Genomic Analysis of the Allotetraploid Frog, *Xenopus laevis*

By
Adam M Session

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor Of Philosophy
in
Molecular and Cell Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Daniel S. Rokhsar, Chair
Professor Richard Harland
Professor Michael Eisen
Professor Rasmus Nielsen

Spring 2015

Abstract

Genomic Analysis of the Allotetraploid Frog, *Xenopus laevis*

by

Adam M Session

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Daniel S. Rokhsar, Chair

Duplication has long been recognized as an evolutionary source of novelty. The relaxation of purifying selection following duplication allows for normally deleterious mutations to persist long enough to give rise to novel phenotypes. Whole-genome duplications (WGDs) are a specific type of duplication, in which a species suddenly finds itself with two copies of all of its genomic loci. While the fate of most of the duplicated loci is to be lost, those that persist are thought to underlie the innovations seen in groups with a history of polyploidy, such as flowering plants, yeast, *Paramecium*, and vertebrates. These ancient events give us an idea of how WGDs can drive the radiation of large and diverse phyla, but do not give us any information on the genomic response immediately following polyploidy. This thesis provides insights into the origins of polyploidy and its effects on genome dynamics.

There are two models for the mechanism of polyploidy: autopolyploidy and allopolyploidy. Autopolyploids are formed by doubling the somatic chromosomes in the zygote or early embryo. Allopolyploids are formed by the hybridization of two related, but genetically distinct, species, followed by chromosome doubling. If there are no extant diploid relatives, it can be difficult to distinguish between these two models. One feature of allopolyploids is the lack of recombination between their homeologous chromosomes. The end result is that any markers that were unique to each species while apart, such as transposable element subfamilies, will be asymmetrically distributed on the progenitor chromosomes in an organism that recently underwent a WGD.

Xenopus laevis is an important vertebrate model in developmental and cell biology that has experienced a recent WGD (~40 million years ago [MYA], based on cDNA alignments (Hellsten, 2007). Its diploid cousin *Xenopus tropicalis* has become a popular genetic model frog. Comparative analysis of these two frog genomes gives us an excellent opportunity to study genome dynamics following whole genome duplication. The discovery of asymmetrically distributed transposon subfamilies supports the model that cross-species hybridization through allotetraploidy is the mechanism underlying the polyploid *Xenopus* radiation. Thus, the sub-genome sequence divergence of 40 MYA dates the divergence of the progenitor species, not the hybridization event. The asymmetric distribution of these elements between homeologous sequences allows us to assign chromosomes to progenitor species, named “A” and “B”, making *X. laevis* a unique system to study sub-genome-specific evolution. The wealth of transcriptome and epigenetic data available for *Xenopus* allows me to assay how these genomic changes affect gene expression as well as gene retention. The combination of these resources with genomic data gives me the resolution needed to date the hybridization both by studying the decay of unitary pseudogenes and by comparative analysis of the transposable elements discussed above.

The sub-genome from progenitor species “A” has more assembled length, longer chromosomes, a higher rate of gene retention, and higher average expression in the adult frog. The B sub-genome has higher synonymous and nonsynonymous mutation rates. The chromosomes orthologous to *X. tropicalis* 9 and 10 are fused in both sub-genomes of *X. laevis*, forming homeologous chromosomes 15 and 18, and deviate from the A/B trends discussed above. The regions of these *X. laevis* chromosomes orthologous to *X. tropicalis* chromosome 10 have a lower density of diagnostic repeats, no sub-genome bias in gene retention, and have a higher silent substitution rate. This divergence from the rest of the genome is not shared by the

regions orthologous to *X. tropicalis* 9. I hypothesize that the short length of *X. tropicalis* 10 plays a role in these deviations due to a higher rate of gene conversion on shorter chromosomes.

Chapter 1
Introduction

Developmental biology is the study of the process by which multicellular organisms grow from a single cell into a multicellular organism. Scientists have been cataloging the differences between plants and animals since the time of Aristotle in an attempt to understand what controls the different characteristics, or “attributes”, between organisms (Arist. PA I, trans. Ogle). Based on his observations of chicken embryos, Aristotle proposed that embryos were not preformed, miniature organisms, but instead that form and structure emerged gradually as the embryo developed. This hypothesis was challenged for centuries by early embryologists (Hartsoeker 1694) until the proposal of Cell Theory in the mid-19th century [Schwann 1839]. Cell Theory proposes that all life is made up of cells, and that cells are only born from previously existing cells. By the late-19th century August Weismann proposed that the sperm and egg of animals were “germ cells” that carried the hereditary material, and that the somatic cells of an organism could not contribute to the next generation (summarized in Fig. 1.1, Wolpert, Principles of Development).

Ploidy refers to the number of sets of similar chromosomes contained in a cell. Haploid cells have a single copy of each chromosomes, whereas diploid cells have two copies of each chromosome, and triploid cells have three copies. Theodor Boveri used sea urchin embryos to show that haploid sperm and egg cells fuse their nuclei to create a zygote with the diploid chromosome count of the parents (Boveri, 1902. Boveri, 1907), providing a physical basis for the transmission of genetic characteristics discovered by Mendel (Mendel, 1865. Discussed more later in this chapter). This is an early example of the importance of the interplay between developmental biology and genetics.

In the 1880’s Weismann proposed a model of development in which the nucleus of the zygote contained a number of “determining factors”, which are unequally partitioned into the dividing cells of the embryo (Weismann 1892). This led to experiments by Hans Driesch where he allowed a sea urchin egg to be fertilized and undergo the first cell division. He then separated the cells and found that the surviving cell developed into a smaller, but otherwise normal larva (Fig. 1.2, Wolpert, Principles of Development). This experiment shows that development is a process regulated by determining factors that are not asymmetrically distributed from the first cell division, a more complicated model than the one proposed by Weismann. This concept implies that cells have an intrinsic ability to specify their own fate. As part of this ability, work by Hans Spemann and Hilde Mangold provided evidence for induction, the process by which one cell directs the development of a neighbor, in 1924. Spemann and Mangold transplanted sections from the blastopore of a newt gastrula onto the opposite side of the gastrula of a related species having different pigmentation (Fig 1.3, Wolpert, Principles of Development). The transplanted tissue induced the formation of ectopic neural tube, somites and gutin the host embryo, while the graft self differentiated mostly into notochord. The small region they identified, known as the “organizer”, controlled the organization of the embryonic body. These experiments, along with many others from the turn of the twentieth century, established that development is determined by factors that are present in each cell of the early embryo (or else Driesch’s sea urchins would have developed abnormally), but the different cells of the embryo have different factors that specify cell fates (or else Spemann and Mangold’s transplants would not have induced ectopic tissues). Discussing these factors in more detail requires knowledge of genetics.

Genetics is the study of heredity and variation in living organisms. Humans have been breeding plants and animals for specific traits for thousands of years. The extreme variation seen in domesticated organisms was used to build the foundation of Charles Darwin’s Theory of Evolution [referencing Origin of Species chapter 1]. The origin of modern genetics can be traced to the Augustinian monk, Gregor Mendel. Mendel was interested in how traits, or phenotypes as

they would later be called, were inherited between generations. Mendel used the inheritance of a number of visual traits in pea plants, such as smooth vs. wrinkled seeds, to ask whether the inheritance of individual traits was the same between generations and if the inheritance of different types of traits (for example, seed shape and seed color) was controlled by similar mechanisms. An example cross for seed shape is shown in Figure 1.4. Mendel crossed a pea plant with all wrinkled seed to another with all smooth seeds. The resulting cross (F_1 generation) is made up of plants with all smooth seeds. Self-crossing the F_1 plants generates an F_2 generation, made up of plants with ~3:1 ratio of smooth:wrinkled seeds. Mendel proposed that there are "factors," later called genes, which are inherited from each parent, and (along with environmental factors) specify the phenotype (such as seed shape) of the organism. The collection of genes an organism inherits is known as its genotype.

Mendel went on to detail specific mechanisms of inheritance for different genes, but Mendel had worked with traits of whole organisms. He did not investigate how characteristics are sorted and combined on a cellular level, where reproduction takes place. In 1902, the German scientist Theodor Boveri and the American Walter Sutton, working independently, suggested that chromosomes could be the units of heredity. Chromosomes are linear molecules of DNA that get compacted with proteins into chromatin within the nucleus of eukaryotes. Boveri showed that chromosomes remain organized units through the process of cell division (Boveri 1907), and he demonstrated that sperm and egg cells each contribute the same number of chromosomes to the zygote of an animal (Boveri 1902).

Sutton had also become familiar with the process of "reduction division" (later called meiosis), which gives rise to reproductive germ cells (Sutton 1902). In meiosis, the number of chromosomes is halved in sperm and egg cells (1N), with the original number restored in the zygote, or fertilized egg (2N). This proposed mechanism was consistent with Mendel's idea of segregation, and the pairing of homologous chromosomes is known as bivalent pairing. By taking an experimental approach, Boveri sought to understand whether differences in the inheritance of chromosomes caused differences in the developing embryo. Sea urchin eggs can be fertilized with two sperm, and he showed that the embryos resulting from such double unions possess variable numbers of chromosomes. These included aneuploid embryos, with chromosome counts that are not simple integer multiples of the haploid chromosome count. Boveri found that only those embryos from the double fertilization that had the correct number of chromosomes (36) would develop normally. Boveri and Sutton both recognized that the Mendelian concepts of segregation could be applied to chromosomes based on these data, with chromosomes containing the "factors" of inheritance (genes).

In the early twentieth century, embryology and genetics were still distinct disciplines. The rediscovery of Mendel's ideas sparked interest into how inheritance contributed to evolution, but not to the development of an individual embryo. After showing the structure of DNA to be a double helix (Watson & Crick 1953 Nature), Francis Crick proposed the central dogma of molecular biology, summarized by the phrase "DNA becomes RNA becomes protein" (Francis Crick 1956). The "determining factors" Weismann had proposed were the protein products of gene expression. By the mid-twentieth century, understanding how genotype dictated the phenotype of protein expression in the developing embryo became a major focus of developmental biology. Classical genetic techniques, such as gene knockdowns or overexpression, are now common practice in studying how a gene contributes to the development of an organism. Transitioning into a more genetics-focused methodology proved particularly interesting for the model frog, *Xenopus laevis*.

1.1- *Xenopus laevis* and polyploidy

Historians of biology often stress that model organisms are made, not found. Developing a model system requires a large amount of time and thought to establish the proper laboratory conditions for mating, and establishment of stocks. *Xenopus laevis* was used in the nineteenth century for scientific purposes sporadically in Europe and Africa, and occasionally for

recreational aquaria in Europe. Many of the early embryological studies used newts and axolotl (Spemann 1924). The eggs of newts are larger and more tractable to microsurgery than *X. laevis*, and early scientists were confounded by the phylogenetic position of *X. laevis*, which appeared to be “neither a typical frog nor toad” (Gurdon and Hopwood 2000). This began to shift with British endocrinologist Lancelot Hogben, who established in the 1930’s that ox anterior pituitary extracts induce ovulation in *X. laevis* females(Gurdon and Hopwood 2000). This experiment was soon seen as a potential pregnancy test for humans, and Hogben worked hard to establish *Xenopus* in laboratories around the world as a model for reproductive physiology (Gurdon and Hopwood 2000). By the end of World War II, pregnancy testing had made *Xenopus* a regular laboratory animal. In the mid-twentieth century, the zoologist Pieter D. Nieuwkoop focused on improving laboratory protocols for *X. laevis* to be used as a embryological model. Nieuwkoop noted that the ability to spawn year round made *Xenopus* a more attractive system for embryology than the local Northern hemisphere frogs, which were strictly seasonal, and established Normal Tables (standards of development which play an important role in establishing laboratory animals) for *Xenopus* (Nieuwkoop, 1994). Throughout the mid-twentieth century different species of *Xenopus* were identified and characterized.

The field of molecular biology reoriented around genetic and biochemical approaches after the discovery of the structure of DNA. In the 1970’s several groups studied the DNA content and karyotype of *Xenopus* species and hypothesized that several species in the phylum have undergone Whole Genome Duplications(WGDs, or polyploidy), an event where an organism has twice as much DNA as certain related species. Table 1.1 shows the chromosome number and relative DNA content of several *Xenopus* species compared to *X. laevis* (taken from Thiebaud&Fischberg, 1977). With the exception of *X. tropicalis*, with chromosome count 2N=20 and half the DNA content of *X. laevis*, all *Xenopus* species had 2N counts of 36, 72, or 108 chromosomes, with DNA contents equal to or greater than *X. laevis* (Thiebaud&Fischberg, 1977). The presence of more DNA in those species with 2x or 3x the chromosome count of *X. laevis*, and the prevalence of polyploidy in amphibians, led authors to hypothesize that the genus *Xenopus* had undergone multiple rounds of polyploidy. Fischberg later identified *Xenopus epitropicalis*, with 2N=40 and twice the DNA content of *X. tropicalis* (Tymowska and Fischberg, 1982). Comparing the karyotypes of *X. tropicalis* to *X. epitropicalis,* Fischberg observed that for each quartet of chromosomes in *X. epitropicalis*, there was a duet in *X. tropicalis*. However the banding patterns of the quartet chromosomes of *X. epitropicalis* are not shared between all four chromosomes, and bivalent pairing of chromosomes with similar banding patterns is observed. In other polyploid amphibians, such as *Ceratophrys*, multivalent pairing is observed between chromosome quartets (meaning that more than 2 chromosomes associate with one another, Schmid 1985). Fischberg hypothesized that the bivalent pairing is strong evidence for allopolyploidy (polyploidy by hybridization of multiple species) being the main mechanism of polyploidy in *Xenopus*. The differences between multivalent and bivalent pairing in tetraploids is shown in Figure 1.5. A proposed phylogentic tree for *Xenopus* is in Figure 1.6, and the proposed pathway to polyploidy is included in Figure 1.7 (adapted from Kobel and Dupasquier, 1986).

