparison with XB than the comparison with XL (182 versus 280, $G = 20.95$, $P < 0.001$, two-sided test). But the number of significantly upregulated genes in H_{XLXB} compared to each parental species is not significantly different (210 versus 245, $G = 2.69$, $P = 0.20$, two-sided test). Therefore, the difference in the number of significantly divergently expressed genes in each comparison between a parental species and hybrids is attributable to more genes being upregulated in XB compared to H_{XLXB} than are upregulated in XL compared to H_{XLXB}. Thus, the proportion of divergently expressed genes in XB testis compared to H_{XLXB} testis is about 1.34 times as large as the proportion of divergently expressed genes in XL testis compared to H_{XLXB} testis (Table 4.1). But, as mentioned earlier, some or all of this bias could be because we retained probes in this analysis based on sequences of genes that are expressed in XB testis.

While this 1.34 fold asymmetry in divergent expression between the parental species and hybrids is significant (525 versus 392 genes, $G = 19.36$, $df = 1$, $P < 0.001$), it is in sharp contrast with the 28 fold difference reported in comparisons between testis tissue of XL, XM, and H_{XLXM} where 3995 genes were divergently expressed between XM and H_{XLXM} but only 142 genes were divergently expressed between XL and H_{XLXM} (Table 4.1; Malone et al. 2007). The difference in the proportion of divergently expressed genes in this study compared to Malone *et al.* (2007) is significant. More specifically, a re-sampling test (see Methods) indicates that there is a significantly higher proportion of divergently expressed genes between XL and H_{XLXB} using the XB+XL perfect match probemask than were reported between XL and H_{XLXM} by Malone *et al.* (2007) using a gDNA probemask ($P < 0.001$). Likewise, there is a significantly lower proportion of divergently expressed genes between XB and H_{XLXB} using the XB+XL perfect match probemask than were reported between XM and H_{XLXM} by Malone *et al.* (2007).

With respect to misexpression – which we define as hybrid expression that is not intermediate with respect to the expression of each parental species – using the XB+XL perfect match probemask, we find that only 13 genes are significantly upregulated in testis of H_{XLXB} with respect to testis of both XL and to XB and that 16 genes are significantly upregulated in testis of XL and XB with respect to testis of H_{XLXB}. This difference is not significant ($G = 0.31$, $df = 1$, $P = 0.58$).

Comparison of gDNA hybridizations within and between species

To further explore the basis of the discrepancy in the level of asymmetry of divergent expression revealed by our results using the XB+XL probemask and previous studies, we re-analyzed testis expression data from XL, XB, and H_{XLXB} using the XB/XL gDNA probemask that was based on our new gDNA hybridizations. We also re-analyzed testis and ovary expression data from XL, XM, and H_{XLXM} using the XM/XL gDNA probemask that was based on our new gDNA hybridizations.

We compared our gDNA hybridizations to those of Malone *et al.* (2007) and Malone and Michalak (2008a). We ranked all of the probes on the chip by the gDNA hybridization intensity and then divided these ranks into 25 bins. Comparison to the gDNA ratio of each probe indicates that the median intensity of hybridization was lower in the non-target species (XM or XB) than the target species (XL) for most bins (Figure 4.1). Probes with a gDNA ratio near one tended to have lower gDNA hybridization intensities in both the non-target and the target species than other probes on the chip, and the target species (XL) tends to have a more dynamic relationship between probe intensity and the gDNA ratio. Thus, at least on the Affymetrix GeneChip® *Xenopus laevis* Genome Array, probe selection on the basis of a gDNA hybridization ratio near one appears to have an unintended consequence of retaining probes with low gDNA hybridization intensities in both species. This was true in gDNA hybridizations performed by our lab and also by another lab (Figure 4.1), thus it is not attributable to differences in laboratory procedure.

Our XB gDNA hybridization was less intense than our XM hybridization even though we attempted to control for the amount of gDNA used in the hybridization, and even though these species are equally diverged from XL (Figure 4.1). This variation probably is technical in nature and underscores the challenge of generating comparable gDNA hybridizations for different species. Below we report results derived from analyses based on our gDNA hybridizations for XL, XB, and XM; as detailed below, these results are qualitatively similar to those revealed with the gDNA probemask of Malone *et al.* (2007) and Malone and Michalak (2008a).

Is the ratio of genomic DNA hybridization a reliable way to detect perfect match probes on the *Xenopus laevis* Affymetrix chip?

Probes that perfectly match sequences from XL and XB have a wide range of XB/XL gDNA ratios (Figure 4.2A). Under a best-case scenario, this indicates that using the gDNA ratio as a criterion for probe retention will not retain all perfect match probes. But we also found that other probes that we know mismatch both paralogs of genes in XB have a range of XB/XL gDNA ratios that overlaps extensively with the gDNA ratios of probes that perfectly match both species (Figure 4.2B). This point is also illustrated by examination of four probesets on the *Xenopus laevis* Affymetrix microarray that were designed to interrogate XB transcripts: XlAffx.1.5.S1_at, XlAffx.1.9.S1_at, XlAffx.1.10.S1_at, and XlAffx.1.12.S1_at. Sixty out of the 64 probes in these four probesets do not

perfectly match XL, and these also have a broad range of gDNA ratios (Figure 4.2A). Together these observations indicate that gDNA ratios provide a poor basis for selection of perfect match probes in non-target species on the Affymetrix GeneChip® *Xenopus laevis* Genome Array. In addition to not retaining many probes that perfectly match both species, this approach almost certainly results in the retention of probes that do not perfectly match the non-target species.

Does it matter if some probes with differential performance between treatments are included in the analysis?

When testis expression data from XL, XB, and H_{XLXB} are analyzed using our XB/XL gDNA probemask or using our XM/XL gDNA probemask, we find similar results to the analysis of testis expression data from XL, XM, and H_{XLXM} by Malone *et al.* (2007). This suggests that evolutionary differences between XB and XM, possible differences in the geographic origin of XL animals, and variation in laboratory procedures associated with microarray hybridizations together had a much smaller impact on the results than did the type of probe mask used in the analysis. More specifically, in this analysis the asymmetry in expression divergence is significant and more substantial than results from the XB+XL perfect match probemask such that expression in the hybrid appears much more similar to the target than the non-target species (Table 4.1). This is because using a gDNA probemask instead of a perfect match probemask results in a significantly lower proportion of genes that are divergently expressed in the comparison between XL and H_{XLXB} and a significantly higher proportion of genes that are divergently expressed between XB and H_{XLXB} ($P \leq 0.002$ for both comparisons).

We explored alternative analytical approaches including invariant set (IS) normalization (Li and Wong 2001) and the probe logarithmic intensity error (PLIER) method for calculating signal intensity (Affymetrix 2001). These procedures produce results that are qualitatively similar to those found with RMA normalization with each probemask. The asymmetry in divergent expression in testis between each parental species and the hybrid with the XB+XL perfect match probemask is of similar magnitude in each of these analyses (1.34, 1.45 and 1.39 for RMA, IS, and PLIER, respectively). Likewise, more than twice as much asymmetry in divergent expression in testis is found when RMA, IS, or PLIER normalization are used with gDNA probemasks (i.e. there are more divergently expressed genes between the non-target species and the hybrid than between XL and the hybrid with these probemasks; data not shown). Thus we conclude that the method of normalization also does not account for the substantial differences in results that are obtained from perfect match versus gDNA probemasks.

Rank difference

The nature of the discrepancy between results obtained from these different probemasks is further illuminated by consideration of some of the technical aspects of the analysis. When microarray data are normalized it is generally assumed that the overall distribution of expression intensities within each treatment is similar

(Bolstad et al. 2003; Smyth and Speed 2003; Yang et al. 2002). Moreover, most normalization methods were developed for comparisons between treatments with expression divergence at only a few genes (Gilad and Borevitz 2006). When data are normalized with the quantile method (Bolstad et al. 2003), for example, which was used in this study and in Malone *et al.* (2006), Malone *et al.* (2007) and Malone and Michalak (2008a), the expression intensities of each probe are ranked and replaced by the average intensity of each quantile (each rank). This procedure yields identical distributions of overall expression intensities across treatments, even if they were very different to begin with.

If the overall distribution of expression intensities was similar in each treatment before normalization, it is reasonable to expect that the magnitude and direction of expression divergence should be unbiased – that for a given magnitude of expression divergence, a similar number of genes will be upregulated in one treatment as is upregulated in the other. To test this, we calculated the difference in expression rank for each gene included in the analysis, with the lowest rank corresponding to the gene with the lowest expression as depicted in Figure 4.3. Additionally, the skew of this distribution was quantified by the Pearson skewness coefficient (= 3*(mean-median)/standard deviation). Departure of the observed median rank difference and skew of the distribution of rank differences from the null hypotheses of a median and skew of zero was assessed by comparison to a null distribution generated from 1000 randomized ranks using scripts written in PERL.

When interspecific data from the target species and a non-target species were analyzed using a gDNA probemask, the median rank difference was negative and this median departed significantly and substantially from zero (Table 4.2). The skew of the distribution of rank differences was significantly and substantially positive in these interspecific comparisons (Table 4.2). While these metrics are not independent because the median is used in the calculation of skew, they provide qualitative information about the rank difference distributions in these analyses. Because we calculated the rank difference by subtracting the non-target rank from the target rank, a negative median indicates that the non-target sequences tend to be upregulated to a greater degree than do the target sequences. A positive skew of this distribution (Table 4.2) indicates a tail on the right, suggesting that some probesets have a much higher rank (higher expression) in XL but not the reverse.

In contrast, when intraspecific comparisons were analyzed with gDNA probemasks, the median and skew never departed as substantially from the null expectation as the interspecific comparisons between a target and non-target species, although occasionally the intraspecific departure was significant (Table 4.2). When the XB+XL perfect match probemask was used in the analysis, the median and skew were not significantly different from the null expectation (Table 4.3). While occasional departure from the null in some intraspecific comparisons between different XL tissues probably has a biological basis and could also stem from variation between laboratories in microarray protocol, these comparisons suggest that the substantially negative median and positive skew of the rank difference in

interspecies comparisons analyzed with gDNA probemasks has a technical rather than a biological basis.

Spearman rank correlation

When gDNA probemasks are used, we suspected that differential performance of some probesets in the non-target species could cause a spurious signal of upregulation *and* downregulation compared to another species (Figure 4.3). One class of significantly differently expressed genes – those that appear to be upregulated in the target species (XL) – could result when probes hybridize poorly to transcripts of the non-target species. The other class of significantly differently expressed genes – those that appear to be upregulated in the non-target species (XB or XM) – could result when the ranks of some genes in the non-target species are elevated as a result of the other genes that are interrogated by biased probes having a lower rank (Figure 4.3). A key difference between these two classes of divergently expressed genes is that a larger proportion of the genes that appear upregulated in XL are interrogated by probes with differential performance (bias) between species. In analyses with a gDNA probemask, therefore, we predicted that the expression rank of genes that appear to be significantly upregulated in the non-target species would be highly correlated with the expression rank of these genes in the target species. We expected this correlation to be much higher than the correlation between the ranks of genes upregulated in the target species and the rank of these same genes in the non-target species.

