Duplicate Gene Evolution in a Tetraploid African Clawed Frog

DUPLICATE GENE EVOLUTION IN A TETRAPLOID AFRICAN CLAWED FROG

BY

BRIAN ALCOCK, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF BIOLOGY

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2015) McMaster University

(Biology) Hamilton, Ontario, Canada

TITLE:	Duplicate Gene Evolution in a Tetraploid African Clawed Frog
AUTHOR:	Brian Alcock B.Sc., (Biology) Memorial University, St. John's, Newfoundland
SUPERVISOR:	Dr. Ben J. Evans
NUMBER OF PAGES:	x, 80

Lay Abstract

Whole-genome duplication (WGD) is considered a major source of evolutionary innovation and a driver of speciation. By increasing genetic content and introducing redundancy, selective pressures are reduced and paralogous pairs diverge. We investigate how sex and tissue type contribute to duplicate gene divergence following WGD in a tetraploid African clawed frog. We find evidence for sex-dependent variation in sex-biased expression patterns of duplicate genes in brain, heart and liver, and evaluate how molecular evolution of duplicate genes accounts for expression divergence between sexes. This thesis provides a general framework for investigating sex-biased duplicate gene evolution in an amphibious tetrapod.

Abstract

By increasing genomic size, whole-genome duplication (WGD) is considered a major source of evolutionary innovation and speciation. We examined sequence evolution and expression divergence following WGD in a tetraploid African clawed frog (*Silurana*). We hypothesized that the redundancy generated by WGD might allow for sex-specific and/or tissue-specific divergence, contributing to sexual dimorphism in this frog, and that such changes could be detected at both the expression and sequence levels. We investigated this hypothesis with a transcriptome-based approach, comparing both sexes across brain, heart and liver. We compared molecular evolution and expression divergence of duplicate gene homeologs to singleton genes and to an extant diploid relative, and identified genes with evidence for sex-biased expression. In doing so, we provide evidence for an allopolyploid mechanism of WGD and speciation in *Silurana*. Additionally, we find that female-biased gene expression is more prevalent among duplicate genes than male-biased expression, particularly in brain where expression levels are highest. We similarly identified antagonistically sex-biased homeologs with indication of positive selection. Our results indicate that divergent evolution at both the sequence and expression levels following WGD favors the co-option of female-biased gene expression and may help resolve sexually antagonistic selection in this frog, thereby facilitating the evolution of sexual dimorphism.

Acknowledgements

I'd first like to acknowledge my Evans' lab companions, Hermina Ghenu and Ben Furman, my partner, Anthony Reis, and all my friends throughout McMaster University: thank you for the hours of help and distraction, both academic and recreational, without which I'm certain my time at McMaster would have been far less enjoyable (and perhaps more productive, but who can ever really be sure?)

Next, my supervisory committee members, Dr. Brian Golding and Dr. Ben Bolker. Brian: thank you for sparking an interest in bioinformatics and computational biology that has guided my career ambitions for the past two years; I'm not sure I otherwise would have ever discovered this passion. Ben: the hours of meetings with you were not only a tremendous help towards my thesis but expanded my horizons and, really, changed the way I thought about data, statistics and my project. Without any exaggeration I can confidently say that both of you have strongly contributed to my aspirations for the future.

Finally, to my advisor, Dr. Ben Evans: I could not thank you enough for the opportunity to take this project. While I was expecting something different when I signed up, frankly I could not be more pleased with the direction my project has taken and how it has influenced me. You have been an excellent advisor: patient, affable and relatable, and understanding of the need to balance science and life. We've

spoken many times about how strong a determinant the advisor-student relationship is for a graduate experience, and I am certain I made a great choice. My experience with you is something I will always look for in professional relationships.

Contents

List of Tables

List of Figures

Chapter 1

An Overview of the Roles of Gene Duplication

1.1 Gene Duplication in Evolution

In 1970, Susumu Ohno published his seminal work, *Evolution by Gene Duplication*. In it, he discussed the role of gene duplication in evolution and posited that the genetic material introduced to the genome by gene duplication was a source for evolutionary innovation, novel gene function conversion and speciation [1]. While gene and genome duplication was known to biologists, this marked the first time the ability for duplicate genes to shape evolutionary change had been succinctly and concretely recorded.

Prior to *Evolution by Gene Duplication*, other studies had shown the phenotypic responses to gene duplication. One of the earliest observations came from studies of the *Bar* gene mutations in the fruit fly, *Drosophila melanogaster* [2]. Bridges reported that chromosomal band doubling and tripling in a *Bar* mutant fly was linked to a notable reduction in eye size [2]. With this finding, others began to speculate on

the possible role of gene duplication in evolution [3, 4, 5, 6, 7]. Their studies suggested that gene duplication could drive evolutionary change through a combination of genetic redundancy, relaxed purifying selection and mutation accumulation.

It was during this time that the mechanisms through which duplicate genes arise were described, for example through unequal crossing over or chromosomal duplication. However, it was not until the late 1990s that a strong interest in gene duplication came to the forefront of genetics, a shift that was accompanied by an increasing availability of genomic information and methods. Here, Ohno's suspicion was confirmed: while the null expectation for duplicate-gene fate was often pseudogenization through the acquisition of silencing mutations in gene coding or regulatory regions [8, 9, 10, 11], a surprising number of duplicate-genes persist over time [12, 13]. We would not expect this to be the case, unless selection favors retention of both duplicate-gene copies. As Ohno suggested, this indicates that gene duplication is an important driver of genome evolution and evolution of novel functions [1, 14].

As mentioned above, duplicate-genes are generated by multiple mechanisms. However, a distinction should be made between small-scale gene duplication and large-scale chromosome or whole-genome duplication (WGD). Chromosomal duplication occurs, for instance, through meiotic non-disjunction, generates aneuploid individuals, and is associated with many human genetic disorders (See Section 1.4). WGD is accompanied by an increase in ploidy level, such as from diploid ($2n$) to tetraploid ($4n$), and individuals can be allopolyploid or autopolyploid. Allopolyploidy refers to polyploids generated by the hybridization of two species, while autopolyploidy refers to spontaneous within-genome doubling; in either case, the mechanisms responsible for increased ploidy level, such as polyspermy or gametic nonreduction, are similar

[15, 16]. This is an important distinction, as the probability of duplicate-gene retention for paralogous genes generated by large-scale duplication events is higher than individual gene duplications [13, 17, 18, 19].

The evolution of novel function following gene duplication (termed *neofunctionalization*) is well-documented and understood, and the model for this was originally proposed by Ohno [1, 6]. Following duplication one paralog neutrally accumulates mutations due to relaxed purifying selection from redundancy, while the other paralog is preserved with the ancestral function. However, gene silencing never occurs; instead, beneficial mutations arise and become fixed, allowing for functional divergence and paralog retention. Often, the 'novel function' is not truly novel but rather a variation of the ancestral function, perhaps better suited for a particular environment. For example, Zhang *et al* identified a case of neofunctionalization in a pancreatic *RNase* gene in the douc langur (*Pygathrix nemaeus*). They determined that colobine primates possessed a unique duplication of the *RNase1* gene, and that one paralog (*RNase1B*) underwent nucleotide substitution and positive selection, resulting in functional divergence allowing for *RNase1B* catalytic activity at low pH (a property of the douc digestive system) [20].

Other models exist for the preservation of duplicate-genes following WGD. Simultaneous genetic modification of both copies of a pleiotropic gene can partition ancestral functions across both paralogs, such that both are maintained by selection, [9, 17, 21, 22] a process called subfunctionalization. Force and Lynch described a model in which this occurs by complementary degradation of each paralog — the duplication-degeneration-complementation (DDC) model of subfunctionalization [9, 13]. The DDC model proposes that neutral mutations due to relaxed

purifying selection in both paralogs necessitates the retention of both duplicate-genes to perform all ancestral activities [9, 13]. Additionally, subfunctionalization is not specific to sequence-level genetic modifications, and can occur or be accompanied by protein- and expression-level divergence between paralogous genes [9, 13, 17]. In fact, gene expression variation between paralogous genes appears to occur quickly and frequently [17, 23, 24, 25, 26]. It is also crucial to note that these mechanisms are not independent, and multiple mechanisms contribute, either tandemly or successively, to the stabilization of paralogous genes. Therefore, it is important to consider the breadth of processes that affect paralogous genes following duplication to determine the probability of duplicate-gene retention and the evolution of functional divergence between paralogs.

Furthermore, it is now widely believed that gene duplication occurs across all domains of life; this has been shown in such model species as the fruit fly *D. melanogaster* [27], the common yeast *Saccharomyces cerevisiae* [27, 28, 29], the mustard plant *Arabidopsis thaliana* [24, 30], African clawed frogs in the genera *Xenopus* and *Silurana* [11], and of course *Homo sapiens* [31]. Thus, investigating post-duplication evolution and mechanisms of duplicate-gene retention is of high importance to understanding how functional divergence and speciation has occurred across all domains of life.

1.2 The Vertebrate Two-Rounds Hypothesis

Ohno famously proposed that two rounds of whole-genome duplication occurred immediately prior to the origin of vertebrates [1, 6]. For many years though, this hypothesis has been thoroughly debated [32, 33, 34, 35, 36, 37, 38, 39]. However, evidence for ancient WGD in the evolutionary history of many lineages has emerged

as well, and is generally supported [11, 27, 28, 30], including a potential third round of WGD in teleosts [40]. Such species are termed paleopolyploid; that is, they underwent an ancestral WGD in their lineage, but have since been functionally diploidized [37].

Current research supports one WGD event following the divergence of urochordates but before the divergence of jawless vertebrates, and a second WGD event following the divergence of jawless vertebrates but before the divergence of jawed vertebrates, although this is still debated [41, 42]. Evidence from the sea lamprey genome project suggests one or two rounds of WGD prior to the divergence of lampreys and gnathostomes [43]. This observation is termed the Two-Rounds (2R) Hypothesis, and is believed to have contributed strongly to the adaptive radiation of vertebrates.

Evidence for the 2R hypothesis comes from multiple sources. For example, strong support comes from *HOX* gene clusters in mammals and the cephalochordate amphioxus, where it is found that *HOX* gene clusters are in a $4:1$ ratio [44, 45]. Further evidence has emerged from other gene families, specifically globin genes where the same $4:1$ ratio has been described between humans and *Amphioxus* [41, 46, 47]. Despite criticism of the $4:1$ rule, originating from unexpected tree topologies of duplicate genes [33], these relationships are often explained by divergent evolutionary rates of paralogous sequences [48]. Finally, a study by Dehal and Boore [36] found that the distribution of homeologs (paralogous sequences generated by WGD) in the human genome strongly supports the $4:1$ ratio hypothesis and the 2R hypothesis in vertebrates.

1.3 Whole-Genome Duplication in Plants and Animals

The prevalence of WGD is much higher in plant lineages than in animal lineages [15, 49, 50, 51]. Muller [49] was one of the first to suggest that the disparity of incidence of polyploidy between plants and animals was based in the sex-determining mechanism. However, the lack of polyploidy in dioecious mammals cannot be explained by disruption of sex-determination alone, as many dioecious plant genera contain polyploid species [52]. Subsequently, Orr [50] modified Muller's hypothesis, demonstrating that heteromorphy of mammalian sex chromosomes likely results in lethal genotypes (for example, through unbalanced gene dosage).

That is not to say that polyploid incidence is unheard of in animals. Among parthenogenic animals and animals with homomorphic sex chromsomes, polyploidy is disproportionately represented [53]. Interestingly, African clawed frogs in the genera *Xenopus* and *Silurana*, a focus of this thesis that are noted for a high degree of polyploidy, are known to possess homomorphic sex chromosomes [54]. That is to say, while polyploidy is indeed exceptionally rare in birds and mammals, among animals with other sex-determining systems, polyploidy may be surprisingly well tolerated.

Similar to the vertebrate 2R hypothesis, it has been proposed that two rounds of WGD occurred ancestrally in plants, once before the diversification of seed plants and once before the divergence of angiosperms, implying that 100% of flowering plants are paeleopolyploid [55, 56]. It has been estimated that subsequent polyploidization has occurred in up to 70% of angiosperms [57]. Comparatively, polyploidy in animals is mostly limited to amphibians, including frogs and salamanders [15, 58, 59].

Given the abundance of polyploid plants, it is unsurprising that many everyday plants are polyploid, such as wheat (*Triticum aestivum*), coffee beans (*Coffea arabica*), cotton (*Gossypium hirsutum*), maize (*Zea mays*) and the scientific model species, *Arabidopsis thaliana*. Coupled with the overall high incidence of polyploidy in plants [15], the success of these species as agricultural products is thought to be related to the genomic plasticity offered by polyploidization, and contributed to their domestication by allowing for rapid evolutionary changes following genome duplication [60, 61, 62, 63, 64].

1.4 Gene Duplication in Human Disease

As described above, signals of gene and genome duplication exists in modern humans and provide evidence for the 2R Hypothesis [31, 32, 36, 38, 65]. However, examining gene duplication through the lens of evolutionary catalysis and functional divergence fails to convey the unfortunate detrimental role duplication has in human disease and development.

Well-known instances of gene duplication in humans are trisomic chromosomal duplications resulting in aneuploidy. Non-disjunction of a homologous chromosome pair during meiosis characterizes and often causes defects in normal development. In many cases, autosomal chromosome trisomies exhibit embryonic lethality, although often there is a low but non-zero chance of survivorship through birth. Most notably, Trisomy 21 (Down Syndrome) patients often have a life expectancy of > 50 years in more economically developed countries where access to medical care is available and constant. For most other chromosomal trisomies, survivorship through birth and past infancy are extremely low, including Trisomy 8, Trisomy 13 (Patau Syndrome) and

Trisomy 18 (Edwards Syndrome) [66]. Interestingly, compared to autosomal trisomy disorders, aneuploid sex-chromosome syndromes – such as tetrasomic X, triple X syndrome, Klinefelter syndrome (XXY) and trisomy XYY syndrome – have a much higher individual survivorship and life expectancy. Often, individuals with sex-chromosome aneuploidy develop normally and are otherwise healthy [67].

The neurodegenerative disorders Parkinson's disease and Alzheimer disease also have phenotypic forms associated with gene duplication. Parkinson's disease is characterized by Lewy bodies, aggregates of protein that develop inside nerve cells, a primary component of which is α-synuclein. In some cases of Parkinson's disease, it has been found that duplication or even triplication and quadruplication of the α-synuclein gene (*SNCA*) results in overexpression, and the degree of overexpression is correlated with the intensity of Lewy bodies [68, 69]. In Alzheimer disease, amyloid plaque accumulation is a major factor in pathogenesis and progression. Similar to α-synuclein in Parkinson's disease, duplication of the amyloid β precursor protein (*APP*) is associated with increased severity of amyloid plaque formation [70].

Another example of gene duplication in human disease comes from Charcot-Marie-Tooth (CMT) disease, a high-incidence neuropathy of the peripheral nervous system [71, 72, 73]. In CMT disease, it has been found that a $\approx$ 1.4-MB gene duplication on the small arm of chromosome 17 is present in over 60% of clinical patients [71, 72, 74]. Specifically, gene dosage disruption of the peripheral myelin protein *PMP-22* contained within the region of gene duplication is characteristic of the disease phenotype [75, 76]. Trisomy of the *PMP-22* gene results in demyelinization and nerve conduction deficiency.

Despite the evolutionary role of gene duplication discussed in previous sections,

individual gene duplication events are often not well-tolerated, resulting in embryonic lethality, premature death or diminished quality of life. The study of gene duplication, then, encompasses the role of evolutionary mechanisms of innovation and speciation, as well as the role of gene duplication in human health and disease.