It has become clear that all *Xenopus* species, except *X. tropicalis*, have a recent polyploid history (Kobel and Dupasqueir, 1986). As such, *X. tropicalis* has become a favored model for genetic experiments, while developmental experiments are often still performed in *X. laevis*, which has larger embryos that are more tractable to manipulation.

Recently, whole genome shotgun sequencing has allowed for the study of the entire genomic sequence of an organism. The genomic redundancy following polyploidy can complicate genome assembly, because of the difficulty of cleanly separating different homeologs of similar sequence; however current methods are able to overcome these difficulties, so long as the rate of DNA polymorphisms is sufficient to show differences between homeologs in single DNA sequencing reads (Ming, 2015). One step towards sequencing a

polyploid organism like *X. laevis* is sequencing of related diploids such as *X. tropicalis*, which has been performed (Hellsten, 2010). The object of this thesis is to understand the evolutionary impact of polyploidy on the genomic history of *X. laevis* through comparative analysis with *X. tropicalis*. I will discuss evaluating genome assembly, differentiating between autopolyploid and allopolyploid origins, and what the structural and functional evolutionary trends seen in the *X.laevis* genome reveal about its molecular history. I will also discuss analysis of a allohexaploid grass, *Triticum aestivum*, in the final chapter.

In the remainder of this chapter I will discuss the evolutionary hypotheses surrounding polyploidy and discuss using comparative genomics and development as a tool to study evolution. Finally, I will briefly discuss the significant contributions of my collaborators towards providing the data needed to assemble the *X. laevis* genome, and allow for my analysis.

1.2- Evolutionary impact of WGDs

Whole genome duplications, or polyploidy, are thought to have contributed to the radiation of a number of phyla, including vertebrates, flowering plants, yeast, teleost fish, *Paramecium*, and *Xenopus* (Otto, 2007). Polyploidy introduces redundancy at all genomic loci, which can be resolved in a number of scenarios. A new polyploid that cannot overcome the genomic instability, or has lower survival and/or reproductive rates, may become an evolutionary dead-end that does not survive. Alternatively, if the initial shock following redundancy can be overcome, the polyploid can establish a new species. One initial shock to overcome is that introduced by sexual determination (Comai, 2005. Wertheim, 2013. Otto, 2007). Originally sexual incompatibility in polyploid animals was thought to underlie the increased prevalence of polyploidy in plants relative to animals (Ramsay, 1998). In *Xenopus*, it is already known that *X. laevis* and *X. tropicalis* have different sex-determination loci (Wells, 2011). The prevalence of polyploidy in some lineages of modern fish,amphibians and reptiles, as well as ancient fish and pre-vertebrates, would suggest that meiotic stability is not a significant barrier to animal polyploidy (Otto, 2007).

Regardless of the difficulty in establishing a new species, once a polyploid species establishes a diploid genetic system, the genomic redundancy allows loci to be mutated and lost. While the fate of most duplicated loci is ultimately to be lost, those that remain may gain new functions (neofunctionalization), or partition their function with their duplicate (subfunctionalization). An example of neofunctionalization is the vertebrate glucocorticoid receptor, which evolved specificity from a more promiscuous receptor (Bridgham, 2001). An example of subfunctionalization can be seen with *X. laevis skp1a*, which has partitioned expression domains in the developing embryo (Hellsten, 2007).

Analysis of ancient WGDs, such as yeast, vertebrates, and *Paramecium* suggests that different gene families may be subject to selective pressures following polyploidy to either be retained as two genes that subfunctionalize, or to be reduced to a single-copy locus (Scannell, 2007. Aury, 2006. Putnam, 2008). DNA repair machinery is often reduced to single-copy (Scannell, 2007. Aury, 2006). It could be that DNA repair is such an important process that any mutations affecting protein function in DNA repair machinery would have a dominant-negative effect, impairing the function of the remaining copy. Dominant negative effects are easy to argue, but difficult to prove as we do not know all of the possible deleterious mutations a gene can undergo. Studying the genomes of recent duplicates, such as *X. laevis*, may give us a view into the prevalence of dominant-negative mutations.

By analyzing…...Lynch and Conery have shown that the average lifetime of an isolated gene duplicate is 3–7 million years (Lynch and Conery, 2000). Genomic analyses of polyploids have shown that different lineages have different gene retention rates. For example, ~8% of duplicated genes have remained in yeast over ~100 MY following polyploidization (Scannell, 2007), ~72% have remained in in maize over ~11 MY (Tanksley, 1993), ~52% in rainbow trout over 96 MY (Berthelot, 2014), and ~47% in catastomid fishes over ~50 MY (Ferris and Whitt, 1979). These retention rates are higher than one might expect under a totally neutral model of

gene loss, as proposed by Lynch and Conery. While differences in generation time must contribute to the differing rates of gene loss, some amount of differential gene loss must be due to phenotypic differences between the polyploid phyla. Also, Lynch and Conery's study analysed the fate of individual duplicate genes in a diploid context; in polyploids, all gene loci are duplicated relative to the diploid progenitors, so stoichiometric considerations must also come into play. Some models for gene retention following polyploidy include the requirement for gene balance, heterozygote advantage, and selection for higher levels of gene expression (Yao, 2010. Birchler, 2010. Wertheim, 2013. Otto, 2002. Kondrashov, 2002). There is also evidence that numbers of interaction partners of proteins encoded by yeast duplicates are larger than those of singletons (He and Zhang, 2005). This might imply that the potential for subfunctionalization is strong enough pressure to drive the retention of certain loci.

Although there is debate on the relative importance of the different mechanisms of selection to preserve gene duplicates, it is accepted that duplicated genes contribute significantly to novel functions. Human three-color vision is due to gene duplication (Tan and Li, 1999), as are our complex immune systems (Nei et al. 1997). Is there strong evidence that polyploidization drives evolution in a unique fashion? This is a difficult question to answer, as pointing towards specific loci, such as the four HOX clusters that arose at the base of the vertebrate radiation, as proof that polyploidization drives evolution is not proof that localized duplications could not drive a similar process. Indeed, tandem duplications are implicated in the evolution of the *Drosophila* visual system (Bao and Friedrich, 2009), human three-color vision, and in the evolution of caffeine production in progenitor species of modern coffee plants (Denoeud, 2014). Understanding the role of polyploidization in genome evolution requires studying all the changes in a genome, and how they relate to one another. Additionally, we need a set of null hypotheses built from studying species divergence without polyploidization. These analyses fall under the relatively new field of comparative genomics.

1.3- Comparative genomics and development as a tool to study evolution

Homology, in a biological context, is defined as the shared ancestry between any two structures or sequences. For molecular sequences homology can be partitioned into a number of types. Orthologs are homologs between species that are separated by a speciation event. Paralogs are homologs within a species that are descended from duplication events. In the case of allopolyploidy, two species hybridize the diploid complements of their genomes into a novel organism. This unique history means that orthologs became paralogs. Historically these paralogs have been known as ohnologs, after Ohno who proposed the 2R hypothesis of ancient vertebrate duplications (Ohno, 1970). In this work we will use the most common term in the literature, homeolog, to refer to duplicates specifically formed through polyploidy. A number of tools have been developed to identify homologous sequences between or within organisms, all starting from sequence alignments. The most popular is the Basic Local Alignment Search Tool (BLAST), which is used extensively in this thesis (Altschul, 1990). Similarity between sequences is *prima facie* evidence for some sort of evolutionary homology, but other methods (e.g., using sequence context and/or phylogenetic analysis) are needed to confidently determine orthology, paralogy, or homeology.

All genomic sequences are subject to mutations, but a fraction of them (such as protein-coding genes and their enhancers) are subject to selection. The selection on these sequences can be positive, making an allele more likely to be passed on with each generation and eventually to become fixed in the population. Alternatively the selection for an allele can be negative, making it less likely to be passed on. The selection to keep a sequence the same (maintaining the ancestral allele, while selecting against new ones) is known as purifying selection. The genetic redundancy introduced by duplications weakens purifying selection, allowing normally essential genes to gain potentially deleterious mutations and be lost from the genome. If the extra copy is lost or suffers a deleterious mutation for part of its function, the remaining copy is once again subject to purifying selection. In the case of polyploidy the entire

genome is duplicated. With this type of duplication, all loci are duplicated at once, which causes gene dosage effects. In some cases purifying selection may be maintained on both copies of the duplicated gene so that it is not titrated to the point that it can no longer perform its necessary function.

Sequence changes not only affect protein function, but also gene expression. Differential deployment of orthologous genes across developmental time and space can lead to significant phenotypic differences. In vertebrate adults, often the same genes are deployed by orthologous organs (Chan, 2009). This suggests that the massive differences in form we see in the adults must be due to changes in their development. This makes conceptual sense: developmental trajectories of animals are tightly regulated to ensure the proper adult form, so many different forms could be made by subtly shifting the gene expression during development (From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, 2nd Edition). This makes developmental biology a key tool in comparative genomics, allowing us to compare the time and space variables of gene expression, and the consequences of variable expression to the phenotype.

Comparative genomics of a polyploid organism can be difficult since the homeologs can violate the assumptions of standard orthology tools. Genomic analysis of polyploids is often impeded by the lack of genome sequence, even when a wealth of transcriptome data is available, as is true with *X. laevis* (Hellsten, 2007). We performed whole genome sequencing of the *X. laevis* genome to analyze polyploidization, as well as work to provide a resource to developmental biologists, whose work is aided significantly by understanding how the DNA flanking a gene affects the expression of a gene of interest.

1.4- Whole genome sequencing of a polyploid genome and conservation of synteny (Meraculous and BAC-FISH)

Whole genome shotgun (WGS) sequencing is a method for obtaining nucleotide sequence data without any a priori knowledge of sequence from that species [Weber and Myers, 1997]. Genomic DNA is extracted from the cells of an organism. The DNA is fragmented and converted into libraries that place known sequencing primer binding sites on either side of the fragments of unknown sequence. For Sanger sequencing, this is accomplished by cloning the DNA into plasmid, fosmid, or BAC vectors that can hold inserts of different sizes. For Illumina sequencing, adapters are added directly by ligation. In either method, the unknown insert sequence is determined by taking advantage of the known flanking sequencing primer binding sites to initiate reads through the unknown sequence. The end-sequences of these clones, or “reads”, are then aligned to each other using computational methods, and based on overlaps between reads, long stretches of contiguous sequence without gaps, called “contigs”, can be constructed. High confidence in the alignments and overlap of reads is ensured by sequencing enough DNA such that each nucleotide in the genome is covered on average by multiple reads. A typically reported genome sequencing statistic is coverage: 30X coverages implies that each position in the genome is sampled by 30 reads on average.

In the case of sexually mating organisms it is important to consider that the DNA we isolate is from both sets of chromosomes. This propagates polymorphisms between the chromosomes, and can confound computational methods at the comparison step, so common practice is to inbreed the organism of interest for sequencing, so that both sets of chromosomes are identical and easier to reassemble into an organized genome sequence. Often multicellular genomes contain large amounts of repetitive DNA, which is often organized into large tandem arrays (Lopez-Florez, 2012). These can be difficult to assemble into contigs because similar sequences are present throughout the genome, leading to ambiguities about which unique genomic sequences lie definitively on each side of a specific repeat sequence. Thus these regions, that cannot be assembled definitively, create limits to contig length. However, using end-sequencing of the varying insert sizes discussed above, contigs can then be computationally assembled together into “scaffolds”, which may have gap sequences. This

“paired-end” sequencing allows us to assemble scaffolds across the repetitive arrays discussed above. The largest scaffolds would represent chromosomes, however it can be difficult to assemble entire chromosomes without additional long-range information from genetic map data. This problem is discussed further in Chapter 2.

1.4.1-History of the J strain

We sequenced an inbred *X. laevis* whose history begins with four pairs of frogs that were introduced from Switzerland to the U.S.A in 1948. These frogs were called the 1st generation (gen.) in the pedigree of J strain. Then, four pairs from the 1st generation were introduced from Iowa State University to Gunma University, Japan, in 1948. Initial records for raising frogs in the U.S. and Japan are missing, but for the purpose of counting generations we assume that matings for the next generation was performed once every two years on average. The Katagiri group at Hokkaido University, Japan, also mated frogs once every two years in average for about 20 years. In 1973, their inbred frogs exhibited “no short-term skin rejection” (Tochinai & Katagiri, 1975; Katagiri, 1978), indicating that the MHC locus is almost homozygous. This population was called the G group (G stands for Gunma Univ.) (Katagiri, 1978; Nakamura et al., 1985) or the J group or J line (J stands for Japan), and assumed to be the 10th generation. At the 21st generation after repeated single-pair mating, the J line exhibited “no long-term skin rejection,” indicating that most genes are homozygous (Izutsu & Yoshizato,1993). This population was called the J strain and was used for our genome sequencing.