To test this, we calculated the Spearman's rank correlation (SRC) of the rank in each treatment of (i) genes upregulated in the non-target species and (ii) genes upregulated in the target species. Under our hypothesis that many of the genes that are upregulated in the non-target species are false positives, we expected that the SRC would be much higher in (i) than in (ii). To quantify this expectation, we calculated the absolute value of the difference in the SRC in (i) and (ii) for the interspecies comparisons, and we refer to this difference as δSRC. For comparative purposes, δSRC was calculated for interspecific comparisons between XL and a non-target species, comparisons between each species and a hybrid, and intraspecific comparisons between different tissues of XL, and this was performed for analyses with each type of probemask.

The data support our expectation. When the XB/XL gDNA probemask or the XM/XL gDNA probemask are used in interspecific comparisons, the δSRC of the rank of genes upregulated in the non-target species is substantially higher than that of genes upregulated in the target species or in hybrids (Table 4.2). When comparisons were made between tissue types in XL or within a tissue type of XL and a hybrid using these gDNA probemasks, extreme differences between δSRC of each of these classes of genes were not observed (Table 4.2). A high δSRC was not observed in any of the analyses with the XB+XL perfect match probemask (Table 4.3). Furthermore, we found other signs of technical bias in results generated with

gDNA probemasks, but not the XB+XL perfect match probemask, by comparing the mean rank of significantly upregulated genes (Table 4.4 and 4.5).

Taken together, these observations are consistent with the notion that the use of probemasks based on gDNA ratios on the Affymetrix GeneChip® *Xenopus laevis* Genome Array produces spurious results when comparisons are made directly between species or between a non-target species and a hybrid, irrespective of tissue type. When gDNA probemasks are used, many of the genes that are putatively upregulated in the non-target species are actually false positives whose high ranks are an artifact of the low ranks of poorly performing probesets. Of course, this group of genes may include some genes that are not false positives, but it is not clear which ones these are. We suspect then, albeit with caveats discussed below, that our analysis with the XB+XL perfect match probemask is a closer approximation of biological variation than that found by Malone *et al.* (2007) and Malone and Michalak (2008a).

DISCUSSION

Probe selection by genomic hybridization

A challenge to the implementation of single-species microarrays in comparative transcriptomics is the identification of unbiased probes. Due to differences from the target species, such as sequence divergence, non-target transcripts will exhibit a range of probe hybridization efficiencies that cause technical variation in hybridization intensities. In comparative analyses, normalization may overcompensate for genes with lower than average divergence and undercompensate for genes with higher than average divergence (Gilad et al. 2005). Exacerbating this problem, our analysis of confirmed perfect match probes in a target and a non-target species illustrates that the gDNA ratio is an unreliable metric with which to identify unbiased probes on the Affymetrix GeneChip® *Xenopus laevis* Genome Array. This approach selects probes with low gDNA intensity (Figure 4.1), misses probes that do perfectly match both species (Figure 4.2A), and includes probes that do not perfectly match both species (Figure 4.2B). The implications of this are large and affect fundamental conclusions of the analysis, such as which and how many genes are significantly or not significantly differently expressed. Notably, our analyses suggest that including biased probes in a microarray analysis leads not only to spurious results from these biased probes, but affects conclusions drawn from probes that are interrogated by probes that perform equally well in both species. We anticipate, therefore, that comparisons between species using probes that are selected by gDNA ratios, including the comparison between XB or XM and XL that are presented here, are characterized by a high level of false positives as well as false negatives. Many of the genes from this type of analysis that are putatively upregulated in the target species are actually interrogated by probes that do not perform equivalently in the non-target species. Many of the genes that are putatively upregulated in the non-target species are actually genes whose ranks have been elevated as an artifact of other probes that do not perform

equivalently in the non-target species. It is therefore not only necessary to retain as many perfect match probes as possible, but also to exclude biased probes from microarray analyses.

Gene duplication

Another concern with the application of this microarray to non-target clawed frog species relates to whole genome duplication. Because XL, XB and XM are tetraploid, asymmetry in cross-hybridization between paralogous transcripts could influence results. For example, a probe might hybridize to only one paralog in one species but to both paralogs of genes in another species, either as a result of sequence divergence or because both are expressed in one but not the other. This problem is aggravated by species-specific pseudogenization. Estimates of the percent of duplicated genes in XL that are still expressed (not pseudogenes) range from 77% (Hughes and Hughes 1993) to a probably more accurate estimate of less than 50% (Hellsten et al. 2007; Sémon and Wolfe 2008). Divergence of the ancestor of XL from the ancestor of (XM+XB) occurred about halfway between the time of whole genome duplication and the present (Chain and Evans 2006; Chain et al. 2008; Evans et al. 2004). For this reason, the frequency of expressed orthologous transcripts in XL and non-target species such as XB and XM is far below 100% as a result of "divergent resolution" – the retention of different (non-orthologous) paralogs of genes in each species (Werth and Windham 1991).

That Affymetrix microarrays do not effectively discriminate between different but closely related duplicated genes has been suggested for allopolyploid wheat (Poole et al. 2007). However, we performed a power analysis that indicated that probes on the XL microarray performed consistently in distinguishing expression of each paralog after the application of probemasks with different specificities for a target paralog (i.e. varying numbers of mismatches to the non-target paralog; Chain et al. 2008). But within *Xenopus*, orthologs are more similar to each other than are paralogs derived from genome duplication because genome duplication occurred before speciation. Orthologous but not identical sequences (from different species) thus have greater potential to be able to hybridize strongly but not equivalently to probes directed against one XL paralog, than do co-expressed paralogs within XL. These concerns are relevant to all of the analyses presented here, including the ones that use the XB+XL perfect match probemask.

CONCLUSIONS

Previous work has explored factors in addition to sequence divergence that influence probe hybridization efficiency in different species, such as variation in labeling, overlap of oligonucleotide probes, alternative splicing, sequence homology to non-target transcripts, insertion/deletion differences, and intraspecific polymorphism (Cambon et al. 2007; Gilad and Borevitz 2006; Gilad et al. 2005; Hsieh et al. 2003; Kirst et al. 2006; Poole et al. 2007; Zakharkin et al. 2005). Some or all of these variables might be at play here – sequence divergence, for example,

has already been shown to influence microarray hybridization efficiency in clawed frogs (Sartor et al. 2006). While sequence mismatches might not substantially affect the ability of microarrays to detect misexpression (Cope et al. 2004; Oshlack et al. 2007), it seems probable that sequence mismatches could cause bias if it varies among treatments such as when expression of two species are compared using an array designed for only one of them. Therefore, an experimental design that has consistent bias across treatments, in which one compares 'apples to apples' (Buckley 2007), has the potential to provide useful information from non-target species. Examples of more appropriate experimental designs include (a) using a microarray designed for another species with a non-target species but only comparing intraspecific expression levels within the non-target species, (b) constructing custom arrays for each species (or hybrid) of interest, and (c) building a custom array with probes directed against each species (Oshlack et al. 2007). Another important measure for comparative analyses using single-species arrays is the validation of results using microarray-independent approaches, such as real-time quantitative PCR. The biases suggested by our analyses have implications for studies that deploy Affymetrix microarrays for interspecific comparisons (Malone et al. 2007; Malone et al. 2006; Malone and Michalak 2008a), and could also be a concern for expression studies of species or genes with population structure, high mutation rate, or large effective population size.

MATERIALS AND METHODS

Origin of animals

XB expression data, gDNA, and XB parents of H_{XLXB} were from or were animals from Kenya. The XL expression data, gDNA, and XL parents of H_{XLXB} were laboratory animals that probably are from Cape Province, South Africa, which is the source of most laboratory stocks (Tinsley and McCoid 1996). All of the H_{XLXB} individuals were from the same cross and are therefore full siblings. We did not analyze hybrid tissue from the reciprocal cross (from an XB female and XL male).

The XM expression data, gDNA, and parents of H_{XLXM} in Malone *et al.* (2007) and Malone and Michalak (2008a) were from animals collected in Swaziland, but the XM gDNA that we performed for gDNA hybridization originated from Tanzania. Within XM, mitochondrial DNA variation between these localities is very low so we do not anticipate substantial levels of intraspecific variation in the nuclear genome of this species compared to XL (Evans et al. 2004).

Microarray hybridizations and comparisons

We performed new expression analyses on testis and brain tissue from XL, XB, and H_{XLXB}. For each tissue from each species or hybrid, RNA was isolated using TRIzol® Reagent (Invitrogen Life Technologies) according to the manufacturer's protocol, purified with RNeasy Mini Kit (Qiagen), and its integrity assessed on an Agilent BioAnalyzer. Two micrograms of total RNA was used to

prepare biotin labeled cRNA probes, which were subsequently hybridized to Affymetrix *Xenopus laevis* expression arrays following the manufacturer's protocol.

We performed new gDNA hybridizations using gDNA from XL, XB, and XM and compared these to gDNA hybridizations on XL and XM that were performed by Malone et al. (2007). For our gDNA hybridizations, five micrograms of gDNA from each species was fragmented with Dpn I at 37°C for 3 hours. Fragmented gDNA was purified with Qiagen PCR clean-up kit and the fragment distribution was checked on Agilent Bioanalyzer (Agilent) using the DNA 1000 assay. 50-100 nanograms of fragmented gDNA were then amplified using the BioPrime Labeling System (Invitrogen) following the manufacturers instructions. After completion of the Klenow Pol I catalyzed reaction, the distribution of PCR products was examined on Agilent Bioanalyzer with the DNA 1000 kit. The entire volume of the product (~50 µl) was used in the hybridization reaction on the Affymetrix *Xenopus laevis* Gene Chip. Hybridization, staining, washing and scanning were performed as described in the Expression Analysis Technical Manual. This protocol is similar to that used by Hammond et al. (2005).

After scanning, raw expression data were converted into CEL files using Microarray Analysis Suite version 5 (MAS 5, Affymetrix). For each pairwise comparison, CEL files were pre-normalized with the Robust Multichip Average (RMA) algorithm in RMAexpress (Bolstad 2007) using custom CDF files (probemasks) and the default parameters, which include a median polish and quantile normalization. The normalized data were used in the R statistical package following the protocol in Malone *et al.* (2007). An empirical Bayesian model was used to compute a moderated t-statistic using the limma package from Bioconductor (Smyth 2004). The TopTable function gave a P-value for differential expression for each gene that was adjusted using the Benjamini and Hochberg (1995) method to control for the false discovery rate. cDNA and gDNA hybridizations that we performed have been deposited in the Gene expression omnibus database (Edgar et al. 2002), GEO Series accession number GSE12625. We also analyzed other data from this database (GSM241082-4 (Malone and Michalak 2008b), GSM99995-7 (Sinner et al. 2006), GSM99980-2 (Sinner et al. 2006)). Expression data and genomic hybridizations from XL and XM testis and ovary that were not found in GEO were kindly provided by Pawel Michalak.