1.5 The Scope of This Thesis

In this thesis, I investigate functional divergence following WGD in a currently unnamed tetraploid African clawed frog, *Silurana* new tetraploid. I borrow this name from Evans *et al.*, who referred to it as *S.* new tetraploid 1 [77]. Using a robust transcriptome assembly generated from RNA-sequencing, I identified paralogous gene pairs in over six-thousand *S.* new tetraploid genes. By comparing *S.* new tetraploid paralogs with orthologous sequences in the closely-related diploid species, *S. tropicalis*, I examine the evolutionary relationship between these two species as well as the paralogs within *S.* new tetraploid.

Through this method, the study investigates the mechanism of WGD in *S.* new tetraploid and the genetic divergence attributable to WGD. It is a comprehensive study which examines both molecular evolution of singleton and duplicate genes, and divergent expression of homeologous transcripts following WGD. It validates the hypothesis of an allopolyploid mechanism of speciation in *S.* new tetraploid, indicative of hybridization between diploid ancestors, including the ancestor of the only extant diploid *Xenopus* or *Silurana* lineage. Branch models of molecular evolution reveal differential selective pressures among homeologs and singletons after gene duplication, and we test hypotheses for relaxed purifying selection in *S.* new tetraploid predicted by WGD. Additionally, the study provides evidence for expression-level divergence

attributable to gene duplication between each sex and across multiple tissue types. I then compare this finding to similar studies with post-duplication models of sex-biased expression evolution that show how gene duplication may resolve sexually antagonistic selection.

This study is one of the few to examine gene and expression divergence and evolution following WGD in a tetrapod species, with direct comparison to an extant diploid sister taxa. The use of RNAseq and a transcriptome-based design is a contemporary method through which to study both sequence and expression evolution. Additionally, the study provides a framework for future research on the topic of sexually dimorphic expression evolution following WGD in amphibious tetrapods.

Chapter 2

Sex-specific Transcriptome Evolution in an Allotetraploid African Clawed Frog

2.1 Introduction

2.1.1 Gene duplication and sexual antagonism

Genetic recombination associated with sexual reproduction offers several advantages over asexuality [78] yet creates a genomic arena for sexual antagonism, wherein alleles advantageous for one sex are deleterious for another [79, 80]. Sexual antagonism can be mitigated if alleles acquire sex-biased inheritance due to linkage to sex chromosomes [81, 82], if they acquire a sex-specific splicing mechanism [83], and/or if they acquire a sex-biased pattern of expression, such as being controlled by sex hormones (reviewed in [84]). Gene duplication could also resolve sexual antagonism

by allowing polymorphic ancestral alleles with differing sex-specific fitness effects to independently fix in different paralogs, which can then diverge in expression in a sex-biased fashion. Similarly, sex-biased divergence of paralogous expression could present opportunities for alleles with sexually antagonistic function to evolve after duplication. Gene duplication can also catalyze the movement of genes whose alleles have sexually antagonistic function [81] to or from sex chromosomes [85].

Several factors apart from the resolution of sexual antagonism could promote the functional persistence of duplicated genes. Novel function, or neofunctionalization, could arise by chance in one paralog while the other retains the ancestral function [1, 86]. Examples of neofunctionalization after gene duplication include a segmental duplicate of the *DMRT1* locus, which became a novel female-specific sex determining gene [87, 88]. Alternatively or in addition, both duplicated paralogs could accumulate complementary deleterious mutations that knock out complementary subfunctions in each paralog (subfunctionalization) making both paralogs necessary to perform the entirety of the ancestral functions [9, 13]. Similarly, activity compromising substitutions could also promote the persistence of both paralogs without necessitating that the ancestral gene have multiple functional domains [89]. Paralogs of the *engrailed* gene, for example, show complementary patterns of expression in zebrafish that together resemble the expression pattern of a singleton gene in an outgroup species [9]. Novel function could also evolve before gene duplication and then act to promote duplicate genes; this happened multiple times in the *opsin* gene of primates [90, 91]. Following large scale gene duplication by polyploidization, the retention of both duplicate gene copies could be favored by natural selection in order to balance the dosage of suites of duplicated interacting proteins [9, 13, 92]. This mechanism for

duplicate gene persistence need not involve functional change.

Consistent with this mechanism is the observation that duplicate genes generated from WGD tend to persist longer than duplicates generated by segmental duplication [12, 38, 93]. Duplicated genes with pleiotropic function, multi-domain proteins, high expression levels, or a low rate of evolution also tend to persist longer than those that lack these characteristics [94, 95, 96, 97].

2.1.2 Polyploid African clawed frogs

Our goal is to better understand how transcriptomes evolve after whole genome duplication, including divergence between the sexes and across tissue types. We use as a model an unnamed tetraploid species of African clawed frog (Anura, Pipidae, Xenopus). We refer to this species as *Silurana* new tetraploid, following Evans, 2004 [77]. This species is interesting in several respects. First, EST databases are available from a closely related diploid species (*S. tropicalis*) and a tetraploid species (*X. laevis*; [98, 99]). Second, tetraploidization occurred independently in *S.* new tetraploid and *X. laevis* [100], making *X. laevis* both a useful outgroup and an interesting genome for comparison to *S.* new tetraploid (see Figure 2.1). And third, based on phylogenetic analysis of cloned paralogs, the mechanism of tetraploidization of *S.* new tetraploid is hypothesized to be allopolyploidization [100, 101, 102]. Because allopolyploidization of the *S.* new tetraploid ancestor is hypothesized to have involved a recent diploid ancestor of *S. tropicalis*, phylogenetic analyses have the potential to distinguish which paralog within each pair of duplicated genes was inherited from each diploid ancestor, thereby making possible tests related to subgenome biases in expression divergence.

Molecular studies indicate that *S. tropicalis* lacks a large sex-specific region on the

sex chromosomes [103], and cytogenetic studies suggest that the sex chromosomes of *S.* new tetraploid are also homomorphic [54]. This makes this species a useful subject with which to evaluate expression divergence between males and females whose genomes harbor only a very small genomic region with sex-specific inheritance. Because sex-specific divergence in expression should be a key mechanism with which to resolve sexual antagonism, we expected to recover evidence of pervasive sex-biased expression patterns and also to find that sex-biased expression would be higher in duplicate genes than in singleton genes.

In this study, we first use a large RNA-sequencing (RNAseq) dataset and orthologous sequences from *S. tropicalis* and *X. laevis* to identify functional paralogous pairs and putative singletons (i.e. the surviving paralog of a WGD pair where one was pseudogenized) in this species. We use these data to test the hypothesis that the ancestor of *S.* new tetraploid formed by allopolyploidization between the ancestor of *S. tropicalis* and another putatively extinct diploid ancestor. We then use these same data to explore sequence-level and expression-level divergence between *S.* new tetraploid paralogs, and between *S.* new tetraploid paralogs and *S.* new tetraploid singletons, including comparisons of expression differences in males and females. Our results support an allopolyploid origin of *S.* new tetraploid, identify a strong signature of sexually dimorphic expression across tissue types, and characterize fundamental differences in expression and molecular evolution of duplicate and singleton genes in this frog.

2.2 Methods

2.2.1 Experimental design and RNA-sequencing

Three female and three male adult *S.* new tetraploid individuals were euthanized by transdermal exposure to a saturated solution of tricaine mesylate (tricaine methanosulfonate; MS-222, Sigma). From each carcass we immediately harvested brain, heart and liver tissue, and extracted total RNA from each tissue type using an RNeasy Mini Kit (Qiagen) following the manufacturer's protocol. Complete transcriptomes from each tissue sample were sequenced using Illumina TruSeq $v2$ for paired-end RNA-seq with the 18 transcriptomes (one for each of the three tissue types from each of the six individuals) multiplexed over three Illumina lanes.

Adapter sequences and poor-quality read fragments were removed using Trimmomatic (v0.33; [104]). Transcriptome assembly was performed using a *de novo* approach with Velvet-Oases (v0.2.08; [105, 106]). For each cDNA transcript library, we annotated start and stop codons based on homology with other transcriptomes with high confidence annotations from four species including humans, zebrafish (*Danio rerio*), chicken (*Gallus gallus*), and mouse (*Mus musculus*) [107]. As detailed below, we then analyzed the transcript libraries from each tissue type from each individual.

2.2.2 Identification of paralogous and orthologous sequences

Here, we use the term homeolog to describe paralogous genes formed from a WGD event, and the term segmental duplications to refer to within-genome duplication of individual genes. Because *S.* new tetraploid is a tetraploid, we expect each ortholog in *S. tropicalis* should have two homeologous genes in *S.* new tetraploid, except for *S.* new

tetraploid homeologs in which one became a pseudogene, one or both homeologs were not expressed in the transcriptomes we sequenced, or one or both homeologs were not sequenced due to missing data. If *S.* new tetraploid underwent whole-genome duplication (WGD) via allopolyploidization between a diploid ancestor of *S. tropicalis* and another (unidentified) diploid species, we expect that pairs of *S.* new tetraploid homeologs should consist of one homeolog that is more closely related to the *S. tropicalis* ortholog (hereafter α homeologs) than to the other *S.* new tetraploid homeolog (hereafter β homeologs; Figure 2.2). If *S.* new tetraploid underwent WGD via autopolyploidization, we expect *S.* new tetraploid paralogs to be monophyletic with respect to an *S. tropicalis* ortholog. A summary of these expectations is provided in Figure 2.2. We included as an outgroup ortholog(s) of *X. laevis* identified from an unpublished cDNA database of unique sequences provided to us by the Xenopus Genome Project Consortium, containing 35,543 genes. We included one or two *X. laevis* homeologs for each gene analyzed depending on whether one or two were identified.

Our methods to identify *S.* new tetraploid homeologs were similar to those used previously in other studies [22, 108]. We used an *S. tropicalis* cDNA database from Ensembl ($v7.2$), which contained 19,921 genes, and the aforementioned unpublished *X. laevis* cDNA database, which contained 35,543 genes, to identify homeologs and orthologs. Alternatively spliced transcripts from each of these transcript databases were filtered to include only one exemplar sequence from each gene. We used BLASTn to attempt to match *X. laevis* homeologs with each *S. tropicalis* ortholog. We used each *X. laevis* sequence as a query against the *S. tropicalis* database, and saved one top hit to *S. tropicalis* if the expect score was $< 1 \times 10^{-10}$ and the length of the match was > 100 bp. We then used each *S. tropicalis* sequence as a query against the

X. laevis database, saving one or two top hits with the same criteria. We then used a script written in Perl to match each *S. tropicalis* ortholog with one or two putative *X. laevis* homeologs.

BLASTn was used to match *S. tropicalis* transcripts queries to *S.* new tetraploid transcripts, saving all *S.* new tetraploid hits below an effect value of 1×10^{-10}. This identified a suite of *S.* new tetraploid sequences that were putative homeologs to many of the *S. tropicalis* sequences (13,400 of 19,921 *S. tropicalis* sequences with one or two *X. laevis* homeologs). If multiple *S.* new tetraploid transcripts aligned to the same *S. tropicalis* sequence, we considered all of them to be putative paralogs. *Silurana* new tetraploid genes for which at least one *X. laevis* and one *S.* new tetraploid orthologous sequence could not be inferred were excluded from further analysis. The results from BLASTn were parsed with a custom Python script including functions from Biopython [109].

We used a phylogenetic approach to identify sets of α and β homeologs in *S.* new tetraploid and to distinguish them from other *S.* new tetraploid sequences of unknown origin or phylogenetic affinities. Gene sets consisting of one *S. tropicalis* ortholog, one or two *X. laevis* homeolog(s), and at least one *S.* new tetraploid ortholog were aligned with MAFFT (v7; [110]), allowing the program to automatically select an optimized alignment algorithm, and allowing for sequence reverse-complementation. Identical *S.* new tetraploid sequences from each gene set were clustered with CD-HIT (v4.6; [111, 112]). Using aligned gene sets, we then constructed a maximum-likelihood (ML) tree with RAxML (v8.00; [113]). For each gene, a ML phylogeny was constructed with the default rapid hill-climbing algorithm with a general time reversible (GTR) model, and a gamma distribution of rate heterogeneity (GTR-GAMMA model in RAxML).

Using scripts written in Perl, we then generated a series of backbone constraint trees (BCTs) for each gene set that were designed to evaluate the phylogenetic affinities of the suite of putative *S.* new tetraploid homeologs with respect to the *S. tropicalis* and *X. laevis* sequences. First, we checked for monophyly of *S.* new tetraploid transcripts. Under the hypothesis of allopolyploidy, monophyly of the *S.* new tetraploid sequences suggests that only one *S.* new tetraploid homeolog was present in the *S.* new tetraploid data. Under the hypothesis of autopolyploidy, monophyly of the *S.* new tetraploid sequences suggests that either one or two *S.* new tetraploid homeologs are present. Under the hypothesis of allopolyploidy involving a recent ancestor of *S. tropicalis*, an observation of paraphyly of *S.* new tetraploid transcripts with respect to an *S. tropicalis* ortholog is consistent with the possibility that both *S.* new tetraploid homeologs were present in the *S.* new tetraploid data (i.e., an α and a β homeolog).

Silurana new tetraploid sequences frequently clustered into only one or two lineages, and we inferred transcript homeology from filtering with BCTs consistent with allopolyploidy. However, in some genes we identified transcripts with additional lineages that did not fit our expectations for either autopolyploidy or allopolyploidy. because the *S.* new tetraploid sequences clustered into more than two lineages. These supernumerary lineages could be due to segmental duplications within gene families, chimerical sequences (i.e., errors) from transcriptome assembly, errors in phylogenetic inference, or other unknown causes. For these genes, we selected the two lineages most closely related to *S. tropicalis* ortholog, using a script written in Perl, and assumed that these represented the two *S.* new tetraploid homeologs. The *S.* new tetraploid sequences from the other lineages (which branch closer to *X. laevis* than to *S. tropicalis* in an unrooted phylogeny) were discarded from all further analyses. As detailed

above, when two *S.* new tetraploid homeologs were recovered, we designated the one that was most closely related to *S. tropicalis* as the α homeolog and the other the β homeolog. This exercise identified one or two *S.* new tetraploid homeologs with homeolog relatedness and identity for each of 6,516 gene sets, each of which included one *S. tropicalis* ortholog, and one or two *X. laevis* homeologs.

2.2.3 Molecular evolution of expressed duplicates and singletons

We used branch models [114] implemented by Codeml (v4.7), a program within PAML [115], to test for differences in the rate-ratio of non-synonymous to synonymous substitutions per site (hereafter referred to as the d_N/d_S ratio) within genes for which we identified only one *S.* new tetraploid homeolog (a singleton) and genes for which we identified two *S.* new tetraploid homeologs. To perform analyses, we first selected the longest transcript sequence from one or both *S.* new tetraploid homeologs (depending on whether we identified one or two *S.* new tetraploid homeologs) from each gene set. Gene alignments of these *S.* new tetraploid sequences, the corresponding *S. tropicalis* ortholog, and one or two *X. laevis* orthologs were then realigned using the codon-aware algorithm implemented by the program MACSE (v0.8; [116]). Following this, the stop codon was removed from each sequence, as required by Codeml.

As discussed below, our phylogenetic analyses support the hypothesis of an allopolyploid origin of *S.* new tetraploid because many genes had both an α and a β *S.* new tetraploid homeolog. For these genes, we evaluated the hypothesis that WGD is associated with a relaxation of purifying selection, and therefore a higher d_N/d_S in the *S.* new tetraploid α homeolog than the *S. tropicalis* ortholog. More specifically,

we tested whether d_N/d_S differed between the *S. tropicalis* ortholog and its *S.* new tetraploid α-homeolog. The null model has no difference in the d_N/d_S of these lineages and the alternative allows this parameter to vary among lineages (Figure 2.3a). The significance of the difference between these models was evaluated with a likelihood ratio test for each gene (with the χ^2 statistic obtained from two times the difference in the log-likelihood of each model). A combined probability test was performed across all genes using Fisher's method:

$$\chi^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$$

For genes with two *S.* new tetraploid homeologs, we also evaluated the hypothesis that each *S.* new tetraploid homeolog evolves under a unique level of purifying selection. This was accomplished by evaluating a null model in which the d_N/d_S ratio between homeologs was fixed, and comparing it to an alternative model in which the ratio was estimated separately for each homeolog (Figure 2.3b). As above, we evaluated significance of the alternative model over the null model with Fishers method.