1.4.2- Contributions from collaborators

The genomic DNA, cDNA, and RNAseq libraries for the *X. laevis* genome project were provided by a number of sources. Masanori Taira’s group provided the BAC and RNAseq libraries. Plasmid libraries were prepared by Christian Haudenfeld at Illumina. Fosmid libraries were provided by both the Taira group and the Kitzman group. Hi-C data were kindly provided by Ian Quigley. It had been previously shown that *X. tropicalis* and *X. laevis* chromosomes were largely syntenic by cDNA FISH (Uno, 2013). Using BAC-FISH within *X. laevis* the Taira group identified a number of rearrangements that are discussed more in Chapter 2.

The *X. laevis* draft assembly 7.1 (Xenla7.1, summarized in Table 1.2) was produced at the DOE Joint Genome Institute by Jarrod Chapman with Meraculous, as previously described (Chapman, 2011). Xenla7.1 contains 413,763 scaffolds spanning approximately 2.72 Gbp. Roughly half of the assembled sequence contained in 648 scaffolds ranging in size from 1.1 to 21.56 Mb. The next chapter discusses my assessment of the genome using transcriptome data, and annotation of different genomic elements, before discussing the evolution of those elements in more detail in later chapters.

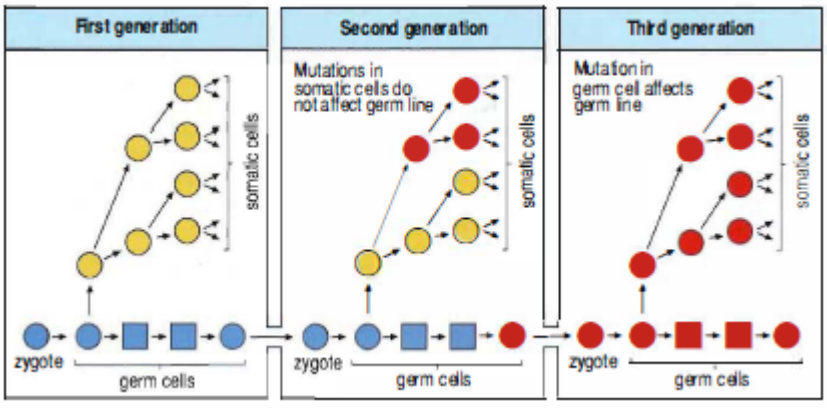


Figure 1.1: The distinction between germ cells and somatic cells
Taken from Principles of Development by Lewis Wolpert (2nd edition). In each generation germ cells give rise to both somatic cells and germ cells, but inheritance is through the germ cells

only. Changes that occur due to mutation in somatic cells can be passed on to their daughter cells but do not affect the germ line.

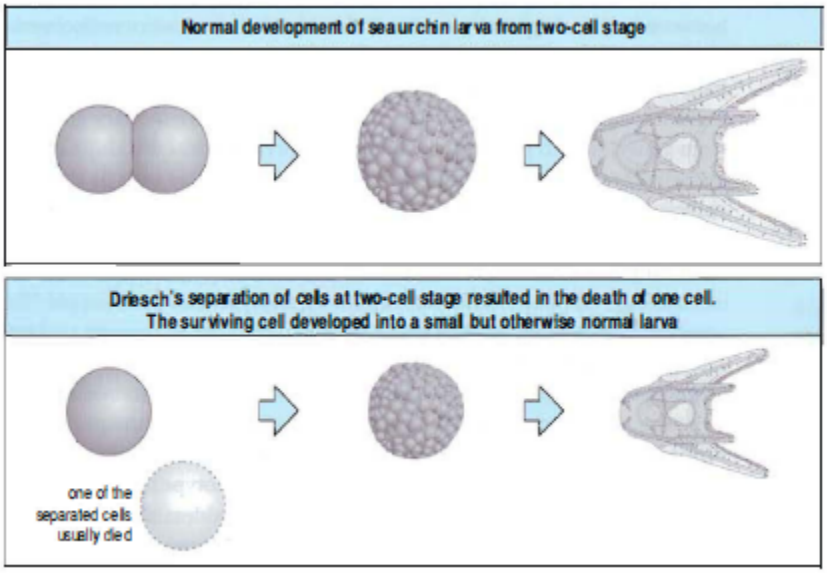


Figure 1.2: The outcome of Driesch's experiment on sea urchin embryos
Taken from Principles of Development by Lewis Wolpert (2nd edition). After separation of cells at the two-cell stage, the remaining cell develops into a small, but whole, normal larva.

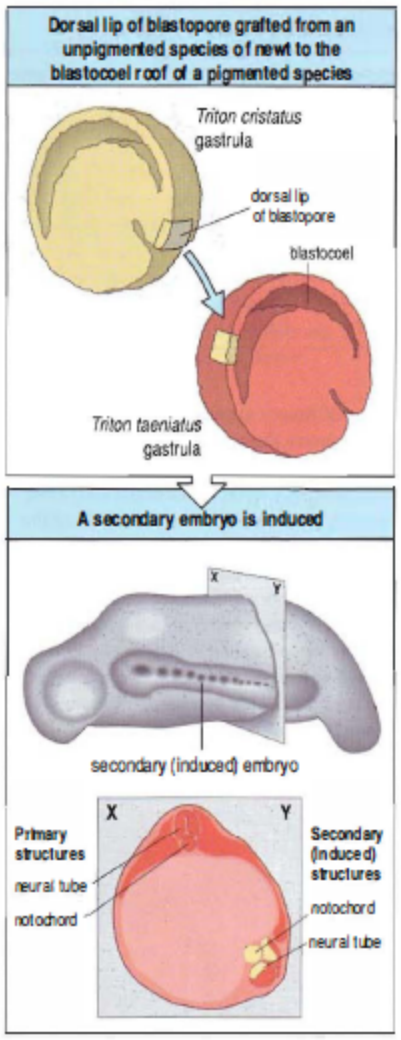


Figure 1.3: The dramatic demonstration by Spemann and Mangold of induction of a new main body axis by the organizer region in the early amphibian gastrula.
Taken from Principles of Development by Lewis Wolpert (2nd edition). A piece of tissue (yellow) from the dorsal lip of the blastopore of a newt (*Triton cristatus*) gastrula is grafted to the opposite side of a gastrula from another, pigmented, newt species (*Triton taeniatus*, pink). The grafted tissue induces a new body axis containing neural tube and somites. The unpigmented graft tissue forms a notochord at its new site (see section in lower panel) but the neural tube and the other structures of the new axis have been induced from the pigmented host tissue.

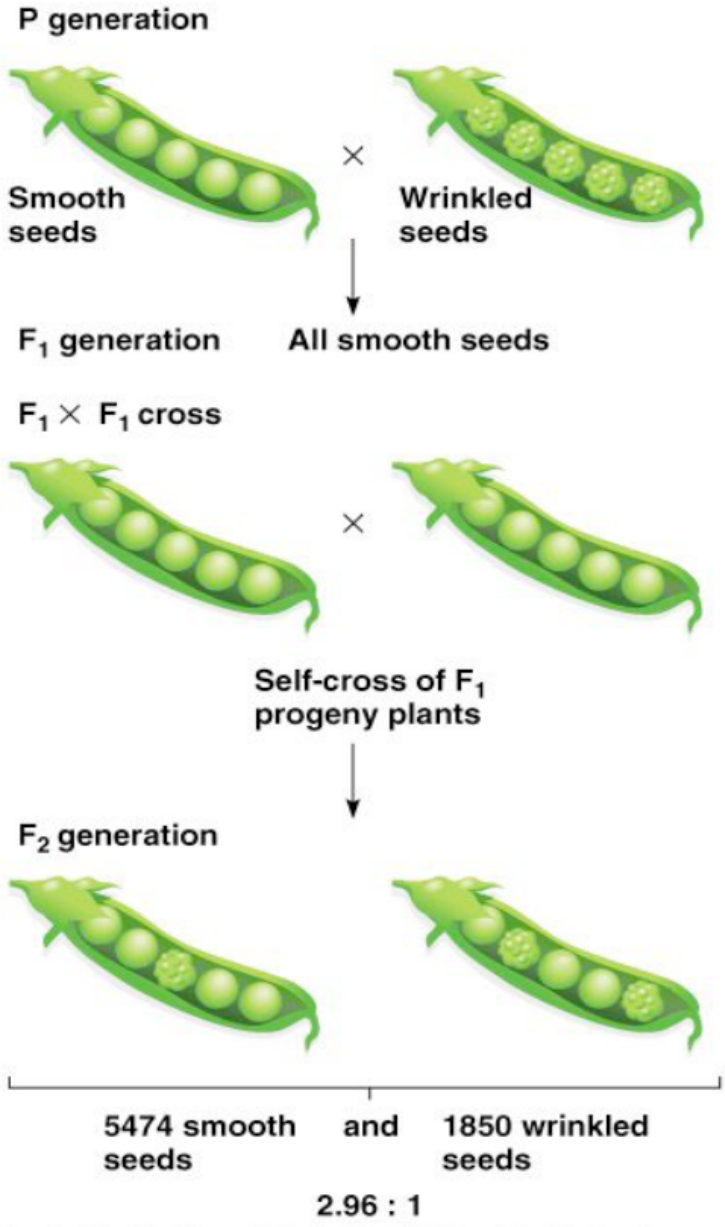


Figure 1.4: Results of one of Mendel's breeding crosses
Taken from iGenetics: A Mendelian Approach by Peter J. Russell. In the parental generation (P), Mendel crossed a true-breeding pea strain that produced smooth seeds with one that produced wrinkled seeds. All the F1 progeny seeds were smooth. The F2 progeny produced both smooth and wrinkled seeds in a 2.96:1 ratio.

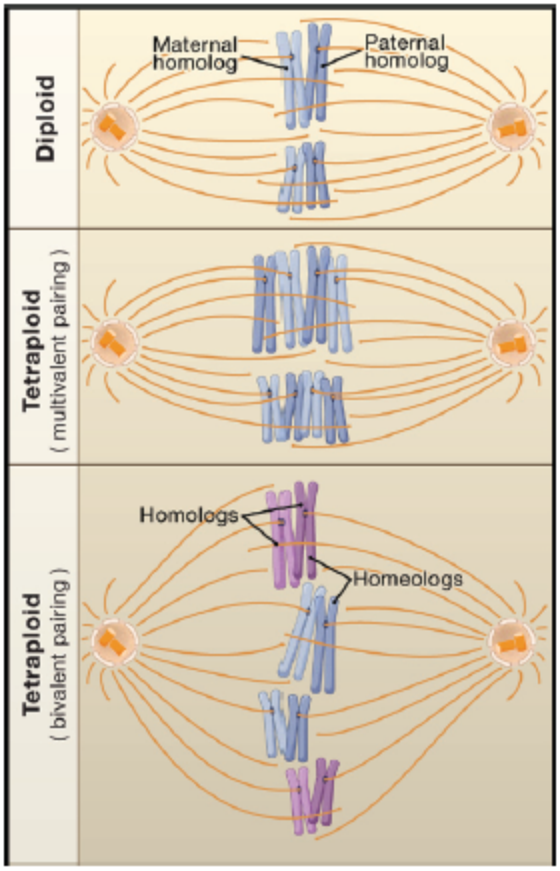


Figure 1.5: Polyploidy Terminology
Taken from Otto, 2007. Two nonhomologous chromosomes are shown (long and short), with each X-shaped chromosome representing a pair of sister chromatids joined at the centromere. In diploids, each chromosome consists of a homologous pair, with one chromosome inherited from the mother and one from the father. In tetraploids (B and C), the chromosomes are further doubled. When the duplicated chromosomes are very similar to one another, they might align randomly in pairs during meiosis (bivalent pairing; not shown) or all align together (multivalent pairing; In either case, gametes may inherit any combination of parental chromosomes (multisomic inheritance), and mutations that arise on one chromosome can spread to all other copies, inhibiting their divergence. When polyploidization involves chromosomes that are sufficiently distinct (that is, “homeologs”; differentiated by blue and purple), the more similar pair of chromosomes tend to align together to the exclusion of the other pair. With strict bivalent pairing, the homeologs behave as distinct chromosomes and segregate independently (disomic inheritance), allowing their divergence. Newly formed autopolyploids typically exhibit multisomic inheritance, whereas newly formed allopolyploids exhibit a variety of patterns of inheritance, depending on the cross (Ramsey and Schemske, 2002).