We used a re-sampling approach to test whether the proportion of divergently expressed genes in different analyses (each with a unique number of genes analyzed) were significantly different. Given two analyses with w and x genes of which y and z are significantly divergently expressed, respectively, using a PERL script we generated 1000 simulated datasets, each with w genes, by re-sampling a distribution of (w+x) total genes with (y+z) genes that are significantly divergently expressed. Where $(y/w) < (z/x)$, the two-sided probability of the null hypothesis of no difference is twice the proportion of these simulated datasets that had a proportion of divergently expressed genes lower than y/w (i.e. more different from z/x). Because some of the genes in these different analyses are the same and should therefore have correlated expression levels, the inclusion of these genes in

this comparison reduces the power to reject the null hypothesis, making this test conservative.

ACKNOWLEDGEMENTS

We thank Pawel Michalak for providing expression data from XL, XM and H_{XLXM} and genomic hybridizations from XL and XM. We also thank Jonathan Dushoff, Wilfried Haerty, John Hammond, Neil Graham, Pawel Michalak, Richard Morton, and the reviewers for advice on analyses, discussions, and comments on this manuscript, and Mohammad Iqbal Setiadi and David Anderson for assistance with rearing animals. This research was supported by the Canadian Foundation for Innovation, the National Science and Engineering Research Council, and McMaster University.

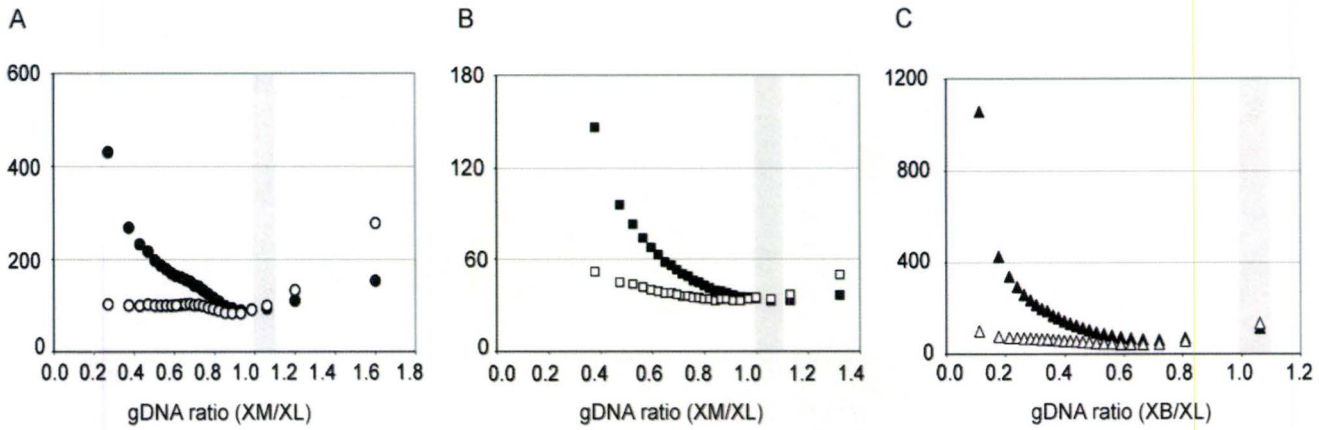


Figure 4.1. Genomic hybridization intensities (gDNA intensity) of XL, XB, and XM vary with respect to the non-target to target ratio of these intensities (gDNA ratio). This graph depicts the median gDNA intensities of all probes on the chip ranked by their gDNA ratio into 25 bins; each bin contains 10,000 probes except the 25th bin, which contains 7852 probes. The area in gray corresponds with the range of gDNA ratios of probes that are retained using the method of Malone et al. (2007). XL gDNA ratios are represented by filled symbols and non-target gDNA ratios are represented by unfilled symbols. Shown are relationships between the median gDNA intensity of each bin and the median gDNA ratio of each bin for (A) our XM and XL gDNA hybridizations, (B) the XM and XL gDNA hybridizations of Malone et al. (2007), and (C) our XB and XL gDNA hybridizations.

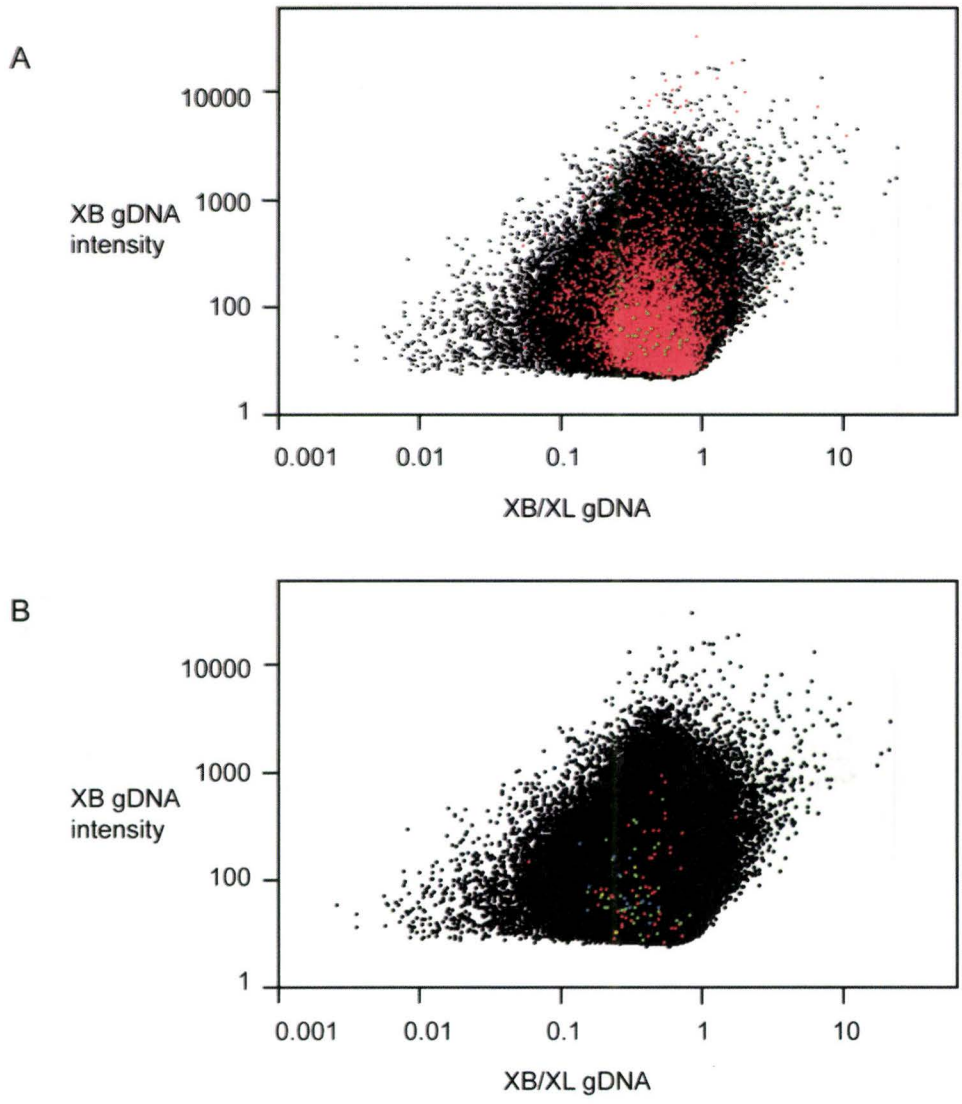


Figure 4.2. The gDNA ratio of probes that perfectly match (PM) XL and XB overlaps extensively with the gDNA ratio of probes that mismatch (MM) one species. (A) XB gDNA intensity versus gDNA ratio of PM probes in XL, XL and XB, and XB. PM probes in XL are in black, PM probes in XL and XB are in red, and PM probes in XB but not XL are in green. (B) XB gDNA intensity versus gDNA ratio of MM probes in XB. For comparative purposes, PM probes in XL are again in black. Probes that mismatch both paralogs of genes in XB with one, two, three, or four base pair differences are indicated in red, blue, green, and yellow respectively.

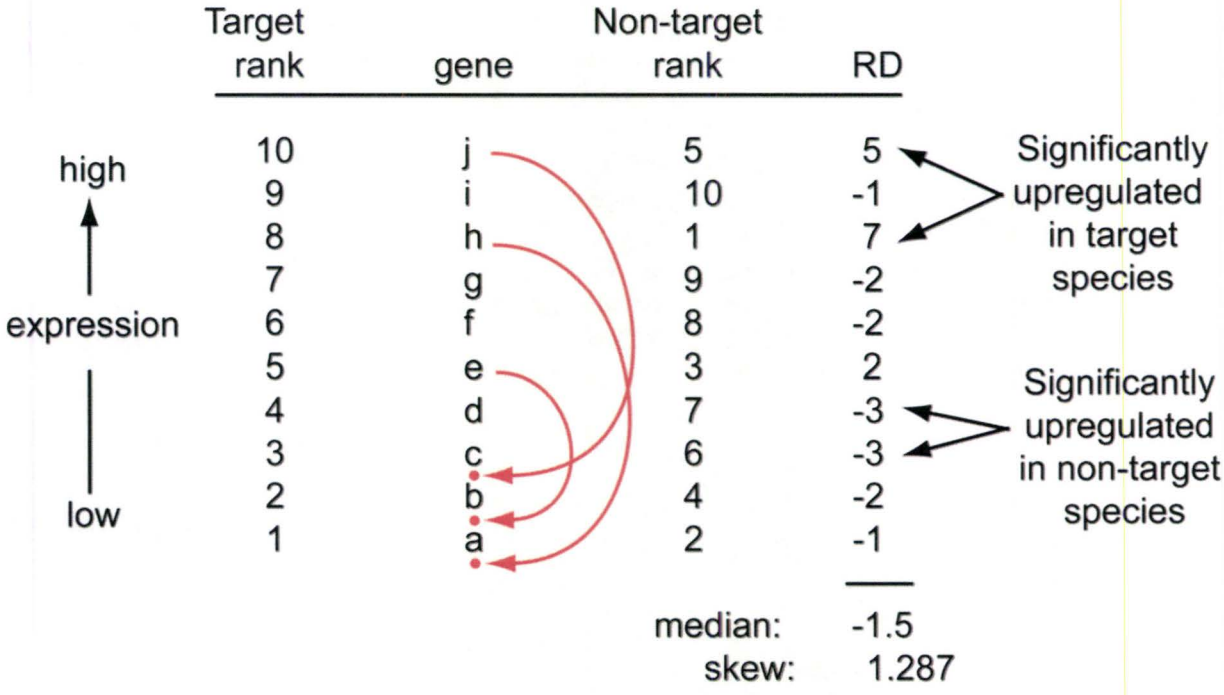
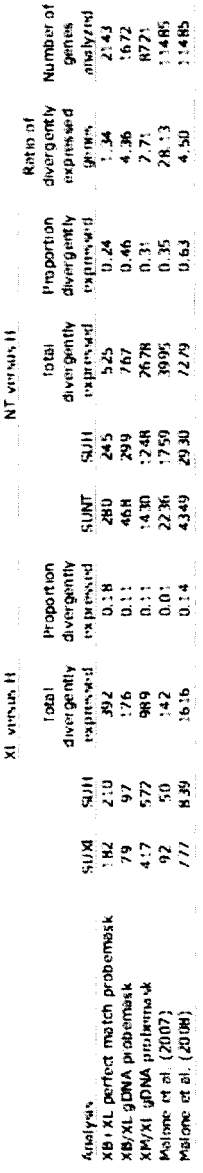