Finally, we used d_N/d_S values to evaluate the hypothesis that *S.* new tetraploid singleton genes are under a different level of purifying selection from their *S. tropicalis* ortholog. As above, the null model had no difference in the d_N/d_S of the *S. tropicalis* and *S.* new tetraploid lineages, while the alternative model allowed this parameter to vary among these two lineages (Figure 2.3c). We again evaluated the significance of the alternative model over the null model with Fisher's method.

2.2.4 Quantification of transcript abundance for expression studies

We used eXpress (v1.5.1; [117]) to quantify transcript abundance based on the RNAseq data. We estimated abundance separately for each of the eighteen transcript libraries (each of three tissue types from each of six individuals).

Bowtie2 (v2.2.2; [118]) was used to index transcript sequences and perform read-mapping. Each library was indexed separately with an offrate of 1. Here, the offrate specifies how many Burrows-Wheeler rows are annotated with their genomic location; an offrate of 1 specifies that every second row is annotated, which maximizes efficiency [119]. Transcript libraries were indexed to facilitate and expedite read mapping. Following this, Bowtie2 was used to align processed RNA-seq reads from each tissue type of each individual to their respective transcriptome library, using a read and reference gap open penalty of 6 and an extend penalty of 5. The minimum alignment score function was set to $f(x) = -0.6 - 0.4x$ where x is the read length. Bowtie2 produces a SAM-formatted table which was piped to eXpress for quantification of transcript abundance. This provided an estimated count value of transcript abundance for each of the ≈4.2-million transcripts across all libraries.

We minimized the count set to include only *S.* new tetraploid transcripts in each of the 6,516 high-confidence gene sets previously identified by BLASTn. With custom Python scripts, we identified for each *S.* new tetraploid transcript the *S. tropicalis* orthologue based on previous steps (See Section 2.2.2) and only included those genes for which homologous sequences had been previously determined. These were genes for which we could identify an orthologous *S. tropicalis* sequence and a homologous

X. laevis sequence, and for which we have estimated non-synonymous and synonymous rates of substitution. We then extracted the total count estimate for each of these transcripts as provided by eXpress, and retrieved the relevant d_N and d_S values estimated previously by Codeml using the longest transcript for each homeolog. Transcript sequences for which d_N and d_S values were absent, because no *S. tropicalis* or *X. laevis* homeolog was identified, were not included in expression analyses.

2.2.5 Models for differential homeolog expression

We performed two separate analyses to test for expression-level divergence in *S.* new tetraploid. The first analysis evaluated expression-level divergence between homeologs among retained duplicate genes in *S.* new tetraploid. *S.* new tetraploid singletons were not included in this analysis. For this, we obtained a count value for each combination of sex, tissue and homeolog type for each gene by summing homologous transcripts from the same library. Libraries from which no transcript count estimate was determined were assumed to have a count of zero. Since a given gene could have many transcripts, either from multiple libraries or within a library, in total the data set contained 38,326 non-zero count estimates for 4,494 unique gene sets.

We then fit a generalized linear mixed-effects model (GLMM) to the count data with a negative binomial conditional distribution, of the form:

$$Y_i \sim \text{NegBinom}(\mu_i, \theta)$$

$$\mu = \exp(X\beta + Zb)$$

$$b \sim \text{MVN}(0, \Sigma)$$

Hence, the model has two parameters: the mean, μ_i and the dispersion, θ. The model contains a three-way interaction term between sex, tissue type and homeolog type modeled as fixed effects, with genes modeled as a random effect. Because sampling intensity may differ between transcript libraries, we added a small value (an offset) to the expression levels of each gene. For each library we calculated an offset equal to the median of the ratio of the counts to the geometric mean of the counts, as described by Anders and Huber, 2010. [120]:

$$\hat{s}_{ij} = \underset{i}{\text{median}} \frac{K_{ij}}{(\prod_{v=1}^{m} K_{ij})^{1/m}}$$

Here, K_{ij} represents the positive non-integer estimated count value, and normalization with this method was performed separately for singleton and duplicate genes. This model was fit with `glmer.nb` from the `lme4` package [121] in R (v3.2.0). We compared the fit of the model with an AIC value and verified the normality of the residuals with a Q-Q plot (Supplemental Table S3; Supplemental Figure S2).

A second analysis of *S.* new tetraploid gene expression evaluated patterns of expression between *S.* new tetraploid singleton and each duplicate gene homeolog. Therefore, we updated our model categories to indicate gene status as either singleton, alpha or beta. Offsets for each library were again calculated using the method from Anders, 2010 [120] (these offset values were different than the first analysis due to summation of homeolog counts in duplicate genes). Then, we again fit a GLMM with a negative binomial distribution, of the same form and with the same parameters, to the count data using `lme4`. As above, this model contained a three-way interaction term for tissue type, gene status (singleton, α or β) and sex modelled as fixed effects, with genes modelled as a random effect.

For each model we evaluated the fixed effects for the three-way interaction term with multiple contrasts. Contrasts were designed to compare β-homeologs to α-homeologs, singleton genes to duplicate genes, males to females, and tissue types (brain, heart, and liver). As above, we evaluated the fit of the model with an AIC value and verified the normality of the residuals with a Q-Q plot (Supplemental Table S4; Supplemental Figure S3).

In addition to total expression (T), we calculated two other non-independent expression summary statistics: expression evenness (E) and expression intensity (I), from Chain *et al.*, 2011 [108]. Expression intensity describes the mean expression of a gene across all tissues rather than by a tissue, and is given by $I = \sum (L_i^2) \,/\, \sum L_i$; therefore, a gene that is highly expressed in only a few tissues will have a high expression intensity. Conversely, expression evenness describes how broadly a gene is expressed, and is given by $E = (\sum L_i)^2 \,/\, \sum (L_i^2)$; therefore, a gene with relatively even expression distribution across tissue types will have an elevated evenness score [108].

2.3 Results

2.3.1 Transcriptome sequencing and assembly

Paired-end, multiplexed RNA Illumina TruSeq sequencing provided 'raw', unprocessed 100 bp RNA reads for eighteen libraries. These reads were processed with Trimmomatic, which removed adapter/barcode sequences that were inserted from the TruSeq $v2$ protocol, and removed short, damaged or otherwise poor-quality reads. After processing, the RNA read data set used for transcriptome assembly contained

a total 414,865,589 reads, across the eighteen libraries (3 males, 3 females, and 3 tissues per individual). Transcriptome assembly was performed separately on each library, and 4,258,556 complete transcript sequences were assembled in total across all libraries. A by-library description of read counts and transcript sequences is provided in Supplementary Table S1. For processed read libraries, the average number of reads per library was 23,048,088 reads. The mean number of transcript sequences per library was 236,586 sequences. The distribution of assembled transcript sequence lengths is summarized in Supplemental Figure S1. The range of the length of assembled transcripts was 101 bp to 92,002 bp and the mean and median length was 2,537 bp and 2,153 bp, respectively.

2.3.2 An allopolyploid origin of *S.* new tetraploid

We identified 13,394 *S. tropicalis* genes for which at least one orthologous *S.* new tetraploid transcript and at least one orthologous *X. laevis* sequence were identified. Multiple *S.* new tetraploid transcripts that best aligned to a single *S. tropicalis* gene were assumed to be genes that were expressed in more than one treatment/library, splice variants, or segmental duplicates. Gene sets containing *X. laevis*, *S. tropicalis* and *S.* new tetraploid sequences were aligned with MAFFT. We filtered these data by requiring a minimum sequence length of 300 bp, allowing for a maximum alignment gap of 40 bp, resulting in a reduced data set size of 6,516 high-quality alignments, which formed the basis of our study.

With a custom phylogeny-based bioinformatics approach described in Section 2.2.2,

we then determined which gene sets supported monophyly of *S.* new tetraploid transcript sequences with respect to *S. tropicalis*, and which supported paraphyly of transcript sequences with respect to *S. tropicalis*. Of the set of 6,516 genes, we found 4,494 genes from which the presence of at least two paraphyletic *S.* new tetraploid lineages were inferred, and 2,022 genes for which *S.* new tetraploid transcript monophyly was inferred. Of the 4,494 genes containing at least two *S.* new tetraploid homeologs, 3,986 contained exactly two *S.* new tetraploid lineages and 508 (11.3%) contained more than two *S.* new tetraploid lineages; for these latter alignments, we pruned the more diverged *S.* new tetraploid lineages as described above, leaving only the two that were most closely related to the *S. tropicalis* ortholog. Of the 4,494 genes containing at least two *S.* new tetraploid lineages, 2,701 also had two *X. laevis* homeologs; the other 1,793 genes had only one *X. laevis* homeolog. Of the 2,022 genes with one *S.* new tetraploid lineage, 1,200 had two *X. laevis* homeologs and 822 had one *X. laevis* homeolog. As described above, *S.* new tetraploid and *X. laevis* are both tetraploid species that originated by independent genome duplication events. In order to test for a correlation between duplicate gene retention in *X. laevis* and *S.* new tetraploid we used Fishers exact test. This test did not recover a significant association for interspecies duplicate-gene retention ($p > 0.5$)

If *S.* new tetraploid originated by autopolyploidization, we expect a monophyletic relationship of *S.* new tetraploid homeologs with respect to the orthologs of other species such as *S. tropicalis* could also arise if *S.* new tetraploid originated via allopolyploidization involving two diploid species that were more closely related to each other than either is to *S. tropicalis*. However, in agreement with previous studies based on a much smaller sample of genome region [17, 101, 102] a preponderance of genes with a

paraphyletic relationship with respect to the *S. tropicalis* ortholog (4,494 out of 6,516; 69%) supports the hypothesis that *S.* new tetraploid originated by allopolyploidization, and that this allopolyploidization involved a diploid ancestor of *S. tropicalis* and another unidentified, and possibly extinct, diploid species.

2.3.3 Expression divergence in *S.* new tetraploid

RNAseq read-mapping performed with Bowtie2 had high efficiency; for each library, we obtained a percent-alignment above 65% with a mean percent-alignment of 72.5%. Total read count values were estimated with eXpress for the 4,258,556 transcripts summed across libraries. Total count values for transcripts included in the 6,516-gene data set were extracted, indexed by *S. tropicalis* Ensembl ID, and homologous *S.* new tetraploid transcripts were grouped by sex, tissue type and homeolog. For genes with multiple homologous transcripts from the same library with the same orthologous gene, the count value was summed across all transcripts.

A boxplot with median expression levels is presented in Figure 2.4. We found that median expressions of singleton genes was higher in each tissue (Brain, B; Heart, H; Liver, L) than either duplicate gene homeolog, in both males ($\widetilde{x}_{B,M} = 5350.0, \widetilde{x}_{H,M} = 1416.5, \widetilde{x}_{L,M} = 1221.0$) and females ($\widetilde{x}_{B,F} = 8359.5, \widetilde{x}_{H,F} = 1539.5, \widetilde{x}_{L,F} = 1181.5$).

To test whether these qualitative differences were significant, we performed two analyses of expression-level divergence in *S.* new tetraploid. First, we tested differential expression between homeologs was attributable to sex and/or tissue type. Second, we tested whether patterns of expression across tissue types differed between duplicate and singleton genes.

To investigate differential expression between *S.* new tetraploid homeologs, we fit

a GLMM with a negative binomial conditional distribution (See Section 2.2.5). We found significant main effects of the sex of the individual ($\chi^2_1 = 880.1$, $p < 0.001$), the tissue type of the library ($\chi^2_2 = 2442.2$, $p < 0.001$), and the origin of the homeolog ($\chi^2_1 = 10.4$, $p = 1.28 \times 10^{-3}$). Additionally, there were significant interaction effects between the sex and the tissue type ($\chi^2_2 = 983.6$, $p < 0.001$), and the tissue type and the homeolog origin ($\chi^2_2 = 98.1$, $p < 0.001$), but not between the sex and the homeolog origin ($\chi^2_1 = 2.6$, $p > 0.05$). Importantly, we also found a significant three-way interaction effect between sex, tissue type and homeolog origin ($\chi^2_4 = 776.7$, $p < 0.0001$), indicating that the tissue type and homeolog origin interaction was different in males and females. This indicates that the extent of expression divergence between *S.* new tetraploid homeologs is sexually dimorphic. These results are summarized in Supplemental Table S3.

We used multiple contrasts to further interpret the significance of the three-way interaction between sex, tissue type and homeolog origin. The first contrast evaluated whether the significant interaction effect between tissue type and homeolog in heart, compared to brain, was significantly different in males than in females. This contrast was significant ($b = 0.894$, $z = 14.15$, $p < 0.001$), revealing that the variation among α and β homeologs in heart tissue, as compared to brain tissue, is greater in males than in females. The second contrast assessed whether the significant interaction effect between tissue type and homeolog in liver, as compared to brain, was also significantly different in males than in females. Again, this contrast was significant ($b = 0.246$, $z = 3.83$, $p < 0.001$), indicating that expression-level variation among homeologs in liver tissue is different in males and females when compared to brain tissue. The third contrast between liver and heart also recovered a significant overall

expression decrease that was sex- and homeolog-dependent ($b = -0.649$, $z = -8.92$, $p < 0.001$). The results of these models are summarized in Figure 2.5. Overall, these results illustrate that the extent of expression divergence between homeologs in each tissue type is more modest in females than in males.

We then fit a second negative-binomial GLMM to investigate patterns of expression between *S.* new tetraploid singleton and duplicate genes. Here our analysis contains both singleton and duplicate genes, with the duplicate genes classified into α and β homeologs based on their evolutionary affinities to *S. tropicalis* as detailed above. As discussed above, for singletons it was not possible to distinguish α from β homeologs using our phylogenetic approach. We refer to this classification (α, β, or singleton) as the gene status. We found significant main effects for the sex of the individual ($\chi^2_1 = 780.19$, $p < 0.001$), the tissue type ($\chi^2_2 = 2540.02$, $p < 0.001$) and the gene status ($\chi^2_2 = 168.05$, $p < 0.001$). Additionally, we found significant interaction effects between sex and the tissue ($\chi^2_2 = 1108.61$, $p < 0.001$), sex and gene status ($\chi^2_2 = 159.16$, $p < 0.001$) and tissue type and gene status ($\chi^2_4 = 389.51$, $p < 0.001$). Finally, we found a significant three-way interaction between sex, tissue type and gene status ($\chi^2_4 = 270.03$, $p < 0.001$). This indicates that the interaction between gene status and tissue type was different in males and females. These results are summarized in Supplemental Table S4.

We again used multiple contrasts to parse the three-way interaction between tissue type, gene status, and sex. These contrasts compared male and female expression for α and β homeologs to singleton genes across each tissue type. The first contrast evaluated whether the non-significant difference between α-homeolog and singleton gene expression in heart, compared to brain, was different in males and females. This

contrast was significant ($b = -0.158$, $z = -2.42$, $p < 0.001$). The second contrast evaluated whether the significant difference between α-homeolog and singleton gene expression in liver, compared to brain, is different in males and females. This interaction was significant ($b = 0.389$, $z = 5.77$, $p < 0.001$), indicating that variation between α-homeologs and singleton gene expression in liver compared to brain differs between the sexes. The third and fourth contrasts both evaluated the difference between β-homeolog and singleton gene expression in males and females. The third contrast evaluated whether the significant difference in β-homeolog expression compared to singleton gene expression in heart compared to brain, was significantly different in males and females, and it was ($b = 0.736$, $z = 11.37$, $p < 0.001$). The fourth contrast tested whether the significant decrease in β-homeolog expression compared to singleton gene expression, in liver as compared to brain, was significantly different in males and females, and it also was ($b = 0.635$, $z = 9.53$, $p < 0.001$). The results of these models are presented in Figure 2.6. Overall, these results indicate that the variation in gene expression is more substantial in females than in males in duplicate genes and in singleton genes. In both duplicate gene homeologs, female brain expression was much higher than expression in other tissues in either males or females.