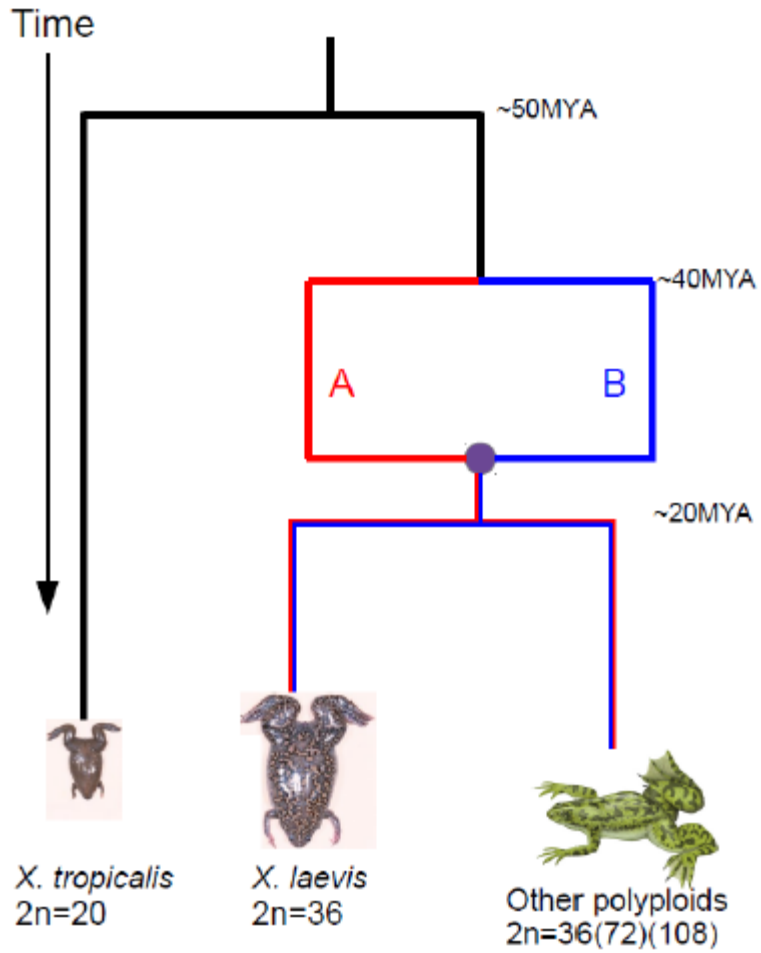


Figure 1.6: *Xenopus* phylogeny
A phylogenetic tree illustrating the unique evolutionary history of *Xenopus*. By comparing the rate of sequence change in protein-coding genes, we calculate ~50MY divergence between *X. laevis* and *X. tropicalis*, and ~40MY divergence between the *X. laevis* progenitors (Hellsten, 2007). The 20MY divergence of the polyploid radiation is based on mitochondrial gene divergence (Evans, 2004).

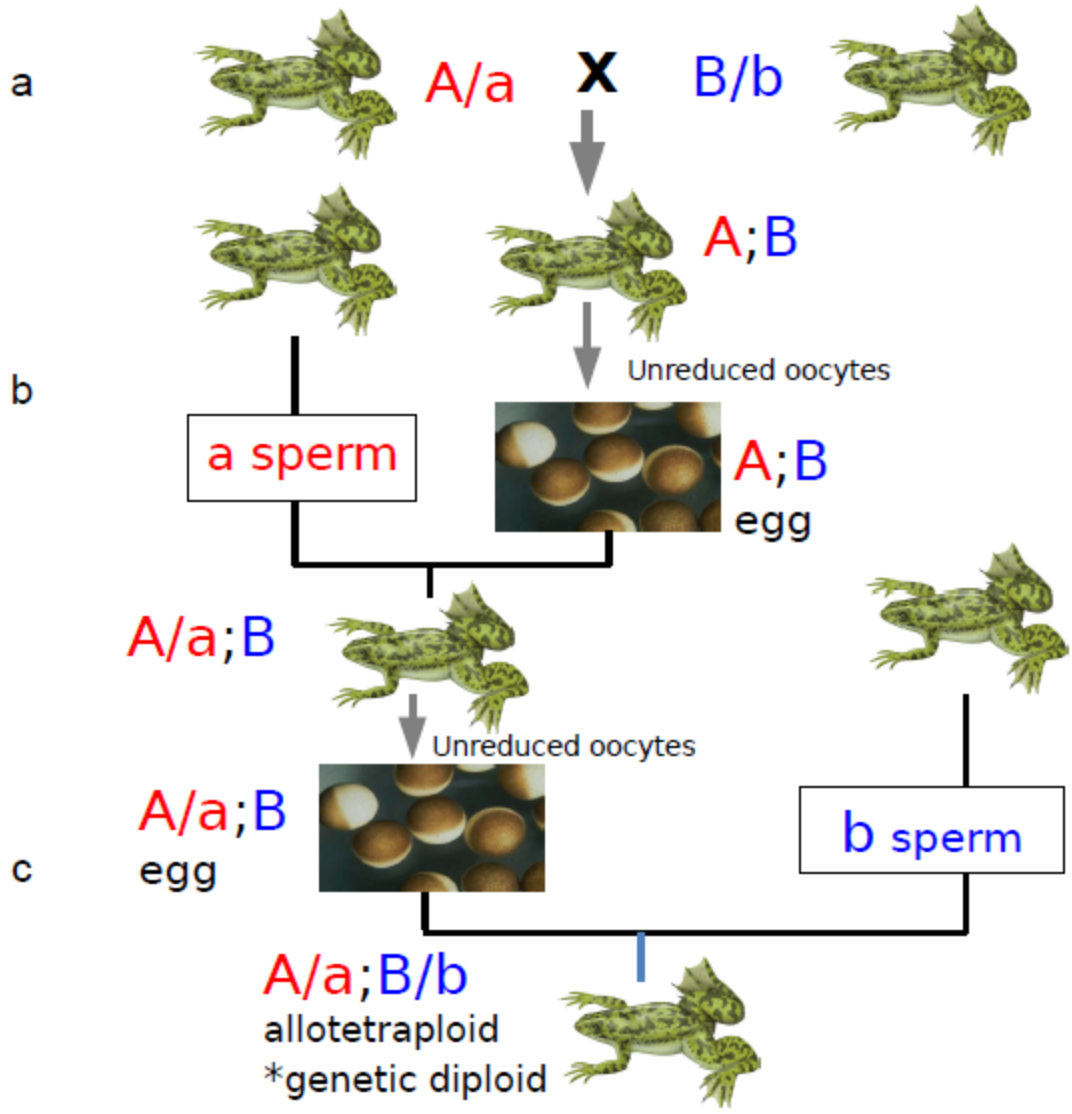


Figure 1.7: Interspecific hybridization and *Xenopus* allopolyploidy
Adapted from Du Pasquier,Kobel 1986. Proposed mechanism for generation of a allotetraploid frog population from two related diploid species. a. Two distinct species of frogs (A and B) mate to form a hybrid whose chromosomes cannot recombine. b. The lack of recombination makes the male hybrids sterile, but the females of the hybrid population lay unreduced oocytes. If the hybrid females mate with one of the two fertile parent species they can create an allotriploid population, where one set of chromosomes cannot recombine (B shown here). c. Similarly to the hybrid, in the allotriploid males are sterile and the females lay unreduced oocytes. If the females mate with males from the complementary parent species, the offspring will be allotetraploid. This final mating restores a genetically diploid state to all chromosomes, and recombination/meiosis can progress as normal in this novel genetically diploid population. Novel polyploid frogs can be made in the lab between modern Xenopus species by this mechanism (Kobel and Du Pasquier, 1986).

Species Subspecies	Chromosome number[a]	Number of measure-ments	DNA content in % of *X.l.l.*			Absolute unit pg/nucleus
			x	S^2	S in %	
Xenopus tropicalis	20	517	56	19.4	7.9	3.55
Xenopus laevis						
X. l. laevis	36	1413	100[b]	35.7	6.0	6.35
X. l. petersi	36	443	101[b]	35.6	5.9	6.35
X. l. victorianus	36	462	101[b]	42.6	6.4	6.35
Xenopus gilli	36	519	100[b]	48.8	7.0	6.35
Xenopus fraseri	36	397	101[b]	48.3	6.9	6.35
Xenopus borealis	36	732	112	43.1	5.9	7.10
Xenopus muelleri	36	615	120	43.6	5.5	7.60
Xenopus clivii	36	929	133	64.3	6.0	8.45
Xenopus vestitus	72	388	202	125.4	5.5	12.83
Xenopus sp.n.	72	487	198	125.8	5.7	12.57
Xenopus ruwenzoriensis	108	547	256	223.7	5.8	16.25
Human lymphocytes	46	536	89	48.4	7.8	7.15[c]

[a] Data from Tymowska and Fischberg (1973) and personal communications
[b] Statistically not different at the 0.05 level
[c] Mean calculated on the basis of data given by Atkin et al. (1965) and Bachmann (1972)

Table 1.1: DNA content of *Xenopus* species
Taken From Thiebaud&Fischberg 1977. *X. tropicalis* has 56 the DNA content of *X. laevis*, whereas the 2N=72 frogs have ~200% the DNA content of *X. laevis*, providing evidence for polyploidy.

	Contigs	Scaffolds
# sequences	648,787	413,763
Length of sequence (MB)	2449.7	2723.8 (10.1% gap)
N/L50	37,644/19.3KB	648/1.1MB

Table 1.2: Summary of XENLA7.1 assembly
Summary statistics of XENLA7.1 genome assembly.

Chapter 2
X. laevis transcriptomics: genome assessment and annotation

A primary goal of developmental biology is to understand how the embryo expresses specific gene sequences. The function of these molecules is often studied by gene knockdown or overexpression. Construction of probes to visualize gene expression (Gall, 1969), or constructs that allow overexpression or knockdown of specific molecules, requires knowledge of the gene sequence. The *X. laevis* community has generated a wealth of transcriptome resources in their experiments to understand the development of the frog, including sequencing 11,515 full-length cDNA clones, ~700,000 expressed sequence tags (ESTs), and over 1 billion RNA-seq reads from different developmental stages and adult tissues.

While protein-coding genes are a large focus of developmental analysis, the parsing of other sequences such as microRNAs (miRNAs) and conserved non-coding elements (CNEs) is important for elucidating the mechanisms of gene regulation. Despite the high level of importance of these sequences, they make up a small fraction, 2–4%, of vertebrate genomes (Lander, 2001). Regardless, they can serve as important markers of genome completeness, by assessing their presence and fragmentation. Conversely, the lack of presence of a cDNA sequence in our assembly, or any related tetrapod, may mark it as being an experimental contaminant that should be removed from a public database.

Transcriptome data is also used to annotate genome sequences, so that members of the research community can easily identify molecules of interest. Annotating an allopolyploid genome raises the unique problem of overcoming the similarity of homeologous sequences. Indeed, highly-similar sequences might be too similar to differentiate by standard methods (Dennis, 2012). Comparison of the polymorphism rate of protein-coding genes against the homeolog divergence is required for confidence in the mapping of small sequences to single loci in the assembly.

In addition to transcribed sequences from the species of interest, data from related species can also be useful for annotation. Related organisms with whole-genome sequence may have protein sequences which are present in our assembly, but may not be sampled in the transcriptomic data. While automated processes exist for overlapping the transcriptome data with related proteomes, assignment of gene orthology is another problem. Orthology is defined as homologous sequences between species. Paralogy is defined as homology between two sequences within a species. Gene orthology can be simple if a gene has no related paralogs, but for larger families we often need to be sure to properly distinguish between the different members of a given family aligning to a specific locus in our assembly. Determining orthology is key in assigning gene names as a resource to the community, and often need to incorporate more variables than simply conserved sequence identity, such as synteny.

Classically, synteny is defined as the physical co-localization of sequences on the same chromosome within an individual or species. More recently, genomic analysis has concentrated on the preservation of gene linkage within “blocks” of orthologous syntenic DNA sequence, referred to as “conserved synteny.” A special case of this situation is conserved collinearity in which gene order and orientation are also preserved. Here we follow current convention and use conserved synteny to refer also to conserved collinearity, which is prevalent within and between *Xenopus*. Genomic rearrangements disrupt conserved synteny between species, however the relationship between small blocks of genes is often maintained, and can be used to infer biological relationships. Stronger-than-expected shared synteny can reflect selection for functional relationships between syntenic genes, such as combinations of alleles that are advantageous when inherited together, or shared regulatory mechanisms (Duret, 2009. Zhao, 2004). In the case of allopolyploidy, selection may work to maintain conservation of synteny

between homeologous chromosomes as a protection against titrating out the dosage of important genes. Alternatively the redundancy introduced by polyploidy may accelerate the rate of disruption of synteny. Distinguishing between these models is difficult, because the rate of rearrangements between different phyla are known to be different (Zhao, 2004). Our ability to come up with a null hypothesis, and to test it, would be greatly aided by a chromosome-scale genome assembly and a wealth of epigenetic data.

The combination of shared synteny and genetic redundancy introduced by polyploidy offers a unique opportunity to study unitary pseudogenes. Pseudogenes are gene sequences which do not produce a functional protein, but can be recognized by their sequence similarity to an annotated protein. Pseudogenes can be formed either through incomplete duplication or gene decay. In diploids, pseudogenes formed by gene decay are rare, as many genes are under purifying selection in diploid organisms. Those that do form will not persist for many generations after the nonfunctionalization becomes fixed in a population, because the accumulation of additional mutations and deletions will obscure their origin as functional genes.

Following polyploidy, however, an entire genome's worth of complete gene duplicates are formed, and the redundancy allows many of them to accumulate nonfunctionalizing mutations to form unitary pseudogenes. While these sequences may be lost from the genome once the nonfunctionalization is fixed, they will be forming at such a high rate that we will expect to capture a large number of them, compared to other analyses (Zhang, 2010). Additionally, if *X. laevis* is currently in a state of losing gene sequences, we will be able to ask questions about the different stages of pseudogenization.