Figure 4.3. An example of how poor performance of a few probes in the non-target species can affect the rank of many genes, even ones that perform equally in both species. Ten genes (a, b, c, d, e, f, g, h, i, and j) are ranked according to their expression intensity. In the non-target species, probes directed against genes e, h, and j perform poorly and have a low rank in the non-target species due to sequence divergence, even though there actually is no expression divergence. This elevates the rank of many other genes, causing an overall negative median rank difference (RD) and a positive skew in the RD distribution. In this example, significantly upregulated genes in the target species tend to have a higher average rank in this species (9) than the significantly upregulated genes in the non-target species do in that species (6.5). Significantly upregulated genes in the target species have a lower average rank in the non-target species (3) than the significantly upregulated genes in the non-target species do in the target species (3.5).

Table 4.1. Proportions of divergently expressed genes differ significantly depending on what probemask is used in the analysis. Results are shown from pairwise comparisons between XL and H (XL versus H) and between a non-target species and a hybrid (NT versus H). All analyses compare testis tissue except the ones from Malone et al. (2008), which compare ovary tissue. For each comparison, the number of significantly upregulated genes in XL (SUXL), significantly upregulated genes in the hybrid (SUH), and significantly upregulated genes in the non-target species (SUNT) is listed. The ratio of divergently expressed genes is equal to the total from the (NT versus H) comparison divided by the total from the (XL versus H) comparison. See text for details of statistical comparisons between these proportions.



Analysis	XL versus H			NT versus H			Ratio of divergently expressed genes	Number of genes analyzed
	SUXL	SUH	Total divergently expressed	Proportion divergently expressed	SUNT	SUH	Total divergently expressed	Proportion divergently expressed		
XB/XL perfect match probemask	1842	2110	892	0.48	280	245	525	0.24	1.84	2143
XB/XL gDNA probemask	79	97	176	0.11	468	299	767	0.46	4.36	1672
XP/XI gDNA probemask	417	572	989	0.11	1430	1248	2678	0.31	2.71	8721
Malone et al. (2007)	92	50	142	0.01	2236	1759	3995	0.35	28.13	11485
Malone et al. (2008)	777	839	1616	0.14	4349	2930	7279	0.63	4.50	11485

Table 4.2. Analyses with gDNA probemasks produce different rank difference distributions in interspecific and intraspecific comparisons. Median and skew of the rank difference distribution and dSRC (see text) are reported. Suffixes after species (XL, XB) refer to the tissue type analyzed: O (ovary), T (testes), T11 (tadpole stage 11), B (brain), and E (egg). Asterisks indicate significant departure from the null. For dSRC, interspecific comparisons and comparisons between a non-target species and a hybrid are higher than other comparisons, and are indicated (C). In all of these cases, the correlation (i) is higher than the correlation (ii).
[a] 1672 probesets, CI median: 0±24, CI skew: 0±0.107
[b] 8721 probesets, CI median: 0±54, CI skew: 0±0.046

Comparisons with XB/XL gDNA probemask[a]			
Interspecific comparisons	median	skew	δSRC
XL_T-XB_T	-54*	0.409*	0.2837[X]
XL_B-XB_B	-58*	0.476*	0.1811[X]
XL_T-XB_B	-51*	0.376*	0.1595[X]
XL_B-XB_T	-54.5*	0.391*	0.2231[X]
Hybrid to parental comparisons			
XL_T-$H(XLXB)_T$	-1	0.016	0.0245
$H(XLXB)_T$-XB_T	-51	0.470	0.2986[X]
Intraspecific comparisons			
XL_T_XL_E	-16	0.140*	0.0794
XL_T-XL_{T11}	-0.5	0.004	0.0304
XL_T-XL_O	17	-0.161*	0.0154
XL_O-XL_{T11}	0	0.000	0.0446
XL_O_XL_E	-12	0.166*	0.0578
XL_E_XL_{T11}	14	-0.178*	0.1325
XL_T_XL_B	12	-0.114*	0.0689
XL_B_XL_E	-13	0.096	0.1578
XL_B_XL_o	11	-0.079	0.0274
XL_B_XL_{T11}	2.5	-0.018	0.1608
Comparisons with XM/XL gDNA probemask[b]			
Interspecific comparisons			
XL_T-XM_T	-269*	0.417*	0.2153[X]
XL_O-XM_O	-390*	0.568*	0.2501[X]
XL_T-XM_O	-250*	0.374*	0.1335[X]
XL_O-XM_T	-338*	0.429*	0.1906[X]
Hybrid to parental comparisons			
XL_T-$H(XLXM)_T$	11	-0.032	0.0129
$H(XLXM)_T$-XM_T	-151*	0.257*	0.1810[X]
XL_O-$H(XLXM)_O$	-32	0.092*	0.0077
$H(XLXM)_O$-XM_O	-187*	0.309*	0.1642[X]
Intraspecific comparisons			
XL_T_XL_E	14	-0.024	0.0091
XL_O-XL_{T11}	14	-0.031	0.0642
XL_T_XL_O	109*	-0.198*	0.0535
XL_T_XL_{T11}	48	-0.081*	0.0043
XL_O-XL_E	-67*	0.175*	0.0477
XL_E-XL_{T11}	77*	-0.192*	0.1313
XL_T_XL_B	17	-0.031	0.0639
XL_B_XL_E	-6	0.009	0.0598
XL_B_XL_O	69*	-0.095*	0.0278
XL_B_XL_{T11}	51	-0.074*	0.0749

Table 4.3. Analysis with the XB + XL perfect match probemaks produces results with similar rank difference statistics in interspecific and intraspecific comparisons. Acronyms and statistics follow Table 4.2.
[a] 2143 probesets, 95% confidence interval (CI) of the median = 0±25, and CI of the skew = 0.000±0.087

Comparisons with XB+XL perfect match probemask[a]			
Interspecific comparisons	median	skew	dSRC
XL_T-XB_T	1	-0.009	0.0029
XL_B-XB_B	1	-0.014	0.0094
XL_T-XB_B	4	-0.035	0.0076
XL_B-XB_T	1	-0.007	0.0045
Hybrid to parental comparisons			
XL_T-$H(XLXB)_T$	4	-0.050	0.0167
XB_T-$H(XLXB)_T$	-2	0.031	0.0197
Intraspecific comparisons			
XL_T_XL_E	-22	0.130*	0.0946
XL_T-XL_{T11}	-13	0.079	0.0558
XL_T-XL_O	-11	0.072	0.0513
XL_O-XL_{T11}	-2	0.018	0.0174
XL_O_XL_E	-11	0.113*	0.0343
XL_E_XL_{T11}	3	-0.030	0.0408
XL_T_XL_B	4	-0.027	0.0008
XL_B_XL_E	-29*	0.154*	0.0619
XL_B_XL_o	-10	0.052	0.0109
XL_B_XL_{T11}	-20	0.110*	0.0710

Table 4.4. Mean rank of significantly upregulated genes in each species based on analysis with probemasks based on gDNA ratios. The rank in XL (RXL), rank in XM (RXM) and rank in XB (RXB) of significantly upregulated genes in XL (SUXL), significantly upregulated genes in XM (SUXM), and significantly upregulated genes in XB (SUXB) is listed along with the difference of these means. Asterisks indicate significant differences after Bonferroni correction for two tests and acronyms follow Table 4.2.

Interspecific comparisons with XM/XL gNDA probemask	SUXL, RXL	SUXM,RXM	Significant?	SUXL, RXM	SUXM, RXL	Significant?
XL-·XM-	6501	5881	*	3347	3857	*
XL$_{2}$·XM$_{0}$	5988	5270	*	3368	3671	*
XL·XM$_{0}$	6196	5713	*	3630	3833	*
XL$_{2}$·XM-	6376	5558	*	3170	3397	*
Hybrid to parental comparisons with XM/XL gDNA probemask						
XL·H(XLXM)$_{1}$	5695	5821		3755	3801	
XM·H(XLXM)$_{1}$	6057	6550	*	4079	3633	*
Intraspecific comparisons with XM/XL gDNA probemask						
XL- XL$_{2}$	6141	6231		3959	4045	
XL$_{2}$·XL$_{1}$	5953	5825		4140	3971	
XL- XL$_{2}$	5783	6100		3951	3860	
XL- XL$_{1;1}$	4097	3962		4011	4019	
XL$_{2}$·XL$_{2}$	5994	5691	*	4254	4299	
XL$_{1}$·XL$_{1;1}$	6256	6095		4436	3859	*
XL- XL$_{2}$	6091	6141		4103	3986	
XL$_{3}$ XL$_{2}$	6183	6202		3893	4054	*
XL$_{2}$ XL$_{2}$	5767	6041	*	3609	3691	
XL$_{2}$ XL$_{1;1}$	6176	6202		3824	3912	
Interspecific comparisons with XB/XL gDNA probemask	SUXL, RXL	XUXB,RXB		SUXL, RXB	SUXB, RXL	
XL·XB-	1219	1164	*	633	782	*
XL$_{2}$·XB$_{2}$	1209	1188		702	871	*
XL·XB$_{3}$	1238	1203		649	789	*
XL$_{3}$·XB-	1230	1188		677	821	*
Hybrid to parental comparisons with XB/XL gDNA probemask						
XL·H(XLXB)-	1134	1201		776	835	
XB$_{1}$·H(XLXB)	1170	1200		884	673	*
Intraspecific comparisons with XB/XL gDNA probemask						
XL- XL$_{2}$	1199	1198		753	781	
XL·XL$_{1;1}$	1190	1217		745	762	
XL·XL$_{2}$	1143	1180		764	747	
XL·XL$_{1}$	1123	1139		747	758	
XL- XL$_{2}$	1141	1069		798	806	
XL$_{2}$ XL$_{1;1}$	1225	1207		873	771	*
XL- XL$_{2}$	1207	1215		800	782	
XL$_{2}$ XL$_{2}$	1214	1206		743	780	
XL$_{3}$ XL$_{2}$	1139	1170		710	704	
XL$_{3}$ XL$_{1;1}$	1194	1206		732	758	

Table 4.5. Mean rank for analyses retaining only confirmed perfect match probes in XL and XB. Abbreviations and symbols follow Table 4.4.