An analysis of expression intensity and expression evenness (See Section 2.2.5) revealed patterns of expression between singleton and duplicate genes. Here, an evenness score of 1 indicates observed expression in only one tissue type and an evenness score of 3 indicates equal expression in all three tissues. Similarly, an intensity score equal to the total expression indicates observed expression in only one tissue type, and an intensity score equal to one-third the total expression indicates equal expression in all tissues. We evaluated the significance of these results with Welch's t-test. We

found that singleton genes ($\mu = 1.76$) were significantly more evenly expressed than duplicate genes ($\mu = 1.63; p < 0.001$), corresponding to increased intensity among duplicate genes. When comparing females to males, we noted a non-significant difference in expression evenness in singleton genes ($p > 0.05$) but a significant difference in duplicate genes ($p < 0.001$). Similarly, we found significantly increased intensity in females compared to males in duplicate genes ($\mu_M = 7670.18, \mu_F = 11,052.89$; $p < 0.001$) but not in singleton genes ($\mu_M = 30,440.30, \mu_F = 29,120.59$; $p = 0.606$). The observed increased intensity in females reflects the high expression divergence observed from our models in brain relative to other tissue types.

2.3.4 Sex-biased expression and evolution

As discussed above, the GLMM fits support sex-biased gene expression in singleton and in duplicate genes. To explore this finding on an individual gene basis, we evaluated gene expression divergence between males and females using Welch's Two Sample t-test. In singleton genes, 296 genes exhibit significantly sexually dimorphic expression in at least one tissue type, including male-biased and female-biased expression. Of these, 226 genes (76.4%) appear female-biased in at least one tissue type, and 70 genes (23.6%) appear male-biased in at least one tissue type. We found that female-biased genes are over-represented in brain (163 genes, 72.1%) compared to heart (40 genes, 17.7%) and liver (36 genes, 15.9%). Male-biased genes were similarly more abundant in brain (41 genes, 57.7%) than heart (17 genes, 23.9%) and liver (13 genes, 18.3%). It is apparent that some genes exhibit sex-biased expression in multiple tissue types. Among female-biased genes, we identified 8 (3.6%) that were biased in both brain and heart, 4 (1.8%) biased in both brain and liver, and 1 (0.4%)

biased in both heart and liver; there were no female-biased genes present in all three tissue types. Among male-biased genes, we did not identify any genes with biased expression in both brain and liver or both heart and liver, and only 1 gene (1.4%) biased in both brain and heart.

Among duplicate genes, sex-biased duplicate genes may be paired with another sex-biased homeolog or an unbiased homeolog. Below we describe our findings from each tissue type with respect to these scenarios.

In brain, we found evidence for sex-biased gene expression in 303 genes. Of these, 292 genes (96.4%) showed female-biased expression and only 23 genes (7.6%) showed male-biased expression. In both cases, sex-biased genes were more often found paired with an unbiased homeolog: 275 female-biased genes (94.2%) were paired with an unbiased homeolog and only 11 genes (3.8%) were paired with another female-biased gene; similarly, 16 male-biased genes (69.7%) were paired with an unbiased homeolog and only 1 (4.3%) was paired with another male-biased homeolog. Expression of 6 genes (2.0%) was antagonistically sex-biased in each homeolog, meaning the sex bias was in the opposite direction for each homeolog.

In heart, sex-biased gene expression was identified in 68 genes. Of these, 46 genes (67.6%) showed female-biased expression and 27 genes (39.7%) showed male-biased expression in at least one homeolog. Sex-biased genes were most often found paired with an unbiased homeolog. In heart, we did not find any genes showing female-biased or male-biased expression in both homeologs. Expression of 5 genes (7.3%) was antagonistically sex-biased in each homeolog.

In liver, sex-biased gene expression was identified in 44 genes. Of these, 34 genes (77.3%) showed female-biased expression and 13 genes (29.5%) showed male-biased

expression in at least one homeolog. No genes with male-biased expression were detected in both homeologs, and only 1 gene (2.9%) showed female-biased expression in both homeologs. Expression of 3 genes (6.8%) was antagonistically sex-biased in each homeolog.

We hypothesized that sex-biased gene expression would be accompanied by underlying sequence changes, increasing the respective d_N/d_S ratio. However, using the d_N/d_S ratios discussed in further detail below, we did not recover support for this hypothesis. In singleton genes, neither female-biased nor male-biased genes had significantly higher d_N/d_S values compared to unbiased genes (Mann-Whitney U test: $p > 0.05$). Among duplicate genes, we did not detect higher d_N/d_S values in the α or β-homeologs of either male-biased or female-biased genes (Mann-Whitney U test: $p > 0.05$), except specifically in female-biased genes expressed in brain (Mann-Whitney U test: $p < 0.01$).

We also observed variance of expression evenness and intensity among sex-biased genes compared to unbiased genes. We found female-biased duplicate genes to be less evenly expressed than unbiased duplicate genes ($\mu_F = 1.37$, $\mu_U = 1.59$; $p < 0.001$); however, we did not find a significant difference in male-biased duplicate genes compared to unbiased genes ($\mu_M = 1.64$, $\mu_U = 1.59$; $p > 0.05$). Comparably we found significantly higher expression intensity in female-biased duplicate genes compared to unbiased genes ($\mu_F = 11,609.56$, $\mu_U = 8,729.14$; $p < 0.001$) but not in male-biased genes ($p > 0.05$). Among sex-biased singleton genes, female-biased genes were less broadly expressed than unbiased singletons ($\mu_F = 1.52$, $\mu_U = 1.74$; $p < 0.001$) but we did not detect a significant difference in male-biased genes ($\mu_M = 1.80$; $p > 0.05$).

Importantly, we observed a significant difference in expression evenness between duplicate and singleton genes for both male-biased and female-biased genes. That is, both female-biased and male-biased genes are less broadly expressed in duplicate genes than in singleton genes ($p < 0.001$).

2.3.5 Molecular evolution of duplicate genes

Using Codeml, we tested for differences in the d_N/d_S rate-ratio in the tetraploid *S.* new tetraploid lineage relative to the diploid *S. tropicalis* lineage that were attributable to and indicative of increased tolerance for genetic modification after WGD. We performed separate analyses for singleton genes, suggesting prior pseudogenization of one paralog, and retained duplicate genes in *S.* new tetraploid.

For 2,022 genes we identified only one *S.* new tetraploid paralog was identified; these genes probably include a combination of *S.* new tetraploid singletons in which either the α or β homeolog was pseudogenized, or retained *S.* new tetraploid genes for which the other homeolog was not sequenced or not identified by our bioinformatic pipeline. For this set of genes, we tested two models with Codeml: a null model in which the branch-specific rate parameters were constant for *S.* new tetraploid and *S. tropicalis*, and an alternative model in which the branch-specific rate parameters differed between the two species (Figure 2.3). Using a likelihood-ratio test and assuming a χ^2 distribution with one degree of freedom (df), we calculated a *p*-value for each of the 2,022 genes comparing the alternative model to the null hypothesis. Across all singleton genes, we calculated that $\chi^2 = 6247.9$ with 4,044 df such that $p = 4.3 \times 10^{-99}$ (Fishers method). We found the median d_N/d_S ratio among singleton *S.* new tetraploid genes to be 0.153 under this model; the median d_N/d_S ratio among

S. tropicalis genes under this model was very similar at 0.155.

As detailed above, we performed two analyses using the aforementioned set of 4,494 duplicate genes. With an assumption of an allotetraploid mechanism of WGD in *S.* new tetraploid, we inferred that duplicate genes originating from different sources could exhibit variant substitution and selective pressues. Assuming allotetraploidy, we performed two separate tests: one, for substitution rate variation between *S. tropicalis* and the *S.* new tetraploid α-homeolog; and two, for substitution rate variation between the *S.* new tetraploid α- and β-homeologs (Figure 2.3). As with the singleton gene tests described above, we calculated a p-value for each duplicate gene by assuming the likelihood ratio was χ^2 distributed with 1 df, and then combining probabilities with Fisher's method. We found that our alternative, more parameterized model allowing for rate variation between *S. tropicalis* and the *S.* new tetraploid α-homeolog was better supported than the null hypothesis with $\chi^2 = 11,743.5, p = 2.0 \times 10^{-78}$ and 8988 df. Additionally, an alternate model with extra rate parameterization between the *S.* new tetraploid α- and β-homeologs was also better supported than a null model with equal rates, with $\chi^2 = 9660.8, p = 4.7 \times 10^{-7}$ and 8988 df. In the alternative model, we found the median d_N/d_S ratio of *S.* new tetraploid α and β homeologs to be 0.128 and 0.152, respectively. The median d_N/d_S ratio for *S. tropicalis* sequences in this model was 0.139.

2.4 Discussion

Whole-genome duplication is a disruptive evolutionary event that can catalyze biological innovation. Here we investigated transcriptome evolution following WGD in an unnamed African clawed frog, *S.* new tetraploid, focusing in particular on sex-biased

expression, expression divergence between homeologs, and expression differences between homeologs and singletons. We recovered strong support for a previously hypothesized allopolyploid mechanism of WGD in *S.* new tetraploid [101, 102]. In contrast to several studies in *Drosophila* and birds which found that gene duplication is associated with evolution of male-biased expression patterns [122, 123, 124] (reviewed in [125]), we recovered evidence for the evolution of female-biased expression following WGD, particularly in brain. This result is similar to findings of female-biased expression in the brain of the chicken (*Gallus gallus*) [126]. Our comparisons of homeologs found more substantial expression divergence between homeologs in brain than in liver or heart, and a prominent influence of sex on homeolog expression divergence with expression divergence between homeologs being most pronounced in female brain (Figure 2.4; Figure 2.5). These contrasts are consistent with the notion that expression subfunctionalization or neofunctionalization are a cause or consequence (or both) of retained expression of homeologs. Our analyses of molecular evolution of *S.* new tetraploid transcripts was consistent with observations in many other taxa [13, 127] that duplicates evolve faster than singletons and that they evolve differently from one another.

2.4.1 An allopolyploid mechanism of WGD in *S.* new tetraploid

Whole genome duplication (WGD) can occur spontaneously within a species (autopolyploidization) or in association with hybridization between species (allopolyploidization). Analysis of a small number of autosomal genes previously suggested that for most polyploid *Xenopus* species allopolyploidy is the probable mechanism of WGD [17] (reviewed in [100]). It has been proposed that *S.* new tetraploid speciated

from its sister species *S. epitropicalis* following an allotetraploid hybridization event between two species, and that one of the diploid ancestors of this pair of tetraploid sister species was also a recent ancestor of *S. tropicalis* [101, 102].

If tetraploidization of *S.* new tetraploid occurred by autopolyploidy, we expected a monophyletic relationship between *S.* new tetraploid homeologs when analyzed with orthologs of other species such as the diploid *S. tropicalis*. Conversely, following allopolyploidization involving a recent diploid ancestor of *S. tropicalis*, we expected *S.* new tetraploid paralogs to be paraphyletic with respect to an ortholog from *S. tropicalis* (Figure 2.2). To evaluate these evolutionary scenarios, we identified 6,516 *S. tropicalis* genes from a 19,921-sequence cDNA database for which we had high-confidence *S.* new tetraploid homologous transcripts and identified a *X. laevis* homolog (either a singleton or duplicate *X. laevis* gene). Of these, we found 4,494 genes (69%) supported paraphyly of *S.* new tetraploid transcripts, indicating putative α- and β-homeologs. This proportion is comparable to, albeit higher than, the observed rate of duplicate gene retention in *X. laevis* [108], as expected because genome duplication occurred earlier in *X. laevis* [101, 102]. The high proportion of homeologs with a paraphyletic relationship with respect to *S. tropicalis* supports an allopolyploid as opposed to an autopolyploid mechanism of gene duplication in *S.* new tetraploid.

An allopolyploid mechanism of WGD in *Silurana* implies a prior hybridization event between two diploid species. While many species are known to form hybrids with other species, the establishment of new polyploid species is much less common, presumably because polyploid offspring may be inviable or have decreased fertility or survivorship compared to their diploid progenitors [128]. However, hybrid individuals may benefit from overdominance/hybrid vigor, and potential examples of this are

evidenced by some polyploid *Xenopus* species being resistant to parasites that infect closely related species with lower ploidy levels [129, 130]. Hybrid vigor, therefore, may bolster the the ability to compete in a particular environment despite decreased fitness [131] (reviewed in [16]). Allopolyploids also benefit from masking the effects of deleterious alleles due to increased gene copies, however this also may slow genomic purging of deleterious alleles. Similarly, beneficial mutations can affect a greater number of gene copies allowing for increased adaptability [15, 132]. In general, speciation by polyploidy occurred more frequently in *Xenopus* than in other amphibians (reviewed in [58, 133]), opening the possibility that the benefits or tolerance of WGD may be atypical.

2.4.2 Sexually dimorphic expression in *S.* new tetraploid

In dioecious species, although the male and female sexes possess nearly identical autosomal genes, there are numerous physiological, behavioral and morphological differences between them. In *Xenopus*, for example, males have a smaller body size, a much larger vocal organ (the larynx), smaller cloacal lobes, and develop nuptial pads on the forearms [134]. The development of sexually dimorphic traits is driven by (i) differential expression between males and females [135] and (ii) differences in gene content between males and females due to genes on sex-specific regions of the sex chromosomes. The latter mechanism is probably of relatively small consequence in *S.* new tetraploid based on the large pseudoautosomal region on the W chromosome of the closely related species *S. tropicalis* [103] and lack of cytological differences between *S.* new tetraploid sex chromosomes [136]. In contrast, sexually dimorphic expression of key genes is well documented in *Xenopus*, including pronounced differences in

laryngeal myosin expression in the larynx and steroid receptors in the brain [134, 137, 138].

A consequence of the necessity of orchestrating sexual differentiation of two sex phenotypes from two almost identical genomes is that alleles may develop sexually antagonistic function [139]. Gene duplication potentially may help resolve the intragenomic conflict by catalyzing expression divergence of paralogs with sex-biased expression and function [124, 140], making the exploration of *S.* new tetraploid transcriptome evolution of particular interest.

To explore sex-biased expression in *S.* new tetraploid, we fitted a negative binomial GLMM to model sex- and tissue-specific expression variation. When data from both sexes are pooled, expression divergence between tissue types accounts for 70.9% and 86.3% of the variation in duplicates and singletons respectively. Thus, expression divergence between tissue types was not more substantial in duplicates compared to singletons. However, expression divergence between tissue types and sexes accounts for 73.3% and 86.6% of the variation in expression across duplicates and singletons respectively. Thus, the addition of sex as a factor in our analysis explained slightly more of the expression variation in duplicates (73.3% with sex versus 70.9% without) compared to singletons (86.6% with sex versus 86.3% without). Most strikingly, expression divergence between the sexes within a tissue type was on the same order of magnitude as that among tissue types within one sex.

We observe similar levels of sex-biased gene expression in both singleton and duplicate genes (14% and 11%, respectively), a result authenticated by similar findings in other studies [124, 141]. However in both singleton and duplicate genes, female-biased or female-specific genes comprised a higher proportion than male-biased genes.

In particular, we find that female-biased genes localized in brain tissue account for 70% of all female-biased duplicate genes and 55% of all female-biased singleton genes.

Evidence for female-biased gene expression abundance following WGD is sparse, with studies typically supporting a preference for male-biased gene evolution [123, 124, 142, 143], although see [126]. In contrast to this, because the sex chromosomes of *S.* new tetraploid probably have a very large pseudoautosomal region based on analysis of the closely related diploid *S. tropicalis* [103], few genes would be expected to have a difference in allele number between males and females.