The wealth of *Xenopus* transcriptome data is known to contain sequences from expressed transposable elements. Transposable elements (TEs) are DNA sequences that are able to move around, or "transpose", themselves throughout the host organism's genome. Class I TEs are retrotransposons, which transpose through an RNA intermediate which is integrated back into the host genome to create a new copy. Class II TEs are DNA transposons, which require a transposase enzyme to cut and paste it into a new location. Biologists are still working to understand the full effects of transposons on their hosts, however they are silenced throughout much of an animal's life cycle, suggesting their constitutive expression can be detrimental, and is selected against.

While these molecules are of minimal interest to developmental biologists, they are sequences that should be annotated and masked in the genome assembly because they are repetitive and could introduce noise to genomic or transcriptomic analysis if left in the annotation as protein-coding gene loci. Despite being noise in protein-coding genomic experiments, TEs can serve as important molecular markers of genomic history. For example, the expansion of specific duplicate genes in primates can be tracked by the expansion of specific ALU elements (Bailey, 2003). In the rat genome paper, the repetitive content of mouse, rat, and human is compared to illustrate that a large number of the repeats of mouse or rat are specific to each organism (Gibbs, 2004). The authors hypothesize that speciation is often accompanied by a genetic bottleneck. Any repeat expansion during the time of this genetic bottleneck has a higher chance of becoming fixed in the species, and thus inherited. While there is no positive selection to keep or maintain these sequences, there is weakened negative selection to remove them. By comparing the complete repeat set between species or populations, we can understand their genomic history by looking for shared and divergent TE subfamily expansions.

The next chapter outlines the usage of external resources to assess the completeness of early genome assemblies, and to annotate the different types of genomic loci. I additionally discuss how different types of annotated sequences allow us to better understand the molecular history of the *X. laevis* genome.

2.1 Transcriptome-based assessment of genome completeness and scaffold length

Early builds of Meraculous produced fragmented assemblies with L50 < 20kb, which is comparable to the average genomic footprint of a *X. tropicalis* gene: ~22kb. I assessed early

assemblies by aligning the 11,515 full-length cDNA clones deposited in NCBI for *X. laevis* to our assembly and asked how many cDNAs were present in our assembly, and how many scaffolds each cDNA was split across (blastn, evalue=1e-10, no DUST masking). I restricted myself to full-length cDNAs for this analysis as individual ESTs could align to identical exons between homeologs, whereas the full-length sequences had an average divergence of ~94% between homeologous sequences (Figure 2.1). 11,472/11,515 (99.6%) aligned to our assemblies across their entire length, even to the earliest assemblies. Table 2.1 tracks the progression from *X. laevis* v5 to the current 7.1 build. The fragmented regions identified by this global analysis were shared with Jarrod Chapman to help improve the gap closing and scaffolding done by Meraculous. The current assembly has all 11,472 mapped cDNAs on a single scaffold.

2.1.1 Identifying cDNA contaminants in NCBI datasets

I aligned the 43 cDNAs that had no placement in the assembly to NCBI's nr database in order to assess whether they were sequences we did not assemble, or possible contaminants in the publicly available data sets. If the top hits for a given cDNA in nr were *X. tropicalis*, or other amphibian sequences, it is likely that we did not assemble the whole *X. laevis* genome, whereas the cDNA grouping with another phylum would indicate that the cDNA was a contaminant in the original experiment. All 43 cDNAs grouped with other phyla, and their top hits are in Appendix Table 1. Briefly, there are three major sources of contamination in the cDNAs, one is from a single Mouse Genome Consortium experiment that sequenced a few *Xenopus* cDNAs in a 96 well plate, and some mouse sequences appear to have contaminated the *Xenopus* well of that experiment. Additionally, some of the remaining sequences map to trypanosomes, or fungi. As trypanosomes are a known gut parasite of frogs, and chytrid is a fungus known to be prevalent in *Xenopus* laboratories, these sources of molecular contamination seem reasonable.

2.2 Repeat annotation

De novo repeat identification is an initial scan of sequence data that seeks to find the repetitive regions of the genome, and to classify these repeats. As short tandem repeats are generally 1–6 base pairs in length and are often consecutive, their identification is relatively simple. Dispersed repetitive elements, on the other hand, are more challenging to identify, due to the fact that they are longer and have often acquired mutations. However, it is important to identify these repeats as they are often found to be transposable elements.

De novo identification of repeats involves three steps: 1) find all repeats within the genome, 2) build a consensus of each family of sequences, and 3) classify these repeats. I identified new families of transposable elements using RepeatModeler (Smit, 2015). First, I detected all fragments of the frog genome coding for proteins similar to catalytic cores of transposases, reverse transcriptases, and DNA polymerases representing all known classes of TEs collected in Repbase (Jurka, 2005). The detected DNA sequences have been clustered based on their pairwise identities by using BLASTclust from the standalone NCBI BLAST package (the pairwise DNA identity threshold was equal to 80%). Each cluster has been treated as a potential family of TEs described by its consensus sequence.

The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. A library of TEs was produced by merging the identified consensus sequences with DNA sequences of *X. laevis* TEs reported previously in literature and collected in Repbase. Using RepeatModeler, we identified genomic copies of TEs similar to the library sequences. They have been clustered based on their pairwise DNA identities using BLASTclust. In each cluster, a consensus sequence was derived based on multiple alignment of the cluster sequences. After refinements of the consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously (Smit, 2015). Identified TEs are deposited in Repbase. The final set of repeats were used by RepeatMasker to mask the assembly. The previously annotated RepBase set of transposons masked ~10% of

the *X. laevis* assembly, while my *de novo* repeat set masked ~40% of the *X. laevis* assembly. Other tetrapod genomes have 40%-50% repeat density, so we are confident that our de novo set is a more complete annotation of the *X. laevis* repeats, as opposed to being too aggressive in masking.

Similar analysis was done to complete the repeat annotation of *X. tropicalis*. The previous assemblies used only the RepBase annotated repeats. A large number of the unnamed transcripts in the previous *X. tropicalis* annotation overlapped with novel transposon families identified by RepeatModeler (Figure 2.11). The removal of excess transposons from the most recent annotation is discussed more in 2.4.

2.2.1 Transposable element subfamilies identify specific sub-genomes of *X. laevis*

The diploid progenitors of *X. laevis* are extinct. If these species were extant, we could directly compare the genomes of the polyploid *X. laevis* to the diploid genomes to search for evidence of allotetraploidy or autotetraploidy (Gill, 2009). Without their genomes we must depend on studying the natural history of genomic evolution to determine between the two models. We know from wheat and soybean allopolyploidy that recombination often does not occur between homeologous chromosomes (Mayer, 2014. Gill, 2009). If there is no recombination over millions of years, the transposable element subfamilies that expanded in each of the progenitor genomes after speciation, but prior to hybridization, will show an asymmetric distribution in the sub-genomes of *X. laevis* today. By identifying these sequences in the longer homeologous scaffolds, we can use the presence of specific sub-families as diagnostic elements of whether a specific scaffolds belongs to one of the two progenitor species.

The sub-genome-specific transposon families were identified by Jarrod Chapman and Oleg Simakov. The RepeatMasker result was used to calculate the total coverage length (bp) of each repeat family on each scaffold. For each repeat family, they analyzed the density of the element in the assembly. Sub-genome-specific elements will be present in approximately half the scaffolds, whereas repeat subfamilies that expanded prior to, or after, the hybridization will have a uniform distribution across all scaffolds. The scaffolds that were mapped on a certain chromosome by BAC-FISH were collected and used as a “pseudochromosome” to calculate the approximate density of the uneven repeats on each chromosome. The density was compared between homoeologous pseudochromosomes (e.g. 1-Long vs. 1-Short) to confirm specificity of the repeats to one of the homoeologous chromosome.

Repeat families confirmed to be specific to either L-pseudochromosomes (type A) or S-pseudochromosomes (type B) were supposed to be partial fragments of sub-genome specific transposons, thus they were used to identify the full-length transposon sequences. Each consensus sequence of type A or type B repeat was used as a query for a BLAST search. HSP (high-scoring segment pair) sequences were collected with their flanking sequences and they were compared by multiple alignment to identify the range where these sequences show homology one another. This homologous range was supposed to correspond to the full length of a type A or type B transposon. All type A or type B transposons were found to belong to the DNA (class II) transposon, thus they were classified by the target site and the terminal inverted repeat (TIR) sequences. The density of the sub-genome-specific transposon families across the different *X. laevis* chromosomes is shown in Figure 2.2. Masanori Taira’s group performed TE-FISH with the most dense type B element to produce the chromsome hybridization picture in Figure 2.3 that shows the TpB_Harb subfamily specific targets the shorter chromosomes of each pair, which correspond to our “B” sub-genome. Our “A” sub-genome corresponds to the “Long” chromosome of each pair, while the “B” sub-genome corresponds to the “Short” chromosome for each pair. I will refer to the “A” sub-genome as “L” from now on, and “B” as “S”.

This sequence of experiments provides strong evidence for an allotetraploidy origin to the *X. laevis* WGD. Under an autotetraploidy model there would be no explanation for asymmetric TE expansion between sub-genomes. Also recombination between the homeologous chromosomes would ensure the signal of the TEs would be lost millions of years after the hybridization.

2.3 De novo annotation of protein-coding genes

The genome sequence has been complemented by ~697,000 *Xenopus laevis* EST sequences from a diverse set of cDNA libraries and more than 1 billion RNAseq reads that sample a useful range of developmental stages and adult organs and tissues, summarized in Appendix Tables 2-3. We relied on raw EST data for gene annotation,rather than the *X. laevis* UniGene clusters, because the clusters were formed without genomic data, and are known to contain misjoins that splice together homeologous sequences. Having a complete genome sequence in hand that separates homeologous sequences avoids this artifact by allowing the ESTs and RNAseq reads to map to their appropriate loci. The ESTs provide a rich resource for the characterization of *X. laevis* genes, and since many libraries were constructed in expression-ready vectors, they also provide an excellent resource for functional experiments with individual clones, or for screening by expression cloning.

Most transcripts were generated from the J strain, but some come from outbred populations. The degree of polymorphism between these libraries is much lower than that between homeologous genes (0.03% vs 6% respectively Figure 2.4), allowing us to confidently map ESTs from various populations and outbred individuals to the assembly. Clustering analysis has enabled the prediction of full-length cDNA clones, their reorganization into non-redundant collections, and their input into various large scale full-insert sequencing programs. These sequencing programs, as well as many smaller efforts, have resulted in the deposition of 697,015 mRNA sequences in GenBank, representing 13,141 genes (data from NCBI-UniGene, *Xenopus laevis* build 94; assuming one UniGene cluster equals one gene). What proportion, however, of these full-insert mRNA sequences contain the full open reading frame is not clear. EST data and full-length sequences are also available in the *Xenopus* Gene Collection.

Using homology-based gene prediction methods and the wealth of *Xenopus* ESTs and cDNAs resources we identified 54,142 candidate protein-coding loci and 72,472 transcripts. This overestimates the actual gene count, partly due to genes extending over multiple small scaffolds, and partly due to our generous inclusion of single-exon gene candidates known to be over-represented in *Xenopus*, likely due to unannotated repetitive elements (Hellsten, 2007). Transcript assemblies were made with PASA (Haas, 2008) from *X. laevis* ESTs/cDNAs using *X. laevis* genome assembly Xenla7.1 as reference and criteria of 98% identity and 50% coverage (*X. laevis* PASA). ESTs/cDNAs were downloaded from NCBI. *X. laevis* genome sequences were repeat-masked by RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org). Both sets of transcript assemblies were aligned to the *X. laevis* repeat-masked genome using blat; *X.tropicalis*, human, mouse, and chicken (ENSEMBL release 77) peptides were aligned using NCBI BLASTX. Putative gene loci were determined based on blat alignments and BLASTX alignments with possible extension of 500 bp at either end. Best ORFs (Open Reading Frames) for transcript assemblies were obtained by studying three-frame translation homology to human peptides (-e 1E-5) or longest ORFs were kept if no homology was found and if the ORF is at least 150 bp long. *X. tropicalis*, human, mouse, and chicken peptides, and transcript assembly ORFs at a given locus were used as protein templates for both GenomeScan (Burge, 1997) and Fgenesh+ (Salamov, 2000) gene predictions along with locus location as range constraint. Gene predictions were fed into *X. laevis* PASA for two rounds of annotation comparison and update. Gene models from *X. laevis* PASA were fed into *X. laevis*

PASA for another 2 rounds of annotation comparison and update. Gene model transcripts are validated if PASA has improved and validated transcripts based on ESTs/cDNA alignments.

Peptides of gene models from *X. laevis* PASA were aligned to *X. tropicalis*, mouse, human, and chicken peptides for homology and synteny analysis. Gene models were discarded if their coding sequence (CDS) overlap with repeats exceeded 20%. After filtering for repeats, all transcripts in a locus were kept if they were validated by PASA runs, whereasonly one transcript (longest CDS length) was kept if it has ESTs/cDNA, homology support, or syntenic orthology to another tetrapod. All candidate loci are supported by EST evidence or peptide homology to human or chicken, with 89.7% being supported over at least 80% of the CDS length by either ESTs and/or sequence homology, and 81% being supported over at least 80% of the CDS length by ESTs/cDNAs alone.