Interspecific comparisons with XB + XL perfect match probemask	SUXL, RXL	SUXB,RXB	Significant?	SUXL, RXB	SUXB, RXL	Significant?
XL_T-XB_T	1308	1274		890	837	
XL_B-XB_B	1330	1296		1040	1023	
XL_T-XB_B	1395	1358		867	831	
XL_B-XB_T	1366	1347		935	915	
Hybrid to parental comparisons with XB + XL probemask						
XL_T-H(XLXB)$_T$	1237	1183		860	765	
XB_T-H(XLXB)$_T$	1224	1260		960	941	
Intraspecific comparisons with XB + XL perfect match probemask						
XL_T_XL_E	1429	1384		830	872	
XL_T-XL_{T11}	1425	1364		834	845	
XL_T-XL_O	1380	1359		852	869	
XL_O-XL_{T11}	1342	1109	*	1086	913	*
XL_O_XL_E	1342	1224	*	954	911	
XL_E_XL_{T11}	1319	1295		899	868	
XL_T_XL_B	1379	1345		882	843	
XL_B_XL_E	1421	1354	*	838	864	
XL_B_XL_O	1347	1358		799	827	
XL_B_XL_{T11}	1424	1367		857	869	

PART II - CONCLUSIONS

Polyploidizations have helped shape a myriad of genomes, including our own. In *Xenopus*, we have found that there is widespread sequence variation and expression divergence between duplicate genes within about fifty million years of whole genome duplication. As a model for early duplicate gene evolution after vertebrate polyploidization, these findings give some insight into the genomic and transcriptomic transitions that can occur following large duplication events. The extensive changes in duplicate gene expression patterns suggest that the regulation of transcript abundance, location, and timing are not only important, but expression can be partitioned across multiple copies of a gene. Furthermore, differential expression can influence the persistence of duplicate genes, such that expression patterns of a gene prior to duplication might affect the odds of certain retention mechanisms acting on duplicate genes.

As seen in the first chapter, most duplicate genes in *Xenopus* have diverged in sequence from each other, as much as they have from a singleton ortholog. Despite their overall rapid evolution compared to singletons, only a small proportion of duplicate genes have molecular patterns that are consistent with models related to neofunctionalization and subfunctionalization at the amino acid level. *Xenopus* duplicate genes may be too old, or our tests not powerful enough to detect clear molecular signatures of such retention mechanisms. Alternatively, other mechanisms like those that occur at the expression level might have preserved both gene copies, a proposal that we further expanded on in chapters two and three. We garnered evidence consistent with this supposition by comparing expression profiles of duplicate genes with those of singleton orthologs. To the extent that an ortholog from the most closely related extant diploid accurately represents ancestral gene expression before WGD, we have found that a large proportion of duplicate genes are expressed in ways consistent with models of neofunctionalization and subfunctionalization of regulation.

If differential regulation of expression of duplicate genes can promote the persistence of both copies, it is expected that genes with particular expression patterns that would allow these changes are more likely to be retained after WGD. By estimating pre-duplication patterns of expression using a diploid species, we find support for these expectations as discussed in chapter three. Genes that are expressed in many tissues have greater odds of being retained after WGD, and also greater odds at partitioning expression in spatiotemporal patterns consistent with regulatory subfunctionalization. However, we detect many more duplicate genes that have gained expression characteristics relative to their singleton ortholog, some in addition to complementary losses. This supports the notion that duplicate genes may be retained via loss or gain of expression domains, and that these events can occur concurrently or sequentially.

Our attempts to find patterns of molecular evolution and expression that are consistent with models of preservation can give some rough idea of the relative importance and characteristics of each model. Our estimates can be more or less accurate as long as subsequent evolutionary changes do not mask initial selective forces and distort the inferences made during this snapshot in time. Nevertheless, it

is quite clear that the bulk of duplicate genes that persist diverge from each other rather quickly following WGD, possibly potentiating their unique and permanent roles in the genome.

Finally, the last chapter of this thesis has critical implications for the use of microarrays in the determination of gene expression divergence between paralogs and between diverged genomes. The risks of technical biases in evaluating gene expression variation between species, hybrids, and even populations are considerable, especially when they are polyploid. The inclusion of even a few microarray probes with poorly hybridizing or cross hybridizing transcripts can affect the global inferences made in expression assays. Innovative ways to tackle such issues are essential, and the consideration of sequence data in assessing probe adequacy is a start. Yet, even when accommodating for differences in sequence identity, the potential for unexpected cross hybridization of transcripts, like those of duplicate genes, can render suspect results. Therefore, after WGD, duplicate genes have impacted the evolution of the *Xenopus* genome and transcriptome, as well as the methods in which we study them. Further efforts to improve the accuracy of expression analyses will improve our ability to determine the relative contributions of the mechanisms that promote the retention of duplicate genes.

PART III – REFERENCES

Adams, K.L. 2007. Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. *J Hered* **98:** 136-141.

Adams, K.L., R. Cronn, R. Percifield, and J.F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences* **100:** 4649-4654.

Adams, K.L. and J.F. Wendel. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8:** 135-141.

Affymetrix. 2001. *Affymetrix GeneChip expression analysis technical manual.* Affymetrix, Santa Clara, CA.

Ahn, S. and S.D. Tanksley. 1993. Comparative linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences* **90:** 7980-7984.

Albertin, W., T. Balliau, P. Brabant, A.-M. Chévre, F. Eber, C. Malosse, and H. Thiellement. 2006. Numerous and rapid nonstochastic modifications of gene products in newly synthesized *Brassica napus* allotetraploids. *Genetics* **173:** 1101-1113.

Allendorf, F. and G. Thorgaard. 1984. Tetraploidy and the evolution of salmonid fishes. In *Evolutionary genetics of fishes* (ed. B.J. Turner), pp. 1-46. Plenum Press, New York.

Allendorf, F.W. and R.G. Danzmann. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* **145:** 1083-1092.

Altschmied, J., J. Delfgaauw, B. Wilde, J. Duschl, L. Bouneau, J.-N. Volff, and M. Schartl. 2002. Subfunctionalization of Duplicate *mitf* Genes Associated With Differential Degeneration of Alternative Exons in Fish. *Genetics* **161:** 259-267.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

Amores, A., A. Force, Y.L. Yan, L. Joly, C. Amemiya, A. Fritz, R.K. Ho, J. Langeland, V. Prince, Y. Wang, M. Westerfield, M. Ekker, and J.H. Postlethwait. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282:** 1711-1714.

Amores, A., T. Suzuki, Y. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J.H. Postlethwait. 2004. Developmental roles of pufferfish *Hox* clusters and genome evolution in ray-fin fish. *Genome Research* **14:** 1-10.

Audic, S. and J.-M. Claverie. 1997. The Significance of Digital Gene Expression Profiles. *Genome Research* **7:** 986-995.

Aury, J.-M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B.M. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Camara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A.-M. Keller, R. Kissmehl, C. Klotz, F. Koll, A. Le Mouel, G. Lepere, S. Malinsky, M. Nowacki, J.K. Nowak, H. Plattner, J. Poulain, F. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Betermier, J. Weissenbach, C.

Scarpelli, V. Schachter, L. Sperling, E. Meyer, J. Cohen, and P. Wincker. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444:** 171-178.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bar-Or, C., M. Bar-Eyal, T.Z. Gal, Y. Kapulnik, H. Czosnek, and H. Koltai. 2006. Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results. *BMC Genomics* **7:** 110.

Becher, M., I.N. Talke, L. Krail, and U. Kramer. 2004. Cross-species microarray transcript profiling reveals high constitutive expression of metal homeostasis genes in shoots of the zinc hyperaccumulator *Arabidopsis holleri*. *Plant J.* **37:** 251-268.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57:** 289-300.

Bisbee, C.A., M.A. Baker, A.C. Wilson, I. Hadji-Azimi, and M. Fischberg. 1977. Albumin phylogeny for clawed frogs (*Xenopus*). *Science* **195:** 785-787.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the *Arabadopsis* genome. *Plant Cell* **12:** 1093-1101.

Blanc, G. and K.H. Wolfe. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *The Plant Cell* **16:** 1679-1691.

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

Bollback, J.P. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7:** 88.

Bolstad, B. 2007. RMA express.

Bolstad, B.M., R.A. Irizarry, M. Astrans, and T.P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinfomatics* **19:** 185-193.

Brunet, F.G., H.R. Crollius, M. Paris, J.-M. Aury, P. Gibert, O. Jaillon, V. Laudet, and M. Robinson-Rechavi. 2006. Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Mol Biol Evol* **23:** 1808-1816.

Buckley, B.A. 2007. Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays. *The Journal of Experimental Biology* **209:** 1602-1606.

Byrne, K.P. and K.H. Wolfe. 2007. Consistent Patterns of Rate Asymmetry and Gene Loss Indicate Widespread Neofunctionalization of Yeast Genes After Whole-Genome Duplication. *Genetics* **175:** 1341-1350.

Byun-McKay, S.A. and R. Geeta. 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. *Trends in Ecology and Evolution* **22:** 338-344.

Cáceres, M., J. Lachuer, M.A. Zapala, J.C. Redmond, L. Kudo, D.H. Geschwind, D.J. Lockhart, and C. Barlow. 2003. Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences* **100:** 13030-13035.

Cambon, A.C., A. Khalyfa, N.G.F. Cooper, and C.M. Thompson. 2007. Analysis of probe level patterns in Affymetrix microarray data. *BMC Bioinformatics* **8:** 146.

Carroll, S.B. 2005. Evolution at Two Levels: On Genes and Form. *PLoS Biology* **3:** e245.

Casneuf, T., S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome Biology* **7:** R13.

Castillo-Davis, C.I. and D.L. Hartl. 2003. GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19:** 891-892.

Chain, F.J.J. and B.J. Evans. 2006. Multiple Mechanisms Promote the Retained Expression of Gene Duplicates in the Tetraploid Frog Xenopus laevis. *PLoS Genetics* **2:** e56.

Chain, F.J.J., D. Ilieva, and B.J. Evans. 2008. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology* **8:** 43.