Sex-biased genes often have expression limited mostly to a particular tissue type or developmental stage as opposed to having a broad and even pattern of expression across tissue types or developmental stages [124, 143, 144]. Our analyses of expression across sexes and tissue types in both singleton and duplicate genes support this observation in *S.* new tetraploid (Figure 2.5; Figure 2.6). Specifically, we found that female-biased expression drives overall higher expression levels in brain tissue compared to heart and liver. Expression in males was much more even, with most male-biased genes found in liver tissue. However, in males we similarly observe the highest expression levels in liver, and this corresponds with the highest proportion of male-biased gene expression. Singleton genes exhibited a similar pattern of expression: brain tissue expression was higher than heart and liver tissue expression overall, and most female-biased expression patterns were in brain and most male-biased genes were in liver.

It is a frequent observation that sex-biased gene expression is negatively correlated with expression breadth [143, 145, 146, 147], and it is suggested that this is a factor that affects evolutionary rates. Here we report a comparable result. We

consistently observed a decrease in expression breadth among female-biased genes in both singletons and duplicates. Other studies have found that male-biased, not female-biased, expression breadth is lowest [145, 148]; however, our result is unsurprising given the abundance of female-biased gene expression we observed in *S.* new tetraploid compared to other species. The limited breadth of expression of female-biased genes, coupled with the observed abundance of female-biased genes, could allow for increased evolutionary rates and could suggest a sex-specific reproductive effect in the brain [149].

2.4.3 Molecular evolution following WGD in *S.* new tetraploid

Following WGD, copies of duplicate genes may be lost (pseudogenization), or retained and retained duplicates may undergo changes in expression or function associated with new (neofunctionalization) or compromised activity (subfunctionalization), or other evolutionary modifications (reviewed in [150]). In theory, redundancy of gene copies is associated with relaxed purifying selection on homeologs [13] and an increased ratio of nonsynonymous to synonymous substitutions per site. Subfunctionalization and neofunctionalization could also be associated with a different rate of evolution between each homeolog.

Based on models of molecular evolution in which natural selection is allowed to vary among branches of a phylogeny, we found evidence for a reduction in the level of purifying selection following WGD. We did not recover support for a bias in the rate ratio of non-synonymous to synonymous substitutions per site among either α- or β-homeologs in *S.* new tetraploid, where the α homeolog is the one more closely related to *S. tropicalis*.

Sémon and Wolfe [97] previously described convergent outcomes of evolution following WGD in *X. laevis* and the zebrafish, *Danio rerio*, in that slowly-evolving genes were more likely to become subfunctionalized and therefore retained in duplicate. We performed a similar analysis with the 6,516 genes we identified between *S. tropicalis* and *S.* new tetraploid. Of 4,494 genes which were retained in duplicate in *S.* new tetraploid, 2,701 homeologous genes (60%) were also retained in *X. laevis*. However, of 2,022 identified singleton *S.* new tetraploid genes, 1,200 were retained duplicate genes in *X. laevis*. In contrast to the results of Sémon and Wolfe [97], we did not find evidence for a significant association between duplicate genes in *X. laevis* and *S.* new tetraploid (Fishers exact test: $p > 0.5$). One explanation for these discordant results is that the divergence time between *S.* new tetraploid and *X. laevis* is much more recent than that between *X. laevis* and *D. rerio*.

Genes with sex biased expression often evolve more quickly than those without sex biased expression, and this is typically evident in male-biased genes ([87, 124, 145]; reviewed in [135]). However, we did not find evidence for accelerated evolution in male-biased genes; instead, compared to unbiased genes, we found elevated d_N/d_S values in female brain tissue (Mann-Whitney U test: $p < 0.05$). We did, however, find evidence for more rapid evolution among antagonistically expression duplicate gene homeologs. Duplicate genes in which homeolog pairs were differentially sex-biased exhibited mean d_N/d_S values up to twice that of the unbiased average. These genes provide strong evidence that duplicate gene evolution alleviates sexually antagonistic selection and allows for divergence of sex-biased gene expression [124, 125, 135, 140]

2.4.4 Conclusions

Polyploidization in the *S.* new tetraploid species offers an opportunity to explore how expression and molecular evolution of duplicate genes differs from singletons, and to examine how and whether sex biases in expression are more prevalent in duplicates as compared to singletons. Analyses presented here support an allopolyploid, as opposed to an autopolyploid, origin for *S.* new tetraploid. We further recovered support for unique expression patterns in duplicates and singletons, wherein the former class of genes tended to have more sex-biased expression and higher expression intensity than the latter class. Perhaps most interestingly, we found several attributes of sex-biased expression, including strongly female-biased expression patterns in the brain and antagonistically sex-biased homeologs among duplicate genes. The abundance of female-biased gene expression in the brain is correlated with decreased breadth of expression and increased evolutionary rate among female-biased genes, a result that is consistent with observations from other studies, suggesting a female-specific reproductive function for these genes in the brain. We propose that WGD in *S.* new tetraploid allows for the resolution of sexually antagonistic selection by reducing selective constraints, resulting in the co-option of duplicate gene homeologs for sex-specific roles, particularly in the brain of females, and increasing sex-biased gene expression.

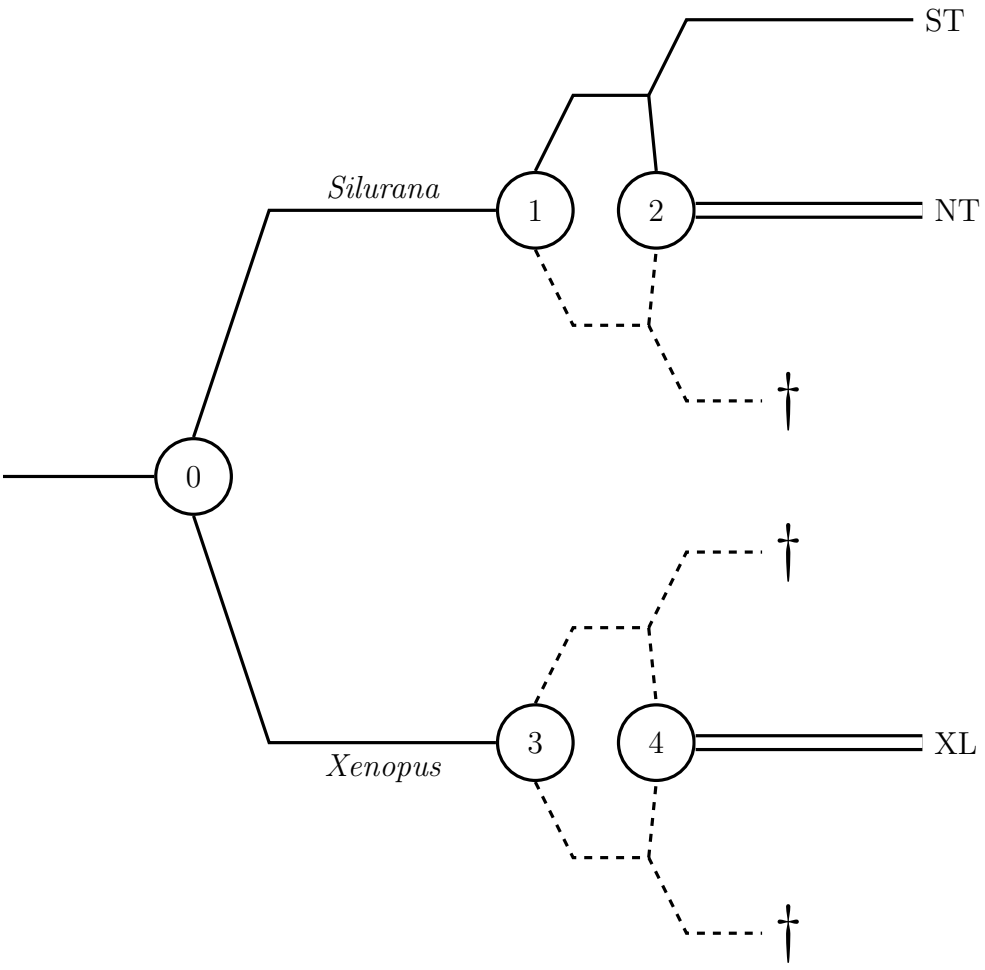


Figure 2.1: **Phylogenetic relationships of species and paralogues in this study.** Phylogenetic relationships are depicted among species for a diploid, *S. tropicalis* (ST), and two tetraploids, *X. laevis* (XL) and *S.* new tetraploid (NT). Speciation by allopolyploidization occurred independently in *Silurana* and *Xenopus*. Tetraploid lineages are indicated with a double line, and extinct lineages are indicated by a dagger and dashed lines. Numbered nodes indicate: (0) divergence of the genera *Silurana* and *Xenopus*, (1) divergence of the diploid ancestors of *Silurana*, (2) allotetraploidization in *Silurana*, (3) divergence of the diploid ancestors of *Xenopus*, (4) allotetraploidization in *Xenopus*.

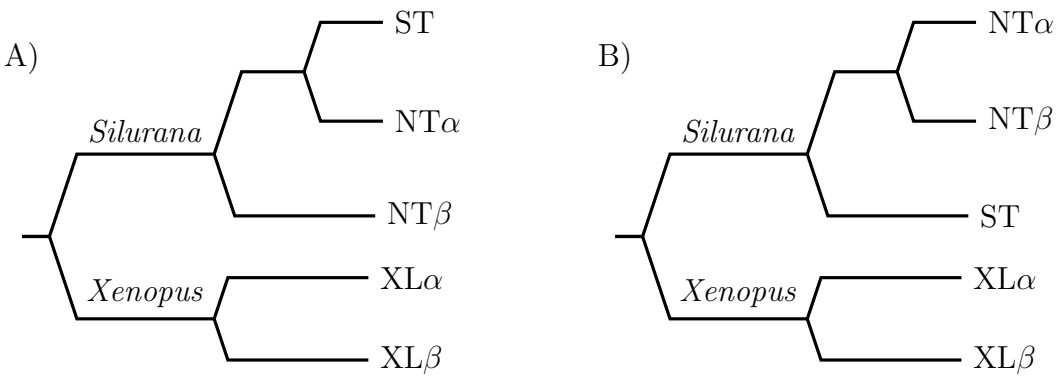


Figure 2.2: **Expectations from Allo- and Auto-polyploidy.** Species in this study are indicated on the tree: *Silurana tropicalis* (ST), *Xenopus laevis* (XT) and *S.* new tetraploid (NT). Homeolog pairs are indicated where appropriate (α and β) *a*) Allopolyploidy from species hybridization results in paraphyly of duplicate gene homeologs (NTα and NT β), where one homeolog is more closely related to ST than the other homeolog. *b*) Autopolyploidization results in monophyly of duplicate gene paralogues (NTα and NTβ). In both cases, pseudogenization of paralogous sequences over time diminishes the ability to resolve paralogy.

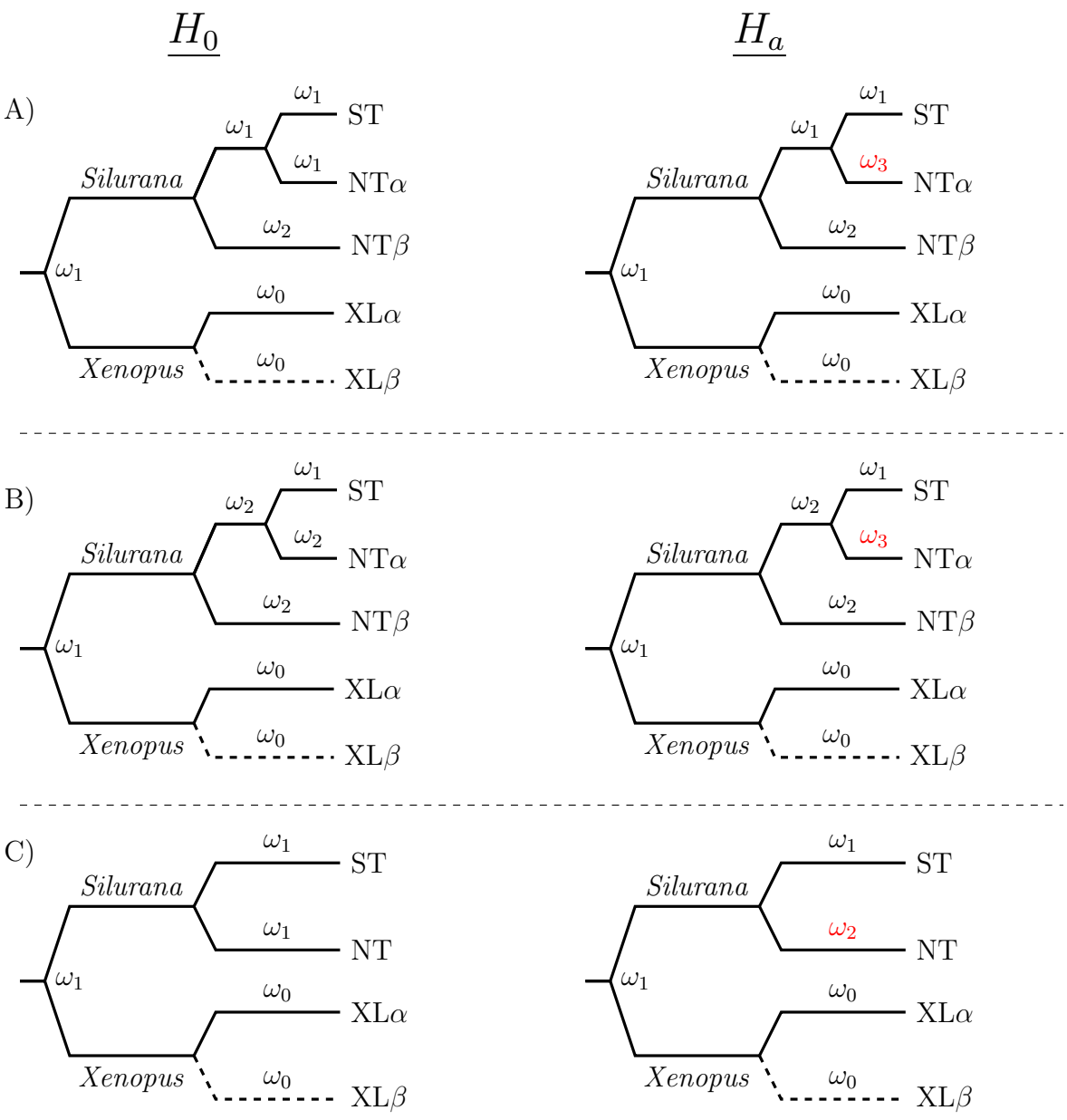


Figure 2.3: **Branch models for testing molecular evolution with Codeml.** For each comparison the alternative hypothesis model (H_a) is compared to the null hypothesis model (H_0) through an LnL ratio for each gene. Rate-ratios for each branch are specified with ω. Rate ratio changes between H_0 and H_a are highlighted in red. Dashed lines indicate *X. laevis* lineages which may or may not be present for a given gene due to pseudogenization. *a*) Test for relaxed purifying selection in duplicate genes of *S.* new tetraploid. *b*) Test for rate-ratio variation between *S.* new tetraploid α and β homeologs in duplicate genes. *c*) Test for rate-ratio variation between *S. tropicalis* and *S.* new tetraploid singleton genes.