We believe the inferred 54,142 candidate loci to be an overestimate of the true gene count for two main reasons. First, only 25,152 show confident orthology to an *X. tropicalis* protein-coding locus (details below). While the remaining ~19k genes may contain true protein-coding genes whose orthology is difficult to determine, we also have an excess of single-exon genes, also seen in the *X. tropicalis* annotation, which may be enriched for transposable elements (Hellsten, 2007).

2.4 Determining orthology with *X. tropicalis*

To identify orthologs of *X. laevis* genes in *X. tropicalis* we used the BLASTP algorithm from the BLAST+ package with a Smith-Waterman refinement and an e-value cutoff of 1e-10. We accepted alignments $\geq$ 80% identity and $\geq$ 50% length of the *X. laevis* query. The highest % identity alignment within 90% of the maximum BLAST bit score is chosen as the *X. tropicalis* ortholog to a given *X. laevis* protein. We only accept *X. tropicalis* loci with 1 or 2 X. laevis (co)-orthologs (also called homeologs) by these criteria. Finally, we only accept *X. laevis* homeologs whose synteny and sub-genome identity agree with the BAC FISH map, resulting in ~16,050/22,718 (72.9%) *X. tropicalis* protein-coding loci available for analysis.

The (>204) *X. tropicalis* loci with $\geq$ 3 loci aligning are separated into 3 classes. (1) The earlier annotations masked with RepBase contained a number of transposon sequences whose homologous subfamilies were not masked in *X. laevis*. This class is defined by not having a clear syntenic ortholog, and the homologs align to many different sequences across their entire length. (2) *X. laevis* loci where one or both genes are fragmented compared to their *X. tropicalis* ortholog. We are working with the *Xenopus* community to properly annotate these loci. (3) *X. laevis* loci that have had a tandem duplication following the speciation from the *X. tropicalis* ancestor. *Chordin* is an experimentally-validated (Atsushi Suzuki, personal communication) example of this type. While it would be interesting to study all of the tandem duplications of *X. laevis*, we must first classify the first two groups to be sure that we have a confident list for the third.

Using this shared orthology between *X. laevis* sub-genomes we can identify gene pairs resulting from the recent allotetraploidy event. 9,102/16,050 (56.7%) of the protein-coding gene loci still retain both copies in *X. laevis*. This corresponds with the high end of previous estimations based on EST data (Hellsten, 2007) and is much larger than the retention of duplicates from the teleost duplication (3–4%, Jaillon, 2004) and vertebrate-stem duplications (1–2%, Putnam, 2008).

microRNA (miRNA) precursor sequences were identified by aligning experimentally-confirmed *X. tropicalis* miRNA precursor-sequences to the *X. laevis* genome via BLASTN with e-value cutoff 1e-10. The highest % identity of each sequence was chosen as the ortholog. When multiple members of a miRNA family aligned to a single *X. laevis* locus as similar %

identities, synteny of flanking protein-coding genes was considered to determine orthology. With the exception of mir-427, a miRNA known to occur in tandem arrays that are difficult to assemble (Lund, 2009), 156/180 (85%) miRNA gene precursor sequences are retained in both homeologs. The high degree of similarity between their homeologs makes it difficult to confirm expression of both copies through small-RNA sequencing, which can only isolate the precursor sequence. While the primary sequences between miRNA homeologs are divergent enough to distinguish reads between the two copies, they have a short half-life, making them difficult to sequence across their length. RNA-seq data may obtain small fragments of the poly-adenylated primary sequence present in each stage. We queried our RNAseq data for alignments +/- 1kb of the intergenic precursor-miRNA sequence to confirm expression of primary sequence of both homeologs. All duplicated intergenic miRNA pairs show reads aligning to the flanking DNA of both copies; this rate is significantly higher than randomly chosen 2.1 kb segments of the unannotated genome (Figure 2.7). We cannot confirm expression of homeologous intronic miRNAs because it is difficult to distinguish their expression from that of their host genes.

In addition, we found that 557/557 of pan-vertebrate conserved non-coding elements (pvCNEs) (Lee, 2011) are present in the assembly, and 533/557(95.6%) are still present in two copies. We aligned the published human sequences to the elephant shark genome by the same megablast parameters in the original paper. The elephant shark sequences were then used to identify the pvCNEs in different tetrapods. Non-*Xenopus* tetrapod genomes are from Ensembl build 77.

2.5 Using synteny of protein-coding alignments to form chromosome-scaled pseudomolecules

Remarkably, with the exception of the 9/10 fusion, *X. laevis* and *X. tropicalis* chromosomes have basically maintained their integrity without inter-chromosome exchanges since their divergence ~50 Mya. This is consistent with broader stability of amphibian chromosomes (although other frogs have 2N=22 and so other events have happened), but in contrast to, for example, mammals, which typically show dozens of inter-chromosome rearrangements. The extensive collinearity between evolutionarily homologous *X. laevis* L and *X. tropicalis* chromosomes indicates that these represent the ancestral chromosome organization. In contrast, the S sub-genome shows extensive intra-chromosomal rearrangements, evident at the chromosome-scale in Figure 2.6, but also found in shorter rearrangements/inversions. S also shows extensive deletions.

This experimental validation of the collinearity between *Xenopus* chromosomes allowed me to construct chromosome-scale pseudomolecules from scaffolds by studying the conserved synteny of protein-coding orthologs (identified above) between the *Xenopus* species. The presence of diagnostic transposons on the larger scaffolds allows me to be sure that I am not creating chimeric chromosomes between the two sub-genomes of *X. laevis*. Working with super-scaffolds generated by the HiRise algorithm (NH Putnam personal communication, Putnam, 2015) from HiC data (generously shared by Ian Quigley), I wrote an algorithm to link scaffolds, assuming conserved synteny with the *X. tropicalis* chromosomes. For the areas where synteny was disrupted, and not captured within an assembled scaffold (example in Figure 2.7), I followed the following protocol.

I used MCScanX (Wang, 2012) to identify collinear blocks of 3 interrupted genes between the *X. laevis* L, *X. laevis* S, and *X. tropicalis* genomes in the ortholog list generated in 2.4. I restricted myself to these blocks to be certain that the units of synteny would not be subject to the noise of individual elements transposing in the genome. Synteny maps for each L and S sub-genome compared to the full-length *X. tropicalis* chromosome were compared to BAC-FISH maps (Figure 2.6) to recapitulate any breaks in the conserved synteny, specifically on *X. laevis* chromosomes 3S and 8S. Scaffolding was performed on the v7.5 super-scaffolds to

form *X. laevis* v8 (summarized in Table 2.2). Figure 2.8 shows the L/S synteny maps for chromosomes 2 and 8. When compared to the BAC-FISH maps in 2.6, it is clear our assembly captures the experimentally-validated synteny of the BAC-FISH experiments.

2.6 Pseudogene identification and age calculation

We utilized synteny to restrict ourselves to studying only those unitary pseudogenes resulting from the allotetraploidy event (Figure 2.9). Briefly, we searched for syntenic triplets of genes between *X. tropicalis* and one *X. laevis* genome, where the second sub-genome was missing the middle gene (defined as a 2-1-2 pattern). We find 1,277 genomic loci fit this "2-1-2" pattern, and 745/12,77 (58.3%) have a unitary pseudogene sequence found by Exonerate (Slater, 2005). 326/745 (43.7%) loci contain at least one premature stop or frameshift mutation and are used for our analysis. There is no difference in the rate at which we find pseudogenes between the chromosomes.

Prior to comparing the rate of sequence change in the pseudogene to the extant gene sequence, we removed frameshift mutations that would increase error in our estimates of substitution rates. Additionally we only considered codons that contain at least one shared site to avoid saturating our measurements. To estimate the nonfunctionalization time (TN) of a unitary pseudogene, we adapted the method devised by Chou et al. (Chou, 2002). It assumes that non-synonymous mutations in the pseudogene accumulate at the same rate as the extant gene until nonfunctionalization; thereafter, mutations at both synonymous and non-synonymous sites accumulate at the synonymous mutation rate. Sequences homeologous and orthologous to the *X. laevis* pseudogene from *X. laevis* and *X. tropicalis* are used in the calculation, as the quantification of lineage-specific mutation rates at synonymous and non-synonymous sites remote from the inactivating deletion provides the information necessary for the calculation. Given this assumption, the following equality holds: $\bar{\omega} \cdot r_{S1} \cdot (T - T_N) + r_{S1} \cdot T_N = K_{A1}$ in which T is the time since the last common ancestor of the *X. laevis* progenitors (~40MYA, Hellsten 2007), T_N is the time since the unitary pseudogene inactivation to be estimated, $r_{S1} = K_{S1}/T$ is the synonymous substitution rate in the *Xenopus* lineage, $\bar{\omega}$ is the average K_A/K_S ratio in the *Xenopus* lineage, and K_{A1} is the nonsynonymous substitutions per nonsynonymous site in the pseudogene. Rearrange the equation above, we have: $T_N = T \cdot \frac{\omega_1 - \bar{\omega}}{1 - \bar{\omega}}$ in which ω_1 is the K_A/K_S ratio in the pseudogene. The distribution of pseudogene ages is in Figure 2.10.

The onset of pseudogene formation 20–30 MYA correlates with the distributions of sub-genome-specific transposon ages generated by Jarrod Chapman and Oleg Simakov, indicating that the hybridization between the *X. laevis* progenitor species occurred between 20–30 MYA, approximately 10–20 MY after their speciation. Understanding the timing of hybridization naturally raises questions of how the different sub-genomes have evolved, both structurally and functionally, since. These topics are discussed in the following chapters.

Table 2.1: Results of transcriptome scaffolding on early assemblies of *X. laevis*

	L5	L6	L6.1	L7	L7.1
Single scaffold	5,438/10,965 (49.59%)	5,834/10,935 (53.35%)	9,954/11,472 (86.7%)	10,581/11,472 (92.23%)	11,472/11,472 (100%)
#connections/transcript	1175/989	956/759	447/406	201/196	0/0
Average # scaffolds/cDNA	4	3	1.1	1.1	1

Table 2.2 : *X. laevis* v8 assembly summary

Tropicalis chromosome	Length	# orthologs	Repeat density (%)	L assembled length	L mapped length	L % assembled	L % trop	L # orthologs	L % retention
1	1.86E+08	1916	38.4	1.8E+08	1.8E+08	98.8%	101.2%	1601	83.5%
2	1.64E+08	1677	40.1	1.4E+08	1.5E+08	92.2%	94.2%	1356	80.8%
3	1.35E+08	1652	38.1	1.2E+08	1.2E+08	99.9%	91.3%	1266	76.6%
4	1.29E+08	1483	39.1	1.1E+08	1.1E+08	98.4%	88.1%	1284	86.5%
5	1.41E+08	1254	39.9	1.4E+08	1.4E+08	99.7%	101.0%	990	78.9%
6	1.28E+08	967	41.4	1.1E+08	1.3E+08	83.3%	105.6%	799	82.6%
7	1.08E+08	1007	40.1	1.1E+08	1.1E+08	99.8%	100.1%	755	74.9%
8	1.16E+08	1434	37.5	1.1E+08	1.1E+08	98.6%	93.2%	1115	77.7%
9	7.79E+07	821	41.8	1.1E+08	1.1E+08	100.0%	92.7%	697	84.9%
10	3.86E+07	593	39.8					479	80.7%
Total	1.22E+09	12804	39.7	1.1E+09	1.1E+09	96.5%	96.6%	10342	80.7%

Tropicalis chromosome	S assembled length	S mapped length	S fraction assembled	S frac trop	S frac L	S # orthologs	S % retention	Differential Retention pvalue
1	1.67E+08	1.82E+08	91.5%	97.6%	96.4%	1304	68.1%	4.20E-04
2	1.23E+08	1.36E+08	90.7%	82.8%	87.9%	1177	70.1%	0.019
3	1.10E+08	1.10E+08	99.7%	81.8%	89.5%	1090	65.9%	0.014
4	1.03E+08	1.03E+08	100.0%	79.2%	89.9%	1020	68.7%	1.80E-04
5	1.20E+08	1.20E+08	100.0%	84.8%	83.9%	783	62.4%	3.60E-04
6	1.05E+08	1.14E+08	92.6%	88.8%	84.1%	627	64.8%	5.19E-04
7	8.09E+07	9.17E+07	88.2%	84.6%	84.5%	651	64.6%	0.033
8	8.11E+07	8.28E+07	98.1%	71.1%	76.2%	819	57.1%	1.59E-06
9	9.42E+07	9.42E+07	100.0%	80.8%	87.2%	518	63.1%	0.0011
10						453	76.3%	0.42
Total	9.84E+08	1.03E+09	95.2%	84.3%	87.2%	8442	65.9%	1.05E-22