Chapman, B.A., J.E. Bowers, F.A. Feltus, and A.H. Paterson. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 2730-2735.

Chaudhary, B., L.E. Flagel, R.M. Stupar, J.A. Udall, N. Verma, N.M. Springer, and J. Wendel. 2009. Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homoeologs in Polyploid Cotton (Gossypium). *Genetics*: genetics.109.102608.

Chismar, J.D., T. Mondala, H.S. Fox, E. Roberts, D. Langford, E. Masliah, D.R. Salomon, and S.R. Head. 2002. Analysis of result variability from high-density oligonucleotide arrays comparing same-species and cross-species hybridizations. *Biotechniques* **33:** 516-524.

Clark, A.G. 1994. Invasion and maintenance of a gene duplication. *Proceedings of the National Academy of Sciences* **91:** 2950-2954.

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

Conant, G.C. and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Biology* **4:** 2052-2058.

Cope, L.M., R.A. Irizarray, H.A. Jaffee, Z. Wu, and T.P. Speed. 2004. A benchmark for Affymetrix GeneChip expression measures. *Bioinfomatics* **20:** 323-321.

D'Onofrio, G., D. Mouchiroud, B. Aïssani, C. Gautier, and G. Bernardi. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *Journal of Molecular Evolution* **32:** 504-510.

Davis, J.C. and D.A. Petrov. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biology* **2:** 318-326.

Davis, J.C. and D.A. Petrov. 2005. Do Disparate Mechanisms of Duplication Add Similar Genes to the Genome? *Trends Genetics* **21:** 548-551.

Dean, E.J., J.C. Davis, R.W. Davis, and D.A. Petrov. 2008. Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLoS Genet* **4:** e1000113.

Dehal, R. and J. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* **3:** e314.

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

Drummond, D.A., J.D. Bloom, C. Adami, C.O. Wilke, and F.H. Arnold. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 14338-14343.

Duda, T.F. and S.R. Palumbi. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proceedings of the National Academy of Sciences* **96:** 6820-6823.

Dulai, K.S., M. von Dornum, J.D. Mollon, and D.M. Hunt. 1999. The evolution of trichromatic colour vision by opsin gene duplication in New World and Old World primates. *Genome Research* **9:** 629-638.

Duret, L. and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects seleciton intensity but not mutation rate. *Molecular Biology and Evolution* **17:** 68-74.

Edgar, R., M. Domrachev, and A.E. Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30:** 207-210.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32:** 1792-1797.

Ellinger-Ziegelbauer, H., A. Hihi, V. Laudet, H. Keller, W. Wahli, and C. Dreyer. 1994. FTZ-F1-related orphan receptors in *Xenopus laevis*: transcriptional regulators differentially expressed during early embryogenesis. *Molecular and Cellular Biology* **14:** 2786-2797.

Enard, W., P. Khaitovich, J. Klose, F. Zoellner, F. Hessig, R. Giavalisco, K. Nieselt-Strew, E. Muchmore, A. Varki, R. Ravid, G.M. Doxiadis, R.E. Bontrop, and S. Pääbo. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296:** 340-343.

Evans, B.J. 2007. Ancestry Influences the Fate of Duplicated Genes Millions of Years After Polyploidization of Clawed Frogs (Xenopus). *Genetics* **176:** 1119-1130.

Evans, B.J. 2008. Genome evolution and speciation genetics of clawed frogs (*Xenopus* and *Silurana*). *Frontiers in Bioscience* **13:** 4687-4706.

Evans, B.J., D.B. Kelley, D.J. Melnick, and D.C. Cannatella. 2005. Evolution of RAG-1 in polyploid clawed frogs. *Molecular Biology and Evolution* **22:** 1193-1207.

Evans, B.J., D.B. Kelley, R.C. Tinsley, D.J. Melnick, and D.C. Cannatella. 2004. A mitochondrial DNA phylogeny of clawed frogs: phylogeography on sub-Saharan Africa and implications for polyploid evolution. *Molecular Phylogenetics and Evolution* **33:** 197-213.

Fay, J.C. and C.-I. Wu. 2001. The neutral theory in the genomic era. *Current Opinion in Genetics and Development* **11:** 642-646.

Fay, J.C. and C.-I. Wu. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annual Review of Genomics and Human Genetics* **4:** 213-235.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Ferris, S.D. and G.S. Whitt. 1977. Loss of duplicate gene expression after polyploidization. *Nature* **265:** 258-260.

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

Flagel, L., J. Udall, D. Nettleton, and J. Wendel. 2008. Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biology* **6:** 16.

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

Freeling, M. and B.C. Thomas. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16:** 805-814.

Fryxell, K.J. and W. Moon. 2004. CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution* **22:** 650-658.

Galitski, T., A.J. Saldanha, C.A. Styles, E.S. Lander, and G.R. Fink. 1999. Ploidy Regulation of Gene Expression. *Science* **285:** 251-254.

Gaut, B.S. and J.F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* **94:** 6809-6814.

Gibbs, M.J., J.S. Armstrong, and A.J. Gibbs. 2000. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16:** 573-582.

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

Gilad, Y. and J. Borevitz. 2006. Using DNA microarrays to study natural variation. *Current Opinion in Genetics and Development* **16:** 553-558.

Gilad, Y., S.A. Rifkin, P. Bertone, M. Gerstein, and K.P. White. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Research* **15:** 674-680.

Golding, G.B. and A.M. Dean. 1998. The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15:** 355-369.

Goldman, N., J.P. Anderson, and A.G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* **49:** 652-670.

Goodman, M., J. Czelusniak, B.F. Koop, D.A. Tagle, and J.L. Slightom. 1987. Globins: a case study in molecular phylogeny. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 875-890.

Goss, P.J.E. and R.C. Lewontin. 1996. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* **143:** 589-602.

Grigoryev, D.N., S.-F. Ma, B.A. Simon, R.A. Irizarry, S.Q. Ye, and J.G.N. Garcia. 2005. *In vitro* identification and *in silico* utilization of interspecies sequence similarities using GeneChip technology. *BMC Genomics* **6:** 62.

Gu, X. and Z. Su. 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences* **104:** 2779-2784.

Gu, X., Z. Zhang, and W. Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 707-712.

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

Gu, Z., S.A. Rifkin, K.P. White, and W.H. Li. 2004. Duplicate genes increase gene expression diversity within and between species. *Nature Genetics* **36:** 577-579.

Gu, Z., L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, and W.H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **291:** 63-66.

Guan, Y., M.J. Dunham, and O.G. Troyanskaya. 2006. Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics*: genetics.106.064329.

Gurvich, N., M.G. Berman, B.S. Wittner, R.C. Gentleman, P.S. Klein, and J.B.A. Green. 2005. Association of valproate-induced teratogenesis with histone deacetylase inhibition in vivo. *FASEB* **19:** 1166-1168.

Ha, M., E.-D. Kim, and Z.J. Chen. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences* **106:** 2295-2300.

Hakes, L., J. Pinney, S. Lovell, S. Oliver, and D. Robertson. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology* **8:** R209.

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

Hammond, J.P., M.R. Boroadley, D.J. Craigon, J. Higgins, Z.F. Emmerson, H.J. Townsend, P.J. White, and S.T. May. 2005. Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species. *Plant Methods* **10:** 1-10.

Hammond, J.P., H.C. Bowen, P.J. Wite, V. Mills, K.A. Pyke, A.J.M. Baker, S.N. Whiting, S.T. May, and M.R. Broadley. 2006. A comparison of the *Thlaspi caerulescens* and *Thlaspi arvense* shoot transcriptomes. *New Phytologist* **170:** 239-260.

Hancock, J. 2005. Gene factories, microfunctionalization and the evolution of gene families. *Trends in Genetics* **21:** 591-594.

Hastings, K. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *Journal of Molecular Evolution* **42:** 631-640.

He, X. and J. Zhang. 2005a. Gene Complexity and Gene Duplicability. *Current Biology* **15:** 1016-1021.

He, X. and J. Zhang. 2005b. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics* **169:** 1157-1164.

He, X. and J. Zhang. 2006. Higher Duplicability of Less Important Genes in Yeast Genomes. *Mol Biol Evol* **23:** 144-151.

Hegarty, M.J., G.L. Barker, I.D. Wilson, R.J. Abbott, K.J. Edwards, and S.J. Hiscock. 2006. Transcriptome shock after interspecific hybridization in *Senecio* is ameliorated by genome duplication. *Current Biology* **16:** 1652-1659.

Hellsten, U., M. Khokha, T. Grammer, R. Harland, P. Richardson, and D. Rokhsar. 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biology* **5:** 31.

Henrici, A.C. and A.M. Báez. 2001. First occurrence of *Xenopus* (Anura: Pipidae) on the Arabian Penninsula: a new species from the Late Oligocene of the Republic of Yemen. *Journal of Paleontology* **75:** 870-882.

Hsieh, W.-P., T.-M. Chu, R.D. Wolfinger, and G. Gibson. 2003. Mixed-model reanalysis of primate data suggest tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* **165:** 747-757.

Huelsenbeck, J.P., D.M. Hillis, and R. Jones. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. In *Molecular Zoology: Advances, Strategies, and Protocols.* (eds. J.D. Ferraris and S.R. Palumbi), pp. 19-45. Wiley-Liss, New York.

Huelsenbeck, J.P. and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754-755.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London* **256:** 119-124.

Hughes, A.L. 1999. *Adaptive evolution of genes and genomes.* Oxford Press, New York.

Hughes, A.L., T. Ota, and M. Nei. 1990. Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Molecular Biology and Evolution* **7:** 515-524.

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

Huminiecki, L. and K.H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Research* **14:** 1870-1879.

Husband, B.C. and D.W. Schemske. 2000. Ecological mechanisms of reproductive isolation between diploid and tetraploid *Chamerion angustifolium. Journal of Ecology* **88:** 689-701.

Irwin, J.O. 1937. The frequency distribution of the difference between two independent variates following the same Poisson distribution. *Journal of the Royal Statistical Society: Series A.* **100:** 415-416.

Isaacs, H.V., D. Tannahill, and J.M.W. Slack. 1992. Expression of a novel FGF in the *Xenopus* embryo. A new candidate inducing factor for mesoderm formation and anteroposterior specification. *Development* **114:** 711-720.

Jensen, R.A. 1976. Enzyme recruitment in evolution of new function. *Annual Review of Microbiology* **30:** 409-425.

Ji, W., W. Zhou, K. Gregg, N. Yu, S. Davis, and S. Davis. 2004. A method for cross-species gene expression analysis with high-density oligonucleotide arrays. *Nucleic Acids Research* **32:e93.**

Jiang, H., D. Liu, Z. Gu, and W. Wang. 2007. Rapid evolution in a pair of recent duplicate segments of rice. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **308B:** 50-57.