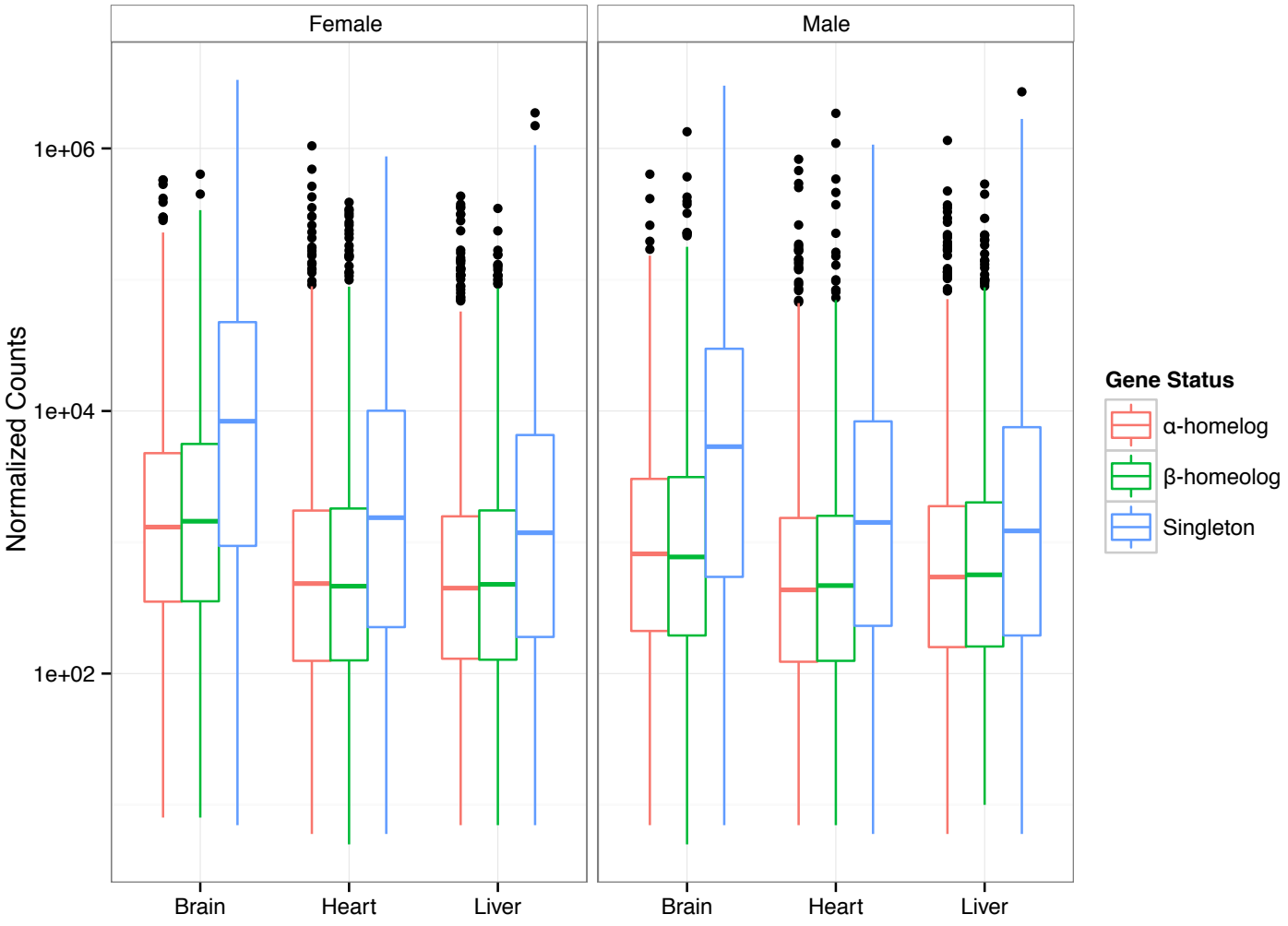
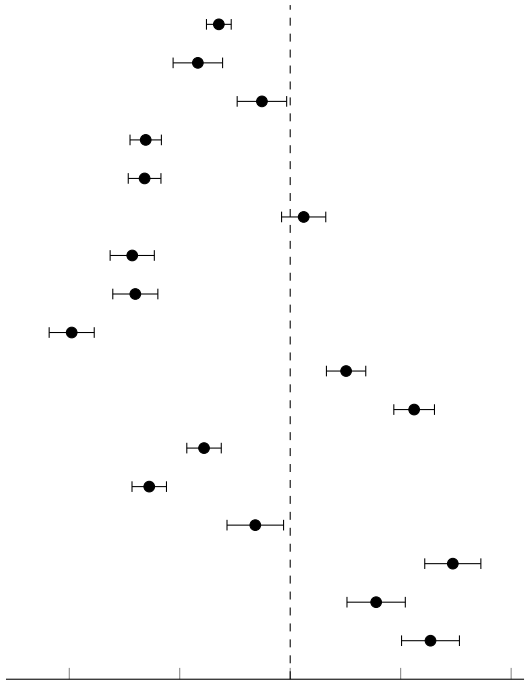


Figure 2.4: **Summary of expression counts of singleton and duplicate genes across both sexes and all tissues.** Displayed values are normalized count values adjusted by a size factor for each library. Singleton genes (blue) and the α (red) and β (green) homeologs of duplicate genes are shown separately. Outlier values are indicated by black points. Median normalized count values are shown with upper and lower quartiles for each library.

		b	$SE_{\bar{x}}$	z
Sex		-0.710	0.027	-26.4
Homeolog		0.290	0.028	-10.5
Heart		-0.585	0.034	-17.0
Liver		-1.353	0.035	-38.8
Heart:Homeolog		-0.776	0.046	-17.0
Liver:Homeolog		-0.287	0.046	-6.2
Heart:Sex		0.090	0.045	2.0
Liver:Sex		0.946	0.046	20.4
Homeolog:Sex		-0.248	0.037	-6.6
Heart:Homeolog:Sex		0.894	0.063	14.1
Liver:Homeolog:Sex		0.246	0.064	3.8
	−1.5 −1 −0.5 0 0.5 1			

Figure 2.5: **Regression coefficients for the fit of a negative binomial GLMM modeling homeolog divergence in *S.* new tetraploid.** Coefficients (b) with standard errors ($SE_{\bar{x}}$) indicate magnitude and direction of the effect. Coefficients with appropriate confidence intervals are plotted on the X-axis. The dashed line at 0 indicates the Y-axis. 95% confidence intervals are given as $1.96 \cdot SE_{\bar{x}}$. The significance and direction of each effect is given by the z-statistic (z). Females, brain tissue and α-homeologs are given as the baseline for sex, tissue type and homeolog comparisons, respectively. Interaction terms for fixed effects are denoted by a colon.



	β	$SE_{\bar{x}}$	z
Sex	-0.323	0.029	-11.3
Alpha	-0.418	0.057	-7.3
Beta	-0.128	0.057	-2.2
Heart	-0.654	0.036	-17.9
Liver	-0.659	0.038	-17.5
Heart:Alpha	0.061	0.051	1.2
Heart:Beta	-0.715	0.051	-14.1
Liver:Alpha	-0.701	0.052	-13.4
Liver:Beta	-0.989	0.052	-18.9
Heart:Sex	0.253	0.045	5.6
Liver:Sex	0.561	0.047	11.9
Alpha:Sex	-0.390	0.040	-9.8
Beta:Sex	-0.638	0.040	-16.0
Heart:Alpha:Sex	-0.158	0.065	-2.4
Heart:Beta:Sex	0.736	0.065	11.4
Liver:Alpha:Sex	0.389	0.067	5.8
Liver:Beta:Sex	0.635	0.067	9.5

Figure 2.6: **Regression coefficients for the fit of a negative binomial GLMM modeling expression divergence between singleton and duplicate genes in *S.* new tetraploid.** Coefficients (b) with standard errors ($SE_{\bar{x}}$) indicate magnitude and direction of the effect. Coefficients with confidence intervals are plotted on the X-axis. The Y-axis is given by the dashed line at 0. 95% confidence intervals are given as $1.96 * SE_{\bar{x}}$. The significance and direction of each effect is given by the z-statistic (z). Females, brain tissue and singletons are given as the baseline for sex, tissue type and gene status, respectively. Interaction terms for fixed effects are denoted by a colon.

2.5 Supplemental Tables and Figures

Tissue	**Individual**	**Number of Transcripts**	**Number of Reads**
Brain	F01	446,588	28,779,998
	F02	451,747	27,585,603
	F03	426,838	25,763,088
	M95	318,218	24,094,162
	M96	408,225	25,438,659
	M97	281,069	20,497,388
Heart	F01	154,231	25,755,812
	F02	193,613	24,064,643
	F03	160,460	23,651,201
	M95	154,058	22,440,134
	M96	167,794	21,580,463
	M97	141,571	19,991,308
Liver	F01	160,594	21,843,062
	F02	151,175	21,312,083
	F03	195,990	24,001,863
	M95	204,814	24,108,551
	M96	82,760	15,870,406
	M97	158,811	18,087,165
Total		**4,258,556**	**414,865,589**

Table S1: **Number of reads and transcripts per library.** Each read library was separately assembled into a transcript library. Read counts are for processed reads and do not contain poor-quality reads removed during processing. Transcriptomes were assembled with the Velvet-Oases algorithm. The mean read library size is 23,048,088 reads. The mean transcript library size is 236,586 transcript sequences.

Tissue	Individual	Number of Reads	Percent Alignment
Brain	F01	28,779,998	77.15
	F02	27,585,603	76.30
	F03	25,763,088	74.59
	M95	24,094,162	71.36
	M96	25,438,659	73.63
	M97	20,497,388	69.34
Heart	F01	25,755,812	69.16
	F02	24,064,643	76.47
	F03	23,651,201	76.93
	M95	22,440,134	73.97
	M96	21,580,463	71.88
	M97	19,991,308	72.26
Liver	F01	21,843,062	66.74
	F02	21,312,083	71.11
	F03	24,001,863	71.18
	M95	24,108,551	75.26
	M96	15,870,406	65.14
	M97	18,087,165	72.10
Mean			**72.5**

Table S2: **Percentage of reads aligned by Bowtie2 for each transcript library.** Percent-alignment indicates the percentage of reads that were successfully aligned from each library to the corresponding transcriptome. Read counts are for processed reads and do not include poor quality reads removed during processing. Reads were first indexed and then mapped with Bowtie2. The mean percent-alignment across all libraries was 72.5%.

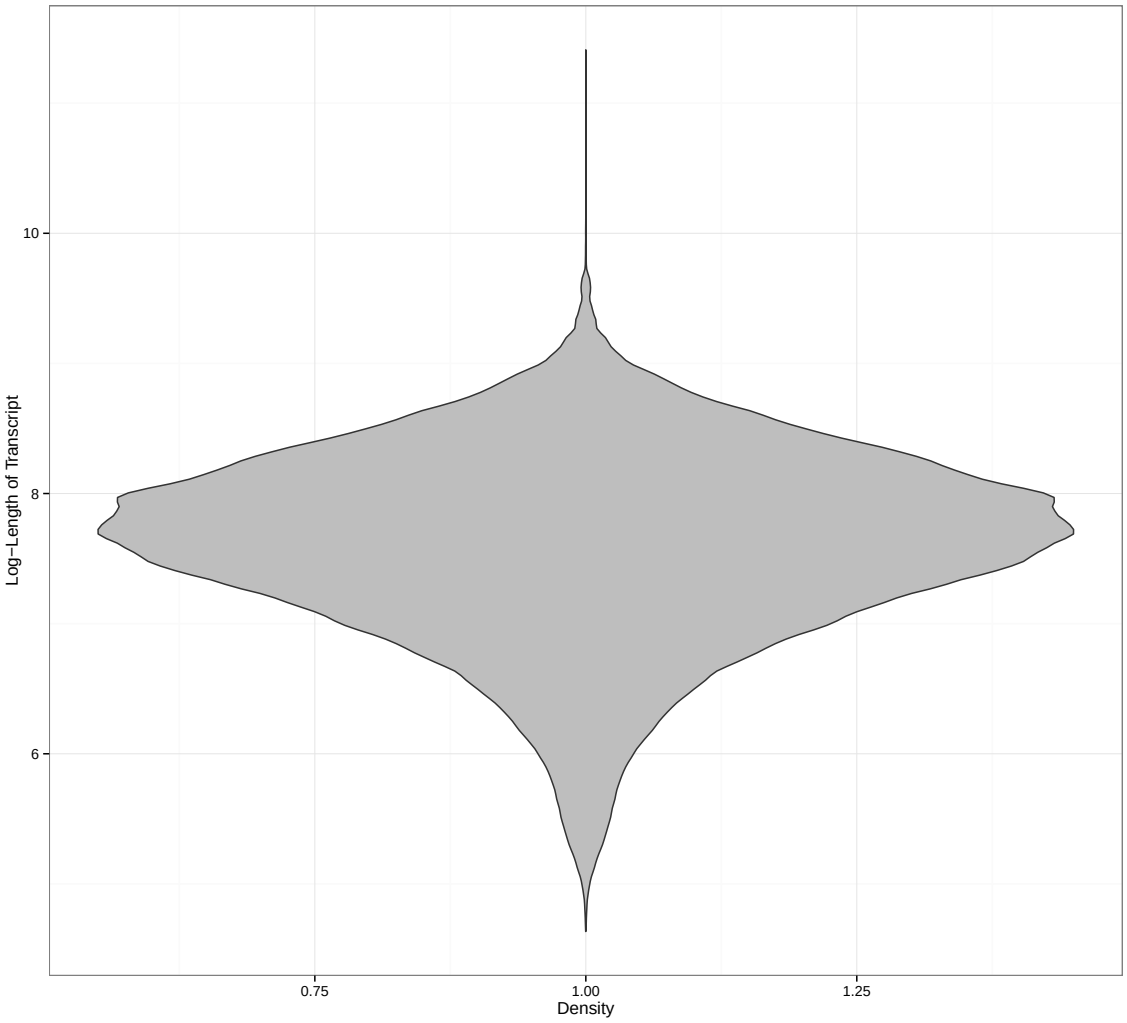


Figure S1: **Violin plot of log-transformed transcript sequence lengths.** 4,258,556 transcripts from 18 transcript libraries are included. The mean transcript length is 2,537 bp, with a median of 2,153 bp. Transcript range from 101 bp to 92,002 bp in length.

	df	lnL	deviance	Chisq	Chi df	Pr(>Chisq)
baseline	2	-3.298×10^5	6.596×10^5			
tissue	4	-3.286×10^5	6.572×10^5	2.442×10^3	2	0.000
homeolog	5	-3.286×10^5	6.572×10^5	1.037×10^1	1	1.278×10^{-3}
sex	6	-3.281×10^5	6.563×10^5	8.801×10^2	1	2.089×10^{-193}
tissue:sex	8	-3.281×10^5	6.562×10^5	9.811×10^1	2	4.967×10^{-22}
tissue:homeolog	10	-3.276×10^5	6.552×10^5	9.836×10^2	2	2.578×10^{-214}
homeolog:sex	11	-3.276×10^5	6.552×10^5	2.610	1	1.062×10^{-1}
tissue:homeolog:sex	13	-3.275×10^5	6.550×10^5	1.836×10^2	2	1.359×10^{-40}

Table S3: **Analysis of variance for homeolog expression model.** Displayed are log-likelihood (lnL), deviance and χ^2 (Chisq) values for each fixed effect and interaction term in the model. Interaction terms are denoted by a colon. 'baseline' refers to a model with only a random effect (no fixed effects or interaction terms). Here, 'homeolog' may be α homeolog or β homeolog. Each fixed effect and interaction term is significant in this model based on a χ^2 test, except for the 'homeolog:sex' interaction term.

	df	lnL	deviance	Chisq	Chi df	Pr(>Chisq)
baseline	2	-5.047×10^5	1.009×10^6			
tissue	4	-5.035×10^5	1.007×10^6	2.540×10^3	2	0.000
status	6	-5.034×10^5	1.007×10^6	1.681×10^2	2	3.221×10^{-37}
sex	7	-5.030×10^5	1.006×10^6	7.802×10^2	1	1.095×10^{-171}
tissue:status	11	-5.028×10^5	1.006×10^6	3.895×10^2	4	5.134×10^{-83}
tissue:sex	13	-5.022×10^5	1.004×10^6	1.109×10^3	2	1.853×10^{-241}
status:sex	15	-5.022×10^5	1.004×10^6	1.592×10^2	2	2.749×10^{-35}
tissue:status:sex	19	-5.020×10^5	1.004×10^6	2.700×10^2	4	3.148×10^{-57}

Table S4: **Analysis of variance for duplicate and singleton gene expression model.** Displayed are log-likelihood (lnL), deviance and χ^2 (Chisq) values for each fixed effect and interaction term in the model. Interaction terms are denoted by a colon. 'baseline' refers to a model with only a random effect (no fixed effects or interaction terms). Here, 'status' refers to the gene status and may be singleton, α homeolog or β homeolog. Each fixed effect and interaction term is significant in this model based on a χ^2 test.