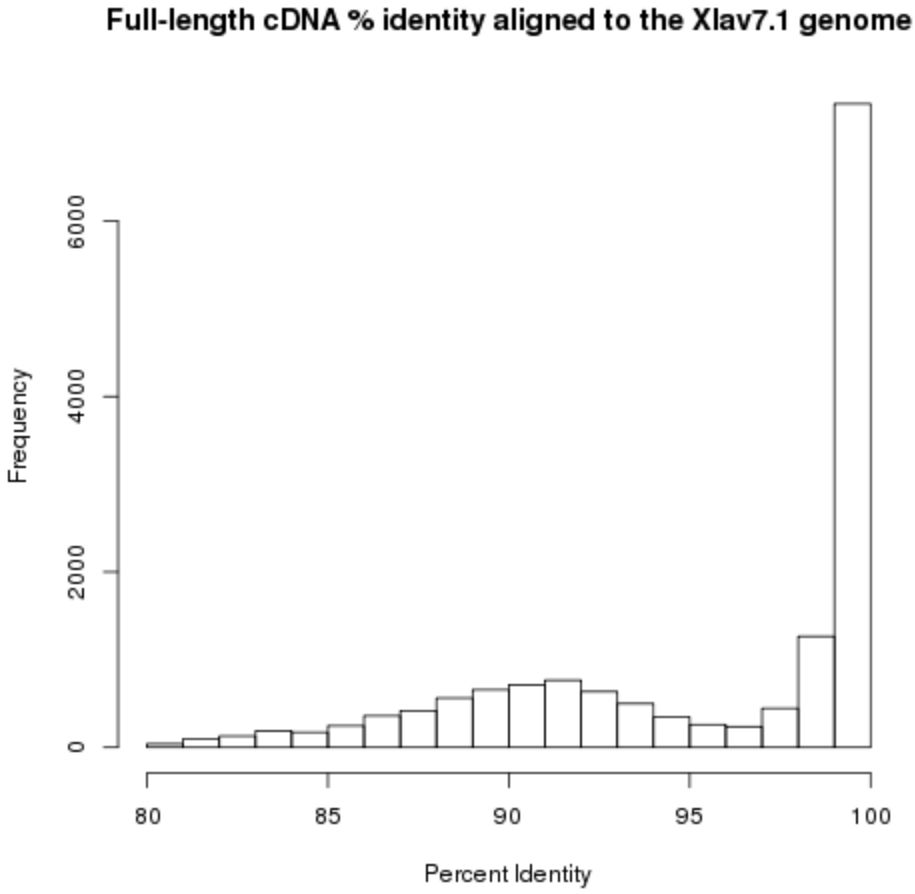


Figure 2.1: % identity distribution of *X. laevis* cDNAs aligned to the assembly

NCBI *X. laevis* cDNAs were aligned to the v7.1 assembly via blastn. The peak between 98-100% comprises cDNAs aligning to themselves in the assembly, the secondary peak comprises cDNAs aligning to their homeologous sequence in the assembly.

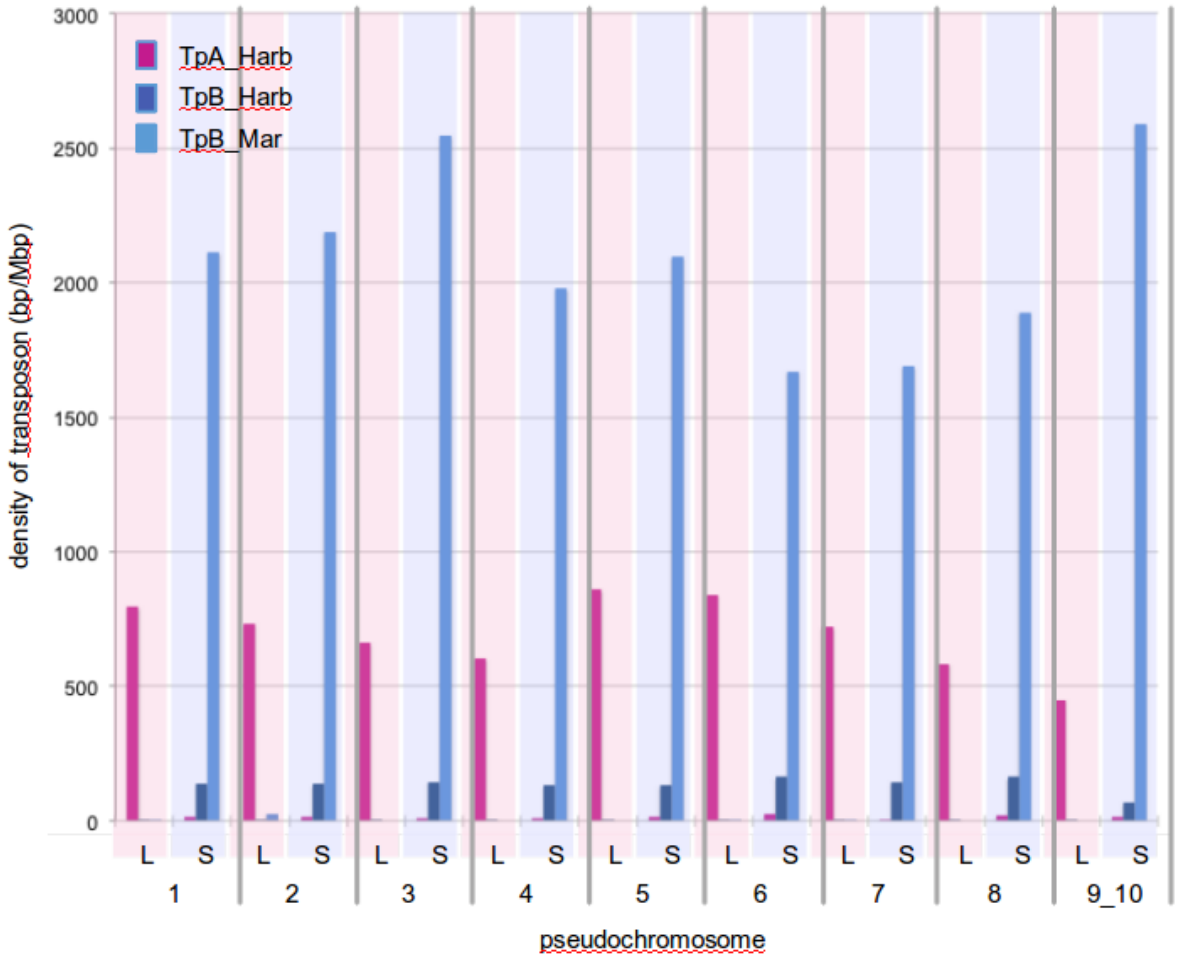


Figure 2.2: Transposable element density across chromosomes in *X. laevis* v7.1
The chromosome-level density of the consensus sequences of the sub-genome-specific transposons. The Harbinger transposon subfamilies are of considerably lower density the the TpB_Mariner class. (Masanori Taira, personal communication)

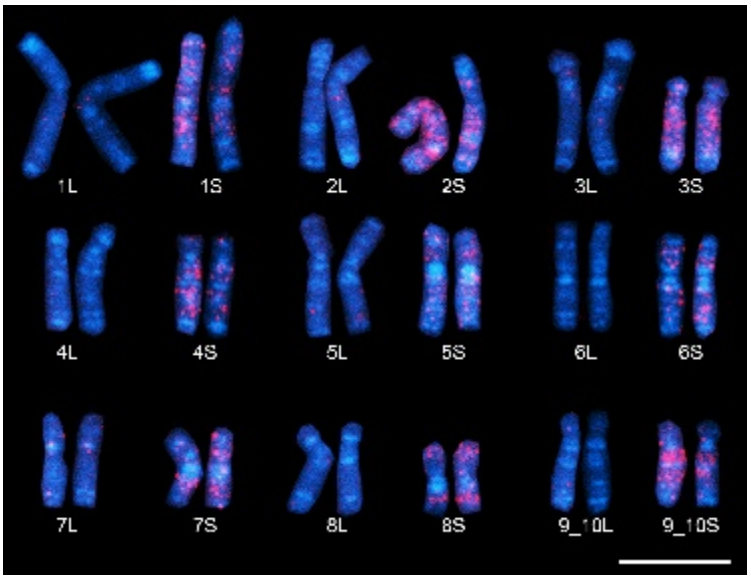


Figure 2.3: Transposable element-FISH with the TpB Mariner probe (by Taira group)
Transposable element (TpB Mariner) fluorescent in situ hybridization of *X. laevis* chromosomes preps by the Taira group. The presence of probe only on the shorter chromosome set of each pair is indicative that the TpB Mariner subfamily is specific to the S sub-genome. (Masanori Taira, personal communication).

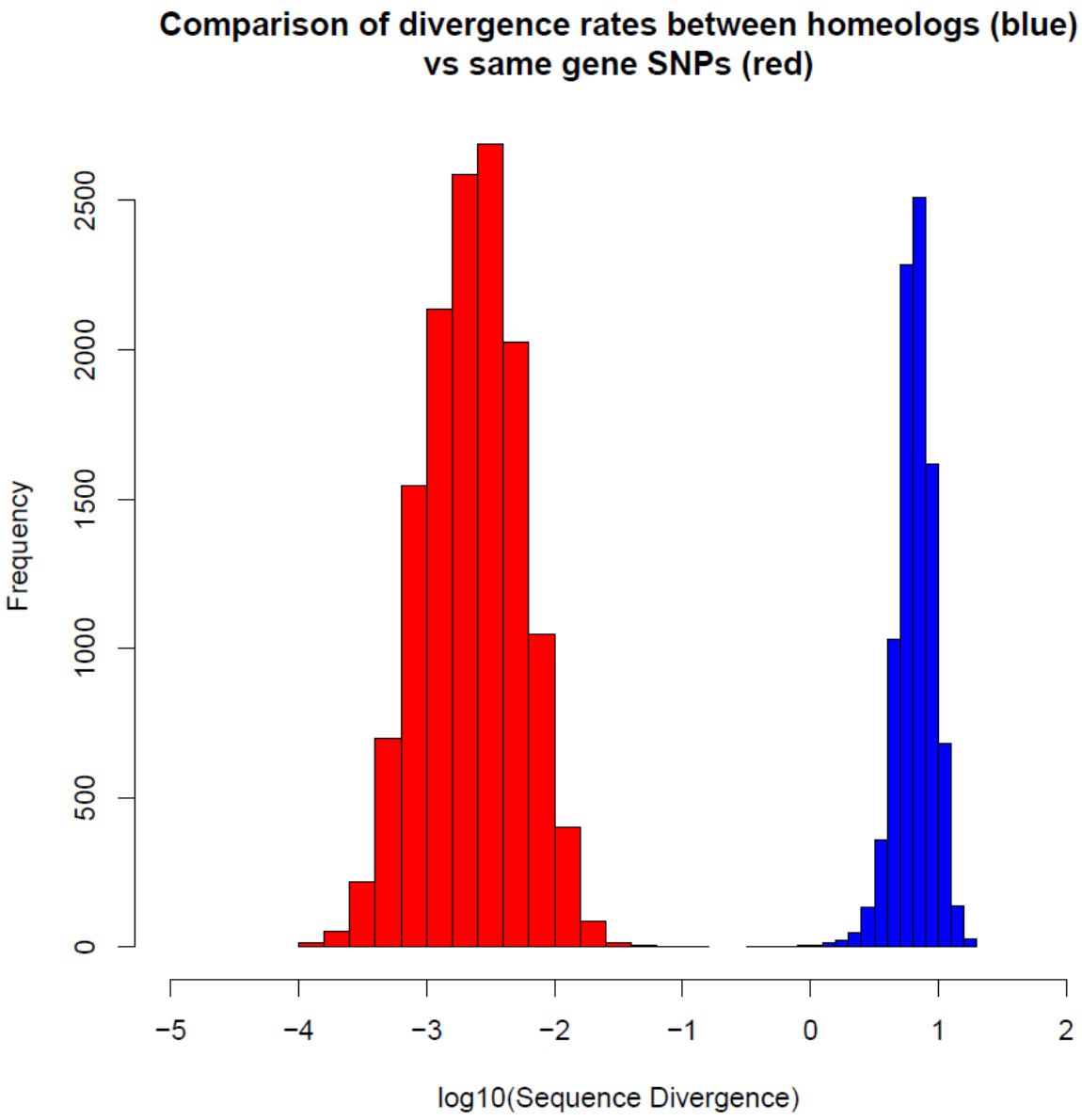


Figure 2.4: Homeolog divergence vs SNP rate
Percent divergence in cDNA sequence between homeologs is shown in blue, percent divergence between populations is shown in red. The two distributions do not overlap, indicating that the most rapidly polymorphic gene in *X. laevis* is more similar than the most similar homoeologs.