Jordan, I.K., L. Mariño-Ramírez, and E.V. Koonin. 2005. Evolutionary significance of gene expression divergence. *Gene* **345:** 119-126.

Jordan, I.K., Y.I. Wolf, and E.V. Koonin. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology* **4:** 22.

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

Katz, L. and C.B. Burge. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research* **13:** 2042-2051.

Khaitovich, P., P. Weiss, M. Lachmann, I. Hellman, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo. 2004. A neutral model of transcriptome evolution. *PLoS Biology* **2:** 682-689.

Kim, J., S.-H. Shiu, S. Thoma, W.-H. Li, and S. Patterson. 2006. Patterns of expansion and expression divergence in the plant polygalacturonase gene family. *Genome Biology* **7**: R87.

Kim, S.-H. and S.V. Yi. 2006. Correlated Asymmetry of Sequence and Functional Divergence Between Duplicate Proteins of Saccharomyces cerevisiae. *Mol Biol Evol* **23**: 1068-1075.

Kimura, M. 1980. Average time until fixation of a mutant allele in a finite population under continued mutant pressure: studies by analytical, numerical, and pseudosampling methods. *Proceedings of the National Academy of Sciences* **77**: 522-526.

Kimura, M. 1983. The neutral theory of molecular evolution. In *Evolution of Genes and Proteins* (eds. M. Nei and R. Koehn). Sinauer Associates., Sunderland.

Kirst, M., R. Caldo, P. Casati, G. Tanimoto, V. Walbot, R.P. Wise, and E.S. Buckler. 2006. Genetic diversity contribution to errors in short oligonucleotide microarray analysis. *Plant Biotechnology Journal* **4**: 489-498.

Knochel, W., E. Korge, A. Basner, and W. Meyerhof. 1986. Globin evolution in the genus *Xenopus*: compatative analysis of cDNAs coding for adult globin polypeptides of *Xenopus borealis* and *Xenopus tropicalis*. *Journal of Molecular Evolution* **23**: 211-223.

Kobel, H. and L. DuPasquier. 1986. Genetics of polyploid Xenopus. *Trends Genetics* **2**: 310-315.

Kobel, H.R. 1996. Allopolyploid speciation. In *The Biology of Xenopus* (eds. R.C. Tinsley and H.R. Kobel), pp. 391-401. Clarendon Press, Oxford.

Kondrashov, F.A. and E.V. Koonin. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics* **20**: 287-290.

Kondrashov, F.A., I.B. Rogozin, Y.I. Wolf, and E.V. Koonin. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3**: RESEARCH0008.

Kuepfer, L., U. Sauer, and L.M. Blank. 2005. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Research* **15**: 1421-1430.

Kuma, K., N. Iwabe, and T. Miyata. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Mol Biol Evol* **12**: 123-130.

Larhammer, D. and C. Risinger. 1994. Why so few pseudogenes in tetraploid species? *Trends in Genetics* **10**: 418-419.

Lercher, M.J., J. Chamary, and L.D. Hurst. 2004. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Research* **14**: 1002-1013.

Lercher, M.J., A.O. Urrutia, and L.D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31**: 180-183.

Li, C. and W.H. Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2:** research0032.

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

Li, W.-H. 1982. Evolutionary change of duplicate genes. *Isozymes* **6:** 55-92.

Li, W.-H. 1985. Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy. In *Population Genetics and Molecular Evolution* (eds. T. Ohta and K. Aoki), pp. 333-352. Springer-Verlag, Berlin.

Li, W.-H. 1997. *Molecular Evolution.* Sinauer Publishers, Sunderland.

Li, W.-H., J. Yang, and X. Gu. 2005. Expression divergence between duplicate genes. *Trends in Genetics* **21:** 602-607.

Liao, B.-Y. and J. Zhang. 2006. Low Rates of Expression Profile Divergence in Highly Expressed Genes and Tissue-Specific Genes During Mammalian Evolution. *Molecular Biology and Evolution* **23:** 1119-1128.

Liao, B.-Y. and J. Zhang. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proceedings of the National Academy of Sciences* **105:** 6987-6992.

Lister, J.A., J. Close, and D.W. Raible. 2001. Duplicate *mitf* Genes in Zebrafish: Complementary Expression and Conservation of Melanogenic Potential. *Developmental Biology* **237:** 333-344.

Liu, B. and J.F. Wendel. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Molecular Phylogenetics and Evolution* **29:** 365-379.

Long, M. and K. Thornton. 2001. Gene duplication and evolution. *Science* **293:** 1551a.

Lynch, M. and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151-1155.

Lynch, M. and J.S. Conery. 2003. The origins of genome complexity. *Science* **302:** 1401-1404.

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

Lynch, M. and V. Katju. 2004. The altered evolutionary trajectories of gene duplicates. *Trends in Genetics* **20:** 544-549.

Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. The Probability of Preservation of a Newly Arisen Gene Duplicate. *Genetics* **159:** 1789-1804.

MacEachern, S., J. McEwan, A. Mather, A. McCulloch, P. Sunnucks, and M. Goddard. 2006. Testing the neutral theory of molecular evolution using genomic data: a comparison of the human and bovine transcriptome. *Genetics Selection Evolution* **38:** 321-341.

Maddison, D.R. and W.P. Maddison. 2000. MacClade. Sinauer Associates, Sunderland.

Maisey, J.G. 2000. Continental break up and the distribution of fishes of Western Gondwana during the Early Cretaceous. *Cretaceous Research* **21:** 281-314.

Makova, K.D. and W.H. Li. 2003. Divergence in the Spatial Pattern of Gene Expression Between Human Duplicate Genes. *Genome Research* **13:** 1638-1645.

Malone, J.H., T.H. Chrzanowski, and P. Michalak. 2007. Sterility and gene expression in hybrid males of *Xenopus laevis* and *X. muelleri*. *PLoS One* **2:** e781.

Malone, J.H., D.L. Hawkins, and P. Michalak. 2006. Sex biased gene expression in a ZW sex determination system. *Journal of Molecular Evolution* **63:** 427-436.

Malone, J.H. and P. Michalak. 2008a. Gene expression analysis of the ovary of hybrid females of *Xenopus laevis* and *X. muelleri*. *BMC Evolutionary Biology* **8:** 82.

Malone, J.H. and P. Michalak. 2008b. Physiological sex predicts hybrid sterility regardless of genotype. *Science* **319:** 59.

Martin, D. and E. Rybicki. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16:** 562-563.

Massingham, T., L.J. Davies, and P. Lio. 2001. Analysing gene function after duplication. *BioEssays* **23:** 873-876.

Maynard Smith, J. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34:** 126-129.

McClintock, J.M., M.A. Kheirbek, and V.E. Prince. 2002. Knockdown of duplicated zebrafish *hoxb1* genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development* **129:** 2339-2354.

McLoughlin, S. 2001. The breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Journal of Botany* **49:** 271-200.

Meiklejohn, C.D., J. Parsch, J.M. Ranz, and D.L. Hartl. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proceedings of the National Academy of Sciences* **100:** 9894-9899.

Michalak, P. and M.A.F. Noor. 2003. Genome-Wide Patterns of Expression in *Drosophila* Pure Species and Hybrid Males. *Molecular Biology and Evolution* **20:** 1070-1076.

Moore, R.C., S. Grant, and M.D. Purugganan. 2005. Molecular Population Genetics of Redundant Floral-Regulatory Genes in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **22:** 91-103.

Moore, R.C. and M.D. Purugganan. 2003. The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences* **100:** 15682-15687.

Morin, R.D., E. Chang, A. Petrescu, N. Liao, M. Griffith, R. Kirkpatrick, Y.S. Butterfield, A.C. Young, J. Stott, S. Barber, R. Babakaiff, M.C. Dickson, C. Matsuo, D. Wong, G.S. Yang, D.E. Smailus, K.D. Wetherby, P.N. Kwong, J. Grimwood, C.P. Brinkley, M. Brown-John, N.D. Reddix-Dugue, M. Mayo, J. Schmutz, J. Beland, M. Park, S. Gibson, T. Olson, G.G. Bouffard, M. Tsai, R. Featherstone, S. Chand, A.S. Siddiqui, W. Jang, E. Lee, S.L. Klein, R.W. Blakesley, B.R. Zeeberg, S. Narasimhan, J.N. Weinstein, C.P. Pennacchio,

R.M. Myers, E.D. Green, L. Wagner, D.S. Gerhard, M.A. Marra, S.J.M. Jones, and R.A. Holt. 2006. Sequencing and analysis of 10,967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis reveals post-tetraploidization transcriptome remodeling. *Genome Research* **16:** 796-803.

Müller, W.P. 1977. Diplotene chromosomes of Xenopus hybrid oocytes. *Chromosoma* **59:** 273-282.

Myers, E.W. and W. Miller. 1988. Optimal alignments in linear space. *Computer Applications in the Biosciences*: 11-17.

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

Nei, M. and S. Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nembaware, V., K. Crum, J. Kelson, and C. Seoighe. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Research* **12:** 1370-1376.

Nielsen, R. 2002. Mapping mutations on phylogenies. *Systematic Biology* **51:** 729-739.

Nieuwkoop, P.D. and J. Faber. 1956. *Normal table of Xenopus laevis*. Daudin, Amsterdam.

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

Nuzhdin, S.V., M.L. Wayne, K.L. Harmon, and L.M. McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila. Molecular Biology and Evolution* **21:** 1308-1317.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin.

Ohno, S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244**.

Ohta, T. 1987. Time for acquiring a new gene by duplication. *Proceedings of the National Academy of Sciences* **85:** 3509-3512.

Osborn, T.C., J.C. Pires, J.A. Birchler, D.L. Auger, Z.J. Chen, H. Lee, L. Comai, A. Madlung, R.W. Doerge, V. Colot, and R.A. Martienssen. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics* **19:** 141-147.

Oshlack, A., A.E. Chabot, G.K. Smyth, and Y. Gilad. 2007. Using DNA microarrays to study gene expression in closely related species. *Bioinformatics* **23:** 1235-1242.

Padidam, M., S. Sawyer, and C.M. Fauquet. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265:** 218-225.

Pál, C., B. Papp, and L.D. Hurst. 2001. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* **158:** 927-931.

Papp, B., C. Pál, and L.D. Hurst. 2003a. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194-197.

Papp, B., C. Pál, and L.D. Hurst. 2003b. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends in Genetics* **19:** 417-422.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence analysis. *Proceedings of the National Academy of Sciences* **85:** 2444-2448.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19:** 651-652.

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

Pond, S.K. and S.V. Muse. 2005. Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution* **22:** 2375-2385.

Poole, R., G.L. Barker, I.D. Wilson, J.A. Coghill, and K.J. Edwards. 2007. Measuring global gene expression in polyploidy; a cautionary note from allohexaploid wheat. *Functional and Integrative Genomics* **7:** 207-219.