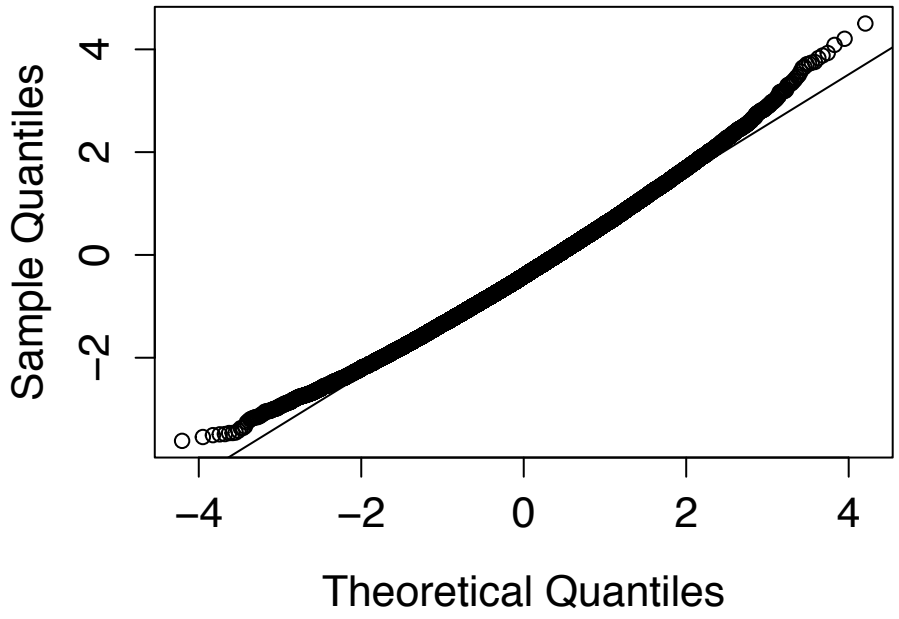


Figure S2: **Q-Q plot for homeolog gene expression model.** Linearity of the theoretical and sample quantiles is indicative of normality of the residuals, an expectation from our model.

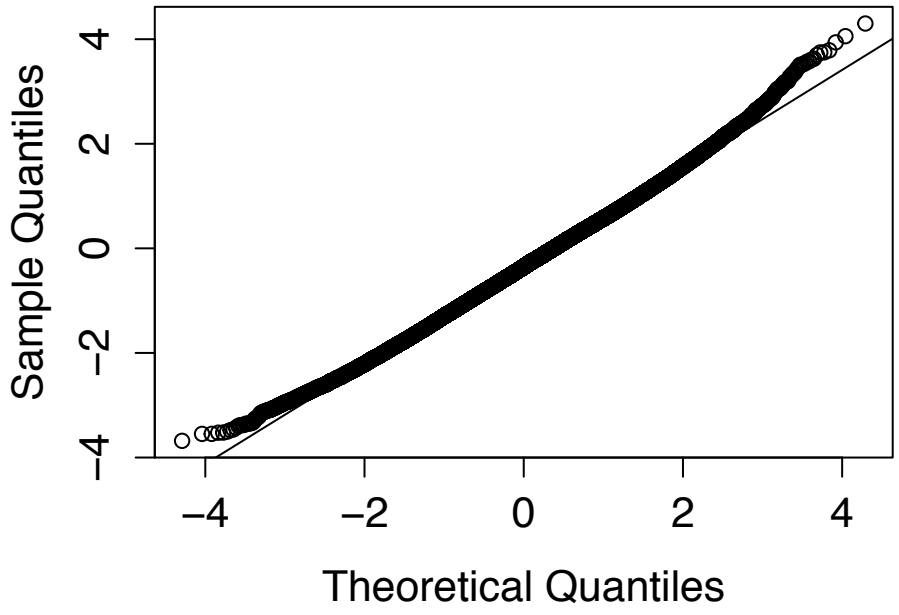


Figure S3: **Q-Q plot for duplicate and singleton gene expression model.** Linearity of the theoretical and sample quantiles is indicative of normality of the residuals, an expectation from our model.

Chapter 3

Concluding Remarks

3.1 Thesis Contributions

This thesis project represents one of only a few studies of evolution following whole-genome duplication (WGD) in a tetrapod animal. The use of a previously identified extant diploid sister taxon (*Silurana tropicalis*) to compare molecular evolution of duplicate and singleton genes in a tetraploid species (*S.* new tetraploid) is particularly unique. Through the use of RNAseq and a transcriptome-based design, we were able to identify homeologous genes in *S.* new tetraploid and orthologous genes in *X. laevis* and *S. tropicalis*, investigate molecular evolution of homeologs and singleton genes in *S.* new tetraploid with comparison to *S. tropicalis*, and model expression divergence between singleton and duplicate genes across multiple tissue types (brain, heart and liver) in *S.* new tetraploid. We then investigated evidence for sex-biased gene expression patterns in *S.* new tetraploid and explored factors contributing to the evolution of sex-biased gene expression and the resolution of sexual antagonism. The study offers a novel method through which to resolve relatedness of homeologous genes,

illustrates how duplicate gene expression divergence can be modeled with generalized linear mixed-effects models, and contributes to the growing body of literature on the topics of duplicate gene evolution, sex-biased expression divergence, and their interaction.

Most studies of sex-biased gene expression evolution are from the fruit fly *Drosophila melanogaster* [82, 124, 145, 151] or birds [85, 126, 152] (reviewed in [135]). This thesis therefore contributes to the relative dearth of studies investigating sex-biased expression in *Xenopus* or other amphibians. Additionally, we found that female-biased gene expression was more common among duplicate genes than male-biased expression, a surprising finding given the literature on rapid evolution and high expression of male-biased genes in *Drosophila* [124, 151]. We also described a disproportionate amount of female-biased gene expression occurring in the brain suggesting differential gene regulation and selective pressures in the brain of females and males, and implying a reproductive function for female-biased genes in the brain. While an explanation for this observation is beyond the scope of this thesis, we note that the *Xenopus* genome contains substantial differences, particularly the ZW sex-determination and the presence of sex-chromosome homomorphy with a large pseudo-autosomal region [103].

A transcriptome-based experimental design lends itself to a comprehensive study, with the ability to simultaneously analyze transcript expression, molecular evolution and function. Additionally, since the transcriptome, unlike much of the genome, contains sex-specific and tissue-specific features, we are able to model multiple factors and their interactions in a single design. While this is not new, investigation of sex-

and tissue-biased transcript expression in duplicate genes is a burgeoning field of research, and we are able to unite findings from sex-biased gene expression in *S.* new tetraploid with expectations from molecular evolution of duplicate genes following allopolyploidization. Though not included in this thesis, we will also determine the genomic location of duplicate genes and sex-biased genes to estimate if genomic distribution of these genes is nonrandom in *S.* new tetraploid, as reported in other species [124, 143, 152], and if so how sex-biased genes cluster in the genome. Additionally, studies of sex-biased expression and gene evolution in *S.* new tetraploid would greatly benefit from further sequencing of sex organs, or over multiple developmental stages. Overall, this study contributes to research on sexual dimorphism, sexually antagonistic selection and sex-biased gene expression in allopolyploid African clawed frogs.

Bibliography

[1] Susumu Ohno et al. *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag., 1970.

[2] Calvin B Bridges. The bar "gene" a duplication. *Science*, 83(2148):210–211, 1936.

[3] JBS Haldane. The part played by recurrent mutation in evolution. *American Naturalist*, pages 5–19, 1933.

[4] SG Stephens. Possible significance of duplication in evolution. *Advances in Genetics*, 4:247–265, 1951.

[5] Susumu Ohno. *Sex chromosomes and sex-linked genes*, volume 1. Springer Science & Business Media, 1967.

[6] Susumu Ohno, Ulrich Wolf, and Niels B Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.

[7] Masatoshi Nei. Gene duplication and nucleotide substitution in evolution. 1969.

[8] W. H. Li. Rate of gene silencing at duplicate loci. A theoretical study and interpretation of data from tetraploid fishes. *Genetics*, 95(1):237–258, 1980.

[9] Allan Force, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[10] Michael Lynch and John S Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, 2003.

[11] Marianne K Hughes and Austin L Hughes. Evolution of duplicate genes in a tetraploid animal, xenopus laevis. *Molecular Biology and Evolution*, 10(6):1360–1369, 1993.

[12] Joseph H Nadeau and David Sankoff. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147(3):1259–1266, 1997.

[13] Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[14] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1984.

[15] Sarah P Otto and Jeannette Whitton. Polyploid incidence and evolution. *Annual Review of Genetics*, 34(1):401–437, 2000.

[16] Sarah P Otto. The evolutionary consequences of polyploidy. *Cell*, 131(3):452–462, 2007.

[17] Frédéric JJ Chain, Dora Ilieva, and Ben J Evans. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology*, 8(1):43, 2008.

[18] Tine Blomme, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology*, 7(5):R43, 2006.

[19] Jerel C Davis and Dmitri A Petrov. Do disparate mechanisms of duplication add similar genes to the genome? *Trends in Genetics*, 21(10):548–551, 2005.

[20] Jianzhi Zhang, Ya-ping Zhang, and Helene F Rosenberg. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics*, 30(4):411–415, 2002.

[21] Austin L Hughes. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B: Biological Sciences*, 256(1346):119–124, 1994.

[22] FJJ Chain and BJ Evans. Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genetics*, 2(4):e56, 2006.

[23] Zhenglong Gu, Dan Nicolae, Henry HS Lu, and Wen-Hsiung Li. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics*, 18(12):609–613, 2002.

[24] Guillaume Blanc and Kenneth H Wolfe. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *The Plant Cell Online*, 16(7):1679–1691, 2004.

[25] Kevin P Byrne and Kenneth H Wolfe. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3):1341–1350, 2007.

[26] Xun Gu, Zhongqi Zhang, and Wei Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences*, 102(3):707–712, 2005.

[27] Gerald M Rubin, Mark D Yandell, Jennifer R Wortman, George L Gabor, Catherine R Nelson, Iswar K Hariharan, Mark E Fortini, Peter W Li, Rolf Apweiler, Wolfgang Fleischmann, et al. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, 2000.

[28] Kenneth H Wolfe and Denis C Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.

[29] Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624, 2004.

[30] Arabidopsis Genome Initiative et al. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796, 2000.

[31] Wen-Hsiung Li, Zhenglong Gu, Haidong Wang, and Anton Nekrutenko. Evolutionary analyses of the human genome. *Nature*, 409(6822):847–849, 2001.

[32] Jürg Spring. Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS Letters*, 400(1):2–8, 1997.

[33] Robert Friedman and Austin L Hughes. Pattern and timing of gene duplication in animal genomes. *Genome Research*, 11(11):1842–1847, 2001.

[34] Xun Gu, Yufeng Wang, and Jianying Gu. Age distribution of human gene families shows significant roles of both large-and small-scale duplications in vertebrate evolution. *Nature Genetics*, 31(2):205–209, 2002.

[35] Robert Friedman and Austin L Hughes. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Molecular Biology and Evolution*, 20(1):154–161, 2003.

[36] Paramvir Dehal and Jeffrey L Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10):e314, 2005.

[37] Kenneth H Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2(5):333–341, 2001.

[38] Angel Amores, Allan Force, Yi-Lin Yan, Lucille Joly, Chris Amemiya, Andreas Fritz, Robert K Ho, James Langeland, Victoria Prince, Yan-Ling Wang, et al. Zebrafish *HOX* clusters and vertebrate genome evolution. *Science*, 282(5394):1711–1714, 1998.

[39] Masanori Kasahara, Jun Nakaya, Yoko Satta, and Naoyuki Takahata. Chromosomal duplication and the emergence of the adaptive immune system. *Trends in Genetics*, 13(3):90–92, 1997.

[40] Axel Meyer and Yves Van de Peer. From 2R to 3R: Evidence for a fish-specific genome duplication (*FSGD*). *Bioessays*, 27(9):937–945, 2005.

[41] Masanori Kasahara. The 2r hypothesis: an update. *Current Opinion in Immunology*, 19(5):547–552, 2007.

[42] Cristian Cañestro. Two rounds of whole-genome duplication: evidence and impact on the evolution of vertebrate innovations. In *Polyploidy and genome evolution*, pages 309–339. Springer, 2012.

[43] Jeramiah J Smith, Shigehiro Kuraku, Carson Holt, Tatjana Sauka-Spengler, Ning Jiang, Michael S Campbell, Mark D Yandell, Tereza Manousaki, Axel Meyer, Ona E Bloom, et al. Sequencing of the sea lamprey (petromyzon marinus) genome provides insights into vertebrate evolution. *Nature Genetics*, 45(4):415–421, 2013.

[44] Cornel Popovici, Magalie Leveugle, Daniel Birnbaum, and François Coulier. Homeobox gene clusters and the human paralogy map. *FEBS Letters*, 491(3):237–242, 2001.

[45] Nicholas H Putnam, Thomas Butts, David EK Ferrier, Rebecca F Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, Eiichi Shoguchi, Astrid Terry, Jr-Kai Yu, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

[46] Federico G Hoffmann, Juan C Opazo, and Jay F Storz. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. *Molecular Biology and Evolution*, 29(1):303–312, 2012.

[47] Jay F Storz, Juan C Opazo, and Federico G Hoffmann. Gene duplication,

genome duplication, and the functional diversification of vertebrate globins. *Molecular Phylogenetics and Evolution*, 66(2):469–478, 2013.

[48] Lars-Gustav Lundin, Dan Larhammar, and Finn Hallböök. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. In *Genome Evolution*, pages 53–63. Springer, 2003.

[49] HJ Muller. Why polyploidy is rarer in animals than in plants. *American Naturalist*, pages 346–353, 1925.

[50] H Allen Orr. "why polyploidy is rarer in animals than in plant" revisited. *American Naturalist*, pages 759–770, 1990.

[51] BK Mable. why polyploidy is rarer in animals than in plants: myths and mechanisms. *Biological Journal of the Linnean Society*, 82(4):453–466, 2004.

[52] M Westergaard. The mechanism of sex determination in dioecious flowering plants. *Advances in Genetics*, 9:217–281, 1958.

[53] James P Bogart. Evolutionary implications of polyploidy in amphibians and reptiles. In *Polyploidy*, pages 341–378. Springer, 1980.

[54] Michael Schmid and Claus Steinlein. Sex chromosomes, sex-linked genes, and sex determination in the vertebrate class amphibia. *Genes and Mechanisms in Vertebrate Sex Determination*, pages 143–176, 2001.

[55] John E Bowers, Brad A Chapman, Junkang Rong, and Andrew H Paterson. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–438, 2003.

[56] Yuannian Jiao, Norman J Wickett, Saravanaraj Ayyampalayam, André S Chanderbali, Lena Landherr, Paula E Ralph, Lynn P Tomsho, Yi Hu, Haiying Liang, Pamela S Soltis, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100, 2011.

[57] Jane Masterson. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, 264(5157):421–423, 1994.

[58] Ben J Evans, R Alexander Pyron, and John J Wiens. Polyploidization and sex chromosome evolution in amphibians. In *Polyploidy and Genome Evolution*, pages 385–410. Springer, 2012.

[59] T RYAN Gregory and BARBARA K Mable. Polyploidy in animals. *The Evolution of the Genome*, 171:427–517, 2005.