Posada, D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular Biology and Evolution* **19:** 708-717.

Posada, D. and K.A. Crandall. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Posada, D. and K.A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences* **98:** 13757-13762.

Postlethwait, J.H., A. Amores, W. Cresko, A. Singer, and Y.L. Yan. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genetics* **20:** 481-490.

Postlethwait, J.H., I.G. Woods, P. Ngo-Hazelett, Y.L. Yan, P.D. Kelly, F. Chu, H. Huang, A. Hill-Force, and W.S. Talbot. 2000. Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. *Genome Research* **10:** 1890-1902.

Prince, V.E. and F.B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics* **3:** 827-837.

Qian, W. and J. Zhang. 2008. Gene Dosage and Gene Duplicability. *Genetics* **179:** 2319-2324.

Rambaut, A. and N.C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* **13:** 235-238.

Ranz, J.M., C.I. Castillo-Davis, C.D. Meiklejohn, and D.L. Hartl. 2003. Sex-dependent gene expression and evolution in the *Drosophila* transcriptome. *Science* **300:** 1742-1745.

Rastogi, S. and D. Liberles. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology* **5:** 28.

Rice, W.R. 1989. Analyzing tables of statistical tests. *Evolution* **43:** 223-225.

Robinson-Rechavi, M. and V. Laudet. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Molecular Biology and Evolution* **18:** 681-683.

Rodin, S. and A. Riggs. 2003. Epigenetic Silencing May Aid Evolution by Gene Duplication. *Journal of Molecular Evolution* **56:** 718-729.

Rodin, S.N., D.V. Parkhomchuk, A.S. Rodin, G.P. Holmquist, and A.D. Riggs. 2005. Repositioning-Dependent Fate of Duplicate Genes. *DNA and Cell Biology* **24:** 529-542.

Rokas, A., B.L. Williams, N. King, and S.B. Carrol. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425:** 798-804.

Salminen, M.O., J.K. Carr, D.S. Burke, and F.E. McCutchan. 1995. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retroviruses* **11:** 1423-1425.

Sammut, B., A. Marcuz, and L. Du Pasquier. 2002. The fate of duplicated major histocompatibility complex class 1a genes in a dodecaploid amphibian, *Xenopus ruwenzoriensis. European Journal of Immunology* **32:** 1593-1604.

Sanderson, M.J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **19:** 101-109.

Sanderson, M.J. 2002. Estimating absolute rates of molecular evolution and divergence times in the absence of rate consistency. *Molecular Biology and Evolution* **19:** 1218-1231.

Sartor, M.A., A.M. Zorn, J.A. Schwanekamp, D. Halbleib, S. Karyala, M.L. Howell, G.E. Dean, M. Medvedovic, and C.R. Tomlinson. 2006. A new method to remove hybridization bias for interspecies comparisons of global gene expression profiles uncovers an association between mRNA sequence divergence and differential gene expression in *Xenopus. Nucleic Acids Research* **34:** 185-200.

Scannell, D.R. and K.H. Wolfe. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* **18:** 137-147.

Sémon, M. and K.H. Wolfe. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proceedings of the National Academy of Sciences* **105:** 8333-8338.

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

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

Shakhnovich, B.E. and E.V. Koonin. 2006. Origins and impact of constraints in evolution of gene families. *Genome Research* **16:** 1529-1536.

Shields, D.C., P.M. Sharp, D.G. Huggins, and F. Wright. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution* **5:** 704-716.

Shiu, S.-H., J.K. Byrnes, R. Ran, P. Zhang, and W.-H. Li. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proceedings of the National Academy of Sciences* **103:** 2232-2236.

Shubin, N.H. and F.A. Jenkins. 1995. An early Jurrasic jumping frog. *Nature* **377:** 49-52.

Sidow, A. 1996. Gen(om)e duplications in the evvolution of early vertebrates. *Current Opinion in Genetics and Development* **6:** 715-722.

Sinner, S., P. Kirilenko, S. Rankin, E. Wei, L. Howard, M. Korfran, J. Heasman, H.R. Woodland, and A.M. Zorn. 2006. Global analysis of the transcriptional network controlling *Xenopus* endoderm formation. *Development* **133:** 1955-1966.

Smit, A.F.A., R. Hubley, and P. Green. 2004. RepeatMasker. http://www.repeatmasker.org.

Smyth, G. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3:** 3.

Smyth, G.K. and T. Speed. 2003. Normalization of cDNA microarray data. *Methods* **31:** 265-273.

Sokal, R.R. and F.J. Rohlf. 2003. *Biometry*. W. H. Freeman and Company, New York.

Soltis, D.E. and P.S. Soltis. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in Ecology and Evolution* **14:** 348-352.

Spring, J. 1997. Vertebrate evolution by interspecific hybridization – are we polyploid? *FEBS Letters* **400:** 2-8.

Stoltzfus, A. 1999. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* **49:** 169-181.

Strähle, U. and S. Rastegar. 2008. Conserved non-coding sequences and transcriptional regulation. *Brain Research Bulletin* **75:** 225-230.

Su, Z., J. Wang, J. Yu, X. Huang, and X. Gu. 2006. Evolution of alternative splicing after gene duplication. *Genome research* **16:** 182-189.

Subramanian, S. and S. Kumar. 2004. Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome. *Genetics* **168:** 373-381.

Swofford, D.L. 2002. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland.

Takahata, N. and T. Maruyama. 1979. Polyploidization and loss of duplicate gene expression: a theoretical study with application to tetraploid fish. *Proceedings of the National Academy of Sciences* **76:** 4521-4525.

Tang, H. and R.C. Lewontin. 1999. Locating regions of differential variability in DNA and protein sequences. *Genetics* **153:** 485-495.

Taylor, J.S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research* **13:** 382-390.

Taylor, J.S., Y. Van de Peer, I. Braasch, and A. Meyer. 2001a. Comparative genomics provides evidence for an ancient genome duplication even in fish. *Philosophical Transactions of the Royal Society of London, Series B* **356:** 1661-1679.

Taylor, J.S., Y. Van de Peer, and A. Meyer. 2001b. Genome duplication, divergent resolution and speciation. *Trends in Genetics* **17:** 299-301.

Tinsley, R.C. and M.J. McCoid. 1996. Feral populations of *Xenopus* outside Africa. In *The Biology of Xenopus* (eds. R.C. Tinsley and H.R. Kobel), pp. 81-94. Clarendon Press, Oxford.

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

Tymowska, J. 1991. Polyploidy and cytogenetic variation in frogs of the genus *Xenopus*. In *Amphibian cytogenetics and evolution* (eds. D.S. Green and S.K. Sessions). Academic Press., San Diego.

Uddin, M., D.E. Wildman, G. Liu, W. Xu, R.M. Johnson, P.R. Hof, G. Kapatos, L.I. Grossman, and M. Goodman. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proceedings of the National Academy of Sciences* **101:** 2957-2962.

Urbina, D., B. Tang, and P.G. Higgs. 2006. The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. *Journal of Molecular Evolution* **62:** 340-361.

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

Veitia, R.A. 2003. Nonlinear effects in macromolecular assembly and dosage sensitivity. *Journal of Theoretical Biology* **220:** 19-25.

Veitia, R.A. 2004. Gene Dosage Balance in Cellular Pathways: Implications for Dominance and Gene Duplicability. *Genetics* **168:** 569-574.

Vinogradov, A.E. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends in Genetics* **20:** 248-253.

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. *Proc Natl Acad Sci U S A* **97:** 6579-6584.

Wagner, A. 2000b. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. *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. Selection and gene duplication: a view from the genome. *Genome Biology* **3:** 1012.1011-1012.1013.

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* **139:** 421-428.

Wang, J., L. Tian, H.-S. Lee, N.E. Wei, H. Jiang, B. Watson, A. Madlung, T.C. Osborn, R.W. Doerge, L. Comai, and Z.J. Chen. 2006. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172:** 507-517.

Watterson, G.A. 1983. On the time for gene silencing at duplicate loci. *Genetics* **105:** 745-766.

Weber, M., E. Harada, C. Vess, E. Roepenack-Lahaye, and S. Clemens. 2004. Comparative microarray analysis of *Arabidopsis thaliana* and *Arabidopsis halleri* roots identifies nicotianamine synthase, a ZIP transporter and other genes as potential metal hyperaccumulation factors. *Plant J.* **37:** 269-281.

Wendel, J.F. 2000. Genome evolution in polyploids. *Plant Molecular Biology* **42:** 225-249.

Werth, C.R. and M.D. Windham. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *American Naturalist* **137:** 515-526.

Woody, O.Z., A.C. Doxey, and B.J. McConkey. 2008. Assessing the Evolution of Gene Expression Using Microarray Data. *Evolutionary Bioinformatics* **4:** 139-152.

Wu, K.H., M.L. Tobias, J.W. Thornton, and D.B. Kelley. 2003. Estrogen receptors in *Xenopus*: duplicate genes, splice variants, and tissue-specific expression. *General and Comparative Endocrinology* **133:** 38-49.

Wyckoff, G.J., C.M. Malcom, E.J. Vallender, and B.T. Lahn. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends in Genetics* **21:** 381-385.

Xu, Q., B.S. Baker, and J.R. Tata. 1993. Developmental and hormonal regulation of the *Xenopus* liver-type arginase gene. *European Journal of Biochemistry* **211:** 891-898.

Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21:** 650-659.

Yang, J., A.I. Su, and W.-H. Li. 2005. Gene Expression Evolves Faster in Narrowly Than in Broadly Expressed Mammalian Genes. *Mol Biol Evol* **22:** 2113-2118.

Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.J. Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple systematic variation. *Nucleic Acids Research* **30:** e15.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13:** 555-556.

Yang, Z., R. Nielsen, N. Goldman, and A. Krabbe Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431-449.

Yim, W., B.-M. Lee, and C. Jang. 2009. Expression diversity and evolutionary dynamics of rice duplicate genes. *Molecular Genetics and Genomics* **281:** 483-493.

Zakharkin, S.O., K. Kim, T. Mehta, L. Chen, S. Barnes, K.E. Scheirer, R.S. Parrish, D.B. Allison, and G.P. Page. 2005. Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics* **6:** 214.

Zhang, J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution* **50:** 56-68.

Zhang, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18:** 292-297.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22:** 2472-2479.

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

Zhang, J. and D.M. Webb. 2003. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proceedings of the National Academy of Sciences* **100:** 8337-8341.

Zhang, J., Y. Zhang, and H.F. Rosenberg. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics* **30**.

Zhang, L. and W.-H. Li. 2004. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. *Mol Biol Evol* **21:** 236-239.

Zhang, P., Z. Gu, and W.H. Li. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biology* **4:** R56.