[60] François Anthony, Benoit Bertrand, O Quiros, A Wilches, Philippe Lashermes, Julien Berthaud, and A Charrier. Genetic diversity of wild coffee (coffea arabica l.) using molecular markers. *Euphytica*, 118(1):53–65, 2001.

[61] Jonathan F Wendel and Richard C Cronn. Polyploidy and the evolutionary history of cotton. *Advances in Agronomy*, 78:139–186, 2003.

[62] Jorge Dubcovsky and Jan Dvorak. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, 316(5833):1862–1866, 2007.

[63] AR Leitch and IJ Leitch. Genomic plasticity and the diversity of polyploid plants. *Science*, 320(5875):481–483, 2008.

[64] Moshe Feldman and Avraham A Levy. Genome evolution due to allopolyploidization in wheat. *Genetics*, 192(3):763–774, 2012.

[65] Jürg Spring. Genome duplication strikes back. *Nature Genetics*, 31(2):128–129, 2002.

[66] Sonja A Rasmussen, Lee-Yang C Wong, Quanhe Yang, Kristin M May, and JM Friedman. Population-based analyses of mortality in trisomy 13 and trisomy 18. *Pediatrics*, 111(4):777–784, 2003.

[67] Jeannie Visootsak and John M Graham Jr. Klinefelter syndrome and other sex chromosomal aneuploidies. *Orphanet J Rare Dis*, 1(42):1–5, 2006.

[68] Andrew Singleton and Katrina Gwinn-Hardy. Parkinson disease and dementia with lewy bodies: a difference in dose? *The Lancet*, 364(9440):1105–1107, 2004.

[69] Andrew B Singleton. Altered α-synuclein homeostasis causing parkinson's disease: the potential roles of dardarin. *Trends in Neurosciences*, 28(8):416–421, 2005.

[70] John Hardy. Amyloid double trouble. *Nature Genetics*, 38(1):11–12, 2006.

[71] James R Lupski, Roberto Montes de Oca-Luna, Susan Slaugenhaupt, Liu Pentao, Vito Guzzetta, Barbara J Trask, Odila Saucedo-Cardenas, David F Barker, James M Killian, Carlos A Garcia, et al. Dna duplication associated with charcot-marie-tooth disease type 1a. *Cell*, 66(2):219–232, 1991.

[72] P Raeymaekers, V Timmerman, E Nelis, P De Jonghe, JE Hoogenduk, F Baas, DF Barker, JJ Martin, M De Visser, PA Bolhuis, et al. Duplication in chromosome 17p11. 2 in charcot-marie-tooth neuropathy type 1a (cmt 1a). *Neuromuscular Disorders*, 1(2):93–97, 1991.

[73] James R Lupski, Jeffrey G Reid, Claudia Gonzaga-Jauregui, David Rio Deiros, David CY Chen, Lynne Nazareth, Matthew Bainbridge, Huyen Dinh, Chyn Jing, David A Wheeler, et al. Whole-genome sequencing in a patient with charcot–marie–tooth neuropathy. *New England Journal of Medicine*, 362(13):1181–1191, 2010.

[74] Liu Pentao, Carol A Wise, A Craig Chinault, Pragna I Patel, and James R Lupski. Charcot–marie–tooth type 1a duplication appears to arise from recombination at repeat sequences flanking the 1.5 mb monomer unit. *Nature Genetics*, 2(4):292–300, 1992.

[75] LJ Valentijn, PA Bolhuis, I Zorn, JE Hoogendijk, N Van den Bosch, GW Hensels, VP Stanton, DE Housman, KH Fischbeck, DA Ross, et al. The peripheral myelin gene pmp–22/gas–3 is duplicated in charcot–marie–tooth disease type 1a. *Nature Genetics*, 1(3):166–170, 1992.

[76] V Timmerman, E Nelis, W Van Hul, BW Nieuwenhuijsen, KL Chen, S Wang, K Ben Othman, B Cullen, RJ Leach, CO Hanemann, et al. The peripheral myelin protein gene pmp–22 is contained within the charcot–marie–tooth disease type 1a duplication. *Nature Genetics*, 1(3):171–175, 1992.

[77] Ben J Evans, Darcy B Kelley, Richard C Tinsley, Don J Melnick, and David C Cannatella. A mitochondrial dna phylogeny of african clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution*, 33(1):197–213, 2004.

[78] Nicholas H Barton and Brian Charlesworth. Why sex and recombination? *Science*, 281(5385):1986–1990, 1998.

[79] William R Rice. Sex chromosomes and the evolution of sexual dimorphism. *Evolution*, pages 735–742, 1984.

[80] GS Van Doorn and M Kirkpatrick. Turnover of sex chromosomes induced by sexual conflict. *Nature*, 449(7164):909–912, 2007.

[81] David Sturgill, Yu Zhang, Michael Parisi, and Brian Oliver. Demasculinization of x chromosomes in the drosophila genus. *Nature*, 450(7167):238–241, 2007.

[82] Michael Parisi, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. Paucity of genes on the drosophila x chromosome showing male-biased expression. *Science*, 299(5607):697–700, 2003.

[83] Corina Schutt and Rolf Nothiger. Structure, function and evolution of sex-determining systems in dipteran insects. *Development*, 127(4):667–677, 2000.

[84] Ieuan A Hughes. Minireview: sex differentiation. *Endocrinology*, 142(8):3281–3287, 2001.

[85] Beatriz Vicoso and Brian Charlesworth. Evolution on the x chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8):645–653, 2006.

[86] Michael Lynch, Martin O'Hely, Bruce Walsh, and Allan Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001.

[87] Jianzhi Zhang. Evolution of dmy, a newly emergent male sex-determination gene of medaka fish. *Genetics*, 166(4):1887–1895, 2004.

[88] Shin Yoshimoto, Ema Okada, Hirohito Umemoto, Kei Tamura, Yoshinobu Uno, Chizuko Nishida-Umehara, Yoichi Matsuda, Nobuhiko Takamatsu, Tadayoshi Shiba, and Michihiko Ito. A w-linked dm-domain gene, dm-w, participates in primary ovary development in xenopus laevis. *Proceedings of the National Academy of Sciences*, 105(7):2469–2474, 2008.

[89] Arlin Stoltzfus. On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2):169–181, 1999.

[90] Shozo Yokoyama and Ruth Yokoyama. Molecular evolution of human visual pigment genes. *Molecular Biology and Evolution*, 6(2):186–197, 1989.

[91] Ana B Asenjo, Jeanne Rim, and Daniel D Oprian. Molecular determinants of human red/green color discrimination. *Neuron*, 12(5):1131–1138, 1994.

[92] Balazs Papp, Csaba Pal, and Laurence D Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003.

[93] Jonathan F Wendel. Genome evolution in polyploids. In *Plant Molecular Evolution*, pages 225–249. Springer, 2000.

[94] Toby J Gibson and Jürg Spring. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends in Genetics*, 14(2):46–49, 1998.

[95] Cathal Seoighe and Kenneth H Wolfe. Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology*, 2(5):548–554, 1999.

[96] Jean-Marc Aury, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M Porcel, Béatrice Ségurens, Vincent Daubin, Véronique Anthouard,

Nathalie Aiach, et al. Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. *Nature*, 444(7116):171–178, 2006.

[97] Marie Sémon and Kenneth H Wolfe. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in xenopus laevis. *Proceedings of the National Academy of Sciences*, 105(24):8333–8338, 2008.

[98] Uffe Hellsten, Richard M Harland, Michael J Gilchrist, David Hendrix, Jerzy Jurka, Vladimir Kapitonov, Ivan Ovcharenko, Nicholas H Putnam, Shengqiang Shu, Leila Taher, et al. The genome of the western clawed frog xenopus tropicalis. *Science*, 328(5978):633–636, 2010.

[99] J Brad Karpinka, Joshua D Fortriede, Kevin A Burns, Christina James-Zorn, Virgilio G Ponferrada, Jacqueline Lee, Kamran Karimi, Aaron M Zorn, and Peter D Vize. Xenbase, the xenopus model organism database; new virtualized system, data types and genomes. *Nucleic Acids Research*, 43(D1):D756–D763, 2015.

[100] Ben J Evans. Genome evolution and speciation genetics of clawed frogs (xenopus and silurana). *Frontiers in Bioscience*, 13:4687–4706, 2007.

[101] Ben J Evans, Darcy B Kelley, Don J Melnick, and David C Cannatella. Evolution of rag-1 in polyploid clawed frogs. *Molecular Biology and Evolution*, 22(5):1193–1207, 2005.

[102] Ben J Evans. Ancestry influences the fate of duplicated genes millions of years after polyploidization of clawed frogs (xenopus). *Genetics*, 176(2):1119–1130, 2007.

[103] Adam J Bewick, Frédéric JJ Chain, Lyle B Zimmerman, Abdul Sesay, Michael J Gilchrist, Nick DL Owens, Eva Seifertova, Vladimir Krylov, Jaroslav Macha, Tereza Tlapakova, et al. A large pseudoautosomal region on the sex chromosomes of the frog silurana tropicalis. *Genome Biology and Evolution*, 5(6):1087–1098, 2013.

[104] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, page btu170, 2014.

[105] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18(5):821–829, 2008.

[106] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[107] BJ Evans and Taejoon Kwon. Molecular polymorphism and divergence of duplicated genes in tetraploid african clawed frogs (xenopus). *Cytogenetic and Genome Research*, 2015.

[108] Frédéric JJ Chain, Jonathan Dushoff, and Ben J Evans. The odds of duplicate gene persistence after polyploidization. *BMC Genomics*, 12(1):599, 2011.

[109] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[110] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[111] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[112] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

[113] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[114] Ziheng Yang. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*, 15(5):568–573, 1998.

[115] Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[116] Vincent Ranwez, Sébastien Harispe, Frédéric Delsuc, and Emmanuel JP Douzery. Macse: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One*, 6(9):e22594, 2011.

[117] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, 2013.

[118] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[119] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 390–398. IEEE, 2000.

[120] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[121] Douglas Bates, Martin Maechler, and Ben Bolker. lme4: Linear mixed-effects models using s4 classes. 2012.

[122] Asher D Cutter and Samuel Ward. Sexual and temporal dynamics of molecular evolution in c. elegans development. *Molecular Biology and Evolution*, 22(1):178–188, 2005.

[123] Florian Gnad and John Parsch. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics*, 22(20):2577–2579, 2006.

[124] Minyoung J Wyman, Asher D Cutter, and Locke Rowe. Gene duplication in the evolution of sexual dimorphism. *Evolution*, 66(5):1556–1566, 2012.

[125] John Parsch and Hans Ellegren. The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews Genetics*, 14(2):83–87, 2013.

[126] Judith E Mank, Lina Hultin-Rosenberg, Erik Axelsson, and Hans Ellegren. Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Molecular Biology and Evolution*, 24(12):2698–2706, 2007.

[127] Michael Lynch and Bruce Walsh. *The origins of genome architecture*, volume 98. Sinauer Associates Sunderland, 2007.

[128] Z Jeffrey Chen. Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, 15(2):57–71, 2010.

[129] Richard C Tinsley and Joseph A Jackson. Speciation of protopolystoma bychowsky, 1957 (monogenea: Polystomatidae) in hosts of the genus shape xenopus (anura: Pipidae). *Systematic Parasitology*, 40(2):93–142, 1998.

[130] JA Jackson and RC Tinsley. Parasite infectivity to hybridising host species: a link between hybrid resistance and allopolyploid speciation? *International Journal for Parasitology*, 33(2):137–144, 2003.

[131] Justin Ramsey and Douglas W Schemske. Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics*, pages 589–639, 2002.

[132] Charlotte Paquin and Julian Adams. Frequency of fixation of adaptive mutations is higher in evolving diploid than haploid yeast populations. *Nature*, 302:495–500, 1983.

[133] Michael Schmid, Ben J Evans, and James P Bogart. Polyploidy in amphibia. *Cytogenetic and Genome Research*, 145(3-4):315–330, 2015.

[134] DARCY B Kelley. Sexual differentiation in xenopus laevis. In *Symposia of the Zoological Society of London*, number 68. London: The Society, 1960-1999., 1996.

[135] Hans Ellegren and John Parsch. The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics*, 8(9):689–698, 2007.

[136] M Schmid and C Steinlein. Chromosome banding in amphibia. xxxii. the genus xenopus (anura, pipidae). *Cytogenetic and Genome Research*, 145(3-4):201–217, 2015.

[137] Diana S Catz, Leslie M Fischer, Maria C Moschella, Martha L Tobias, and Darcy B Kelley. Sexually dimorphic expression of a laryngeal-specific, androgen-regulated myosin heavy chain gene during xenopus laevis development. *Developmental Biology*, 154(2):366–376, 1992.

[138] Laura A Baur, Brian T Nasipak, and Darcy B Kelley. Sexually differentiated, androgen-regulated, larynx-specific myosin heavy-chain isoforms in xenopus tropicalis; comparison to xenopus laevis. *Development Genes and Evolution*, 218(7):371–379, 2008.

[139] G Sander Van Doorn. Intralocus sexual conflict. *Annals of the New York Academy of Sciences*, 1168(1):52–71, 2009.

[140] Tim Connallon and Andrew G Clark. The resolution of sexual antagonism by gene duplication. *Genetics*, 187(3):919–937, 2011.

[141] Julien F Ayroles, Mary Anna Carbone, Eric A Stone, Katherine W Jordan, Richard F Lyman, Michael M Magwire, Stephanie M Rollmann, Laura H Duncan, Faye Lawrence, Robert RH Anholt, et al. Systems genetics of complex traits in drosophila melanogaster. *Nature Genetics*, 41(3):299–307, 2009.

[142] John H Malone, Doyle L Hawkins Jr, and Pawel Michalak. Sex-biased gene expression in a zw sex determination system. *Journal of Molecular Evolution*, 63(4):427–436, 2006.

[143] Agnieszka Lipinska, Alexandre Cormier, Rémy Luthringer, Akira F Peters, Erwan Corre, Claire MM Gachon, J Mark Cock, and Susana M Coelho. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga ectocarpus. *Molecular Biology and Evolution*, page msv049, 2015.

[144] Judith E Mank, Lina Hultin-Rosenberg, Martin Zwahlen, and Hans Ellegren. Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. *The American Naturalist*, 171(1):35–43, 2008.

[145] Raquel Assis, Qi Zhou, and Doris Bachtrog. Sex-biased transcriptome evolution in drosophila. *Genome Biology and Evolution*, 4(11):1189–1200, 2012.

[146] Wilfried Haerty, Santosh Jagadeeshan, Rob J Kulathinal, Alex Wong, Kristipati Ravi Ram, Laura K Sirot, Lisa Levesque, Carlo G Artieri, Mariana F Wolfner, Alberto Civetta, et al. Evolution in the fast lane: rapidly evolving sex-related genes in drosophila. *Genetics*, 177(3):1321–1335, 2007.

[147] Péter Szövényi, Mariana Ricca, Zsófia Hock, Jonathan A Shaw, Kentaro K Shimizu, and Andreas Wagner. Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Molecular Biology and Evolution*, page mst095, 2013.

[148] Richard P Meisel, John H Malone, and Andrew G Clark. Disentangling the relationship between sex-biased gene expression and x-linkage. *Genome Research*, 22(7):1255–1265, 2012.

[149] Willie J Swanson and Victor D Vacquier. The rapid evolution of reproductive proteins. *Nature Reviews Genetics*, 3(2):137–144, 2002.

[150] Jianzhi Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.

[151] Zhi Zhang, Tina M Hambuch, and John Parsch. Molecular evolution of sex-biased genes in drosophila. *Molecular Biology and Evolution*, 21(11):2130–2139, 2004.

[152] Vera B Kaiser and Hans Ellegren. Nonrandom distribution of genes with sex-biased expression in the chicken genome. *Evolution*, 60(9):1945–1951, 2006.