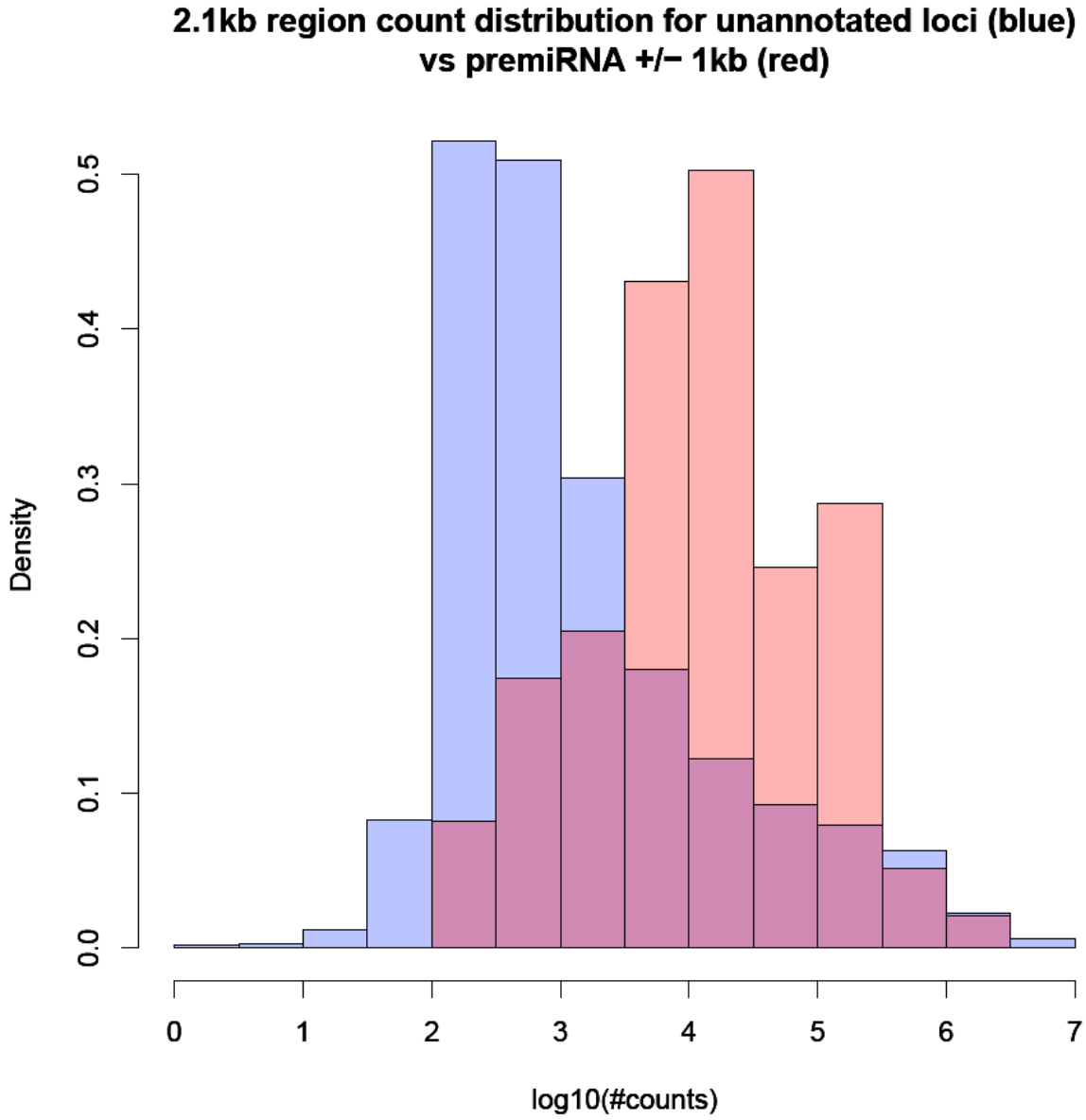


Figure 2.5: miRNA and intergenic sequence expression
The number of reads aligning +/- 1kb of precursor miRNA loci (red) was compared to the read count for 10,000 random unannotated 2.1 kb regions of the genome (blue). All 83 homeologous, intergenic miRNA pairs showed alignment within their regions, as opposed to 4,127/10,000 (41.27%) of the randomly chosen intergenic sequences. The putative primary-miRNA loci have a higher read count than the randomly chosen regions as well.

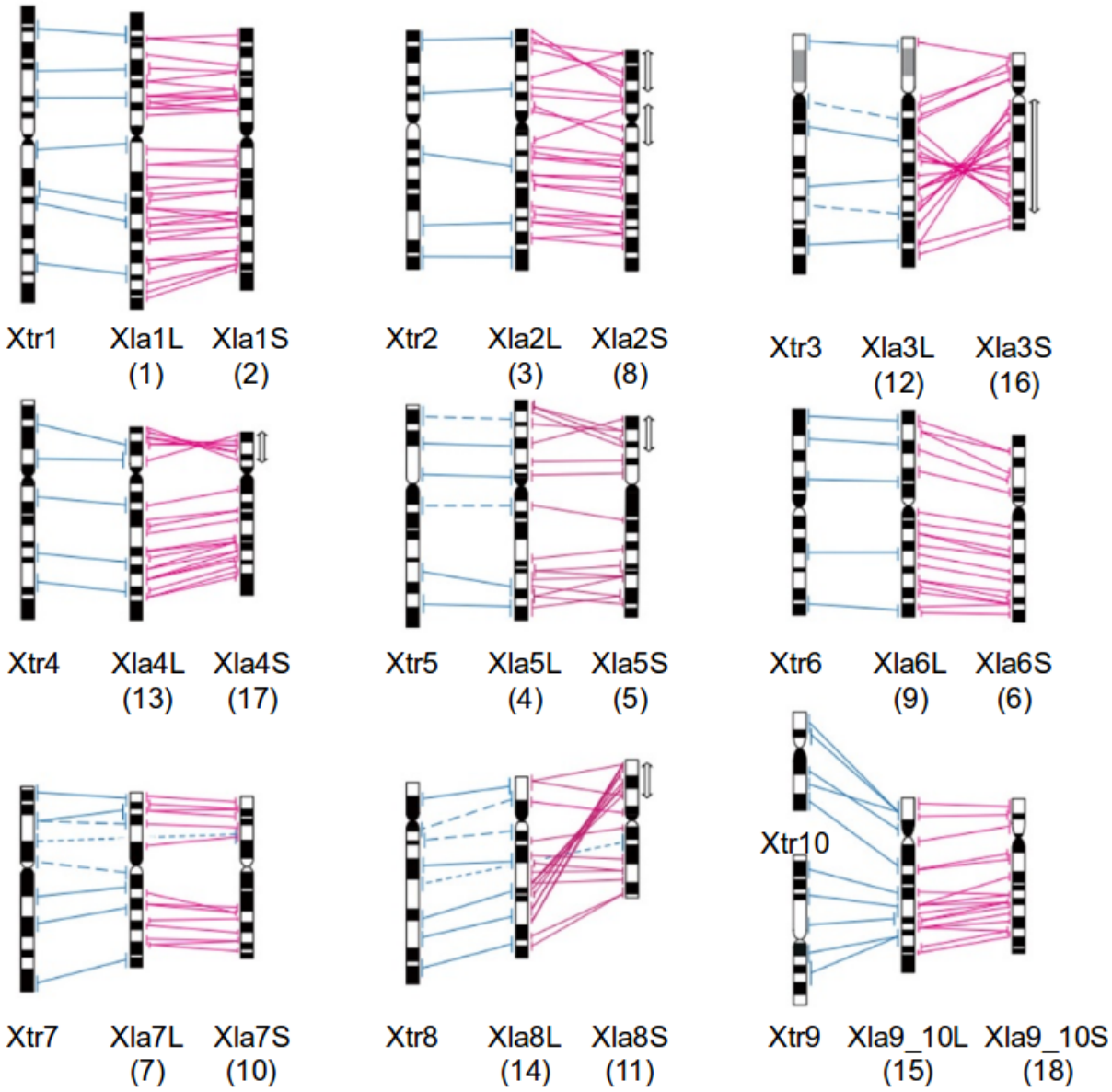
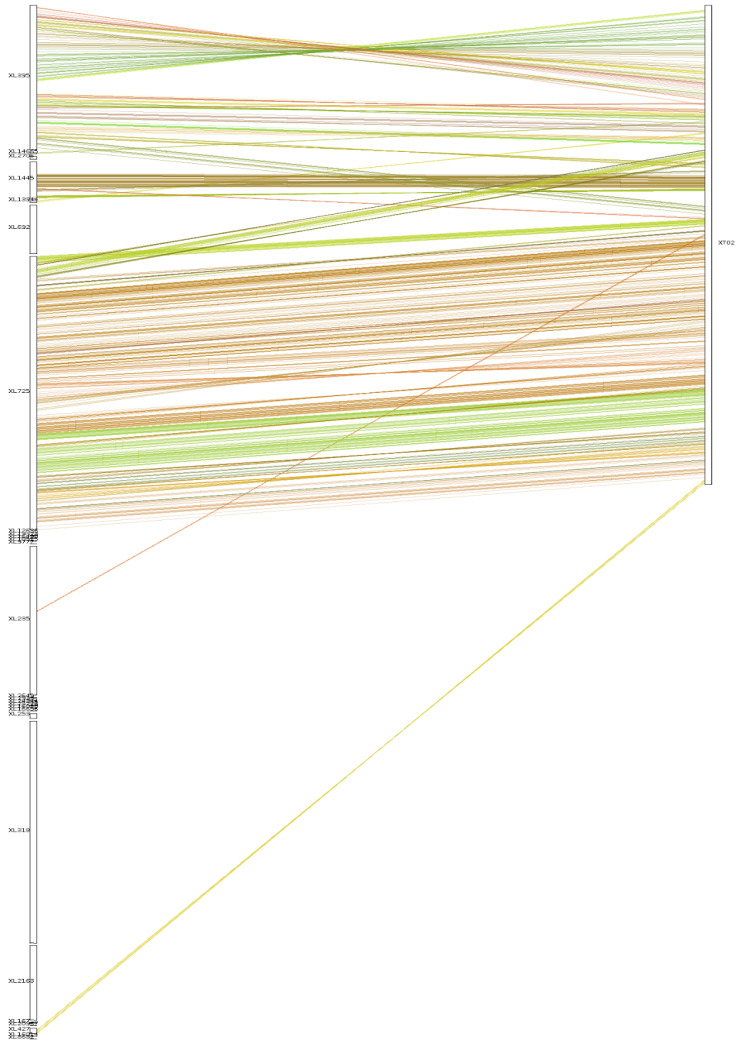
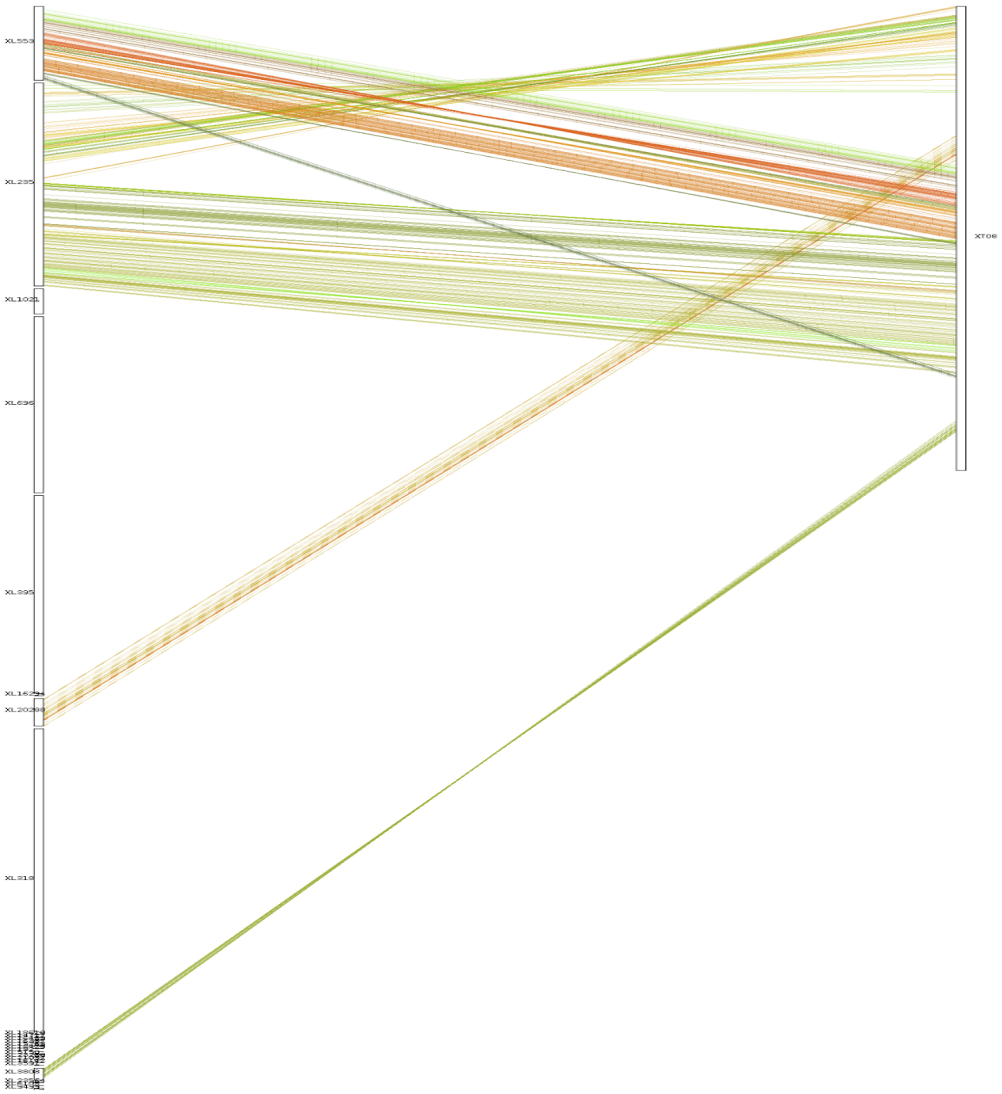


Figure 2.6: cDNA and BAC-FISH of *Xenopus* chromosomes (provided by Taira group) cDNA FISH was performed for known two copy genes of *X. laevis* for both Xenopus species (blue lines). Denser BAC in situs were performed for the *X. laevis* chromosomes only (red lines). (M. Taira, personal communication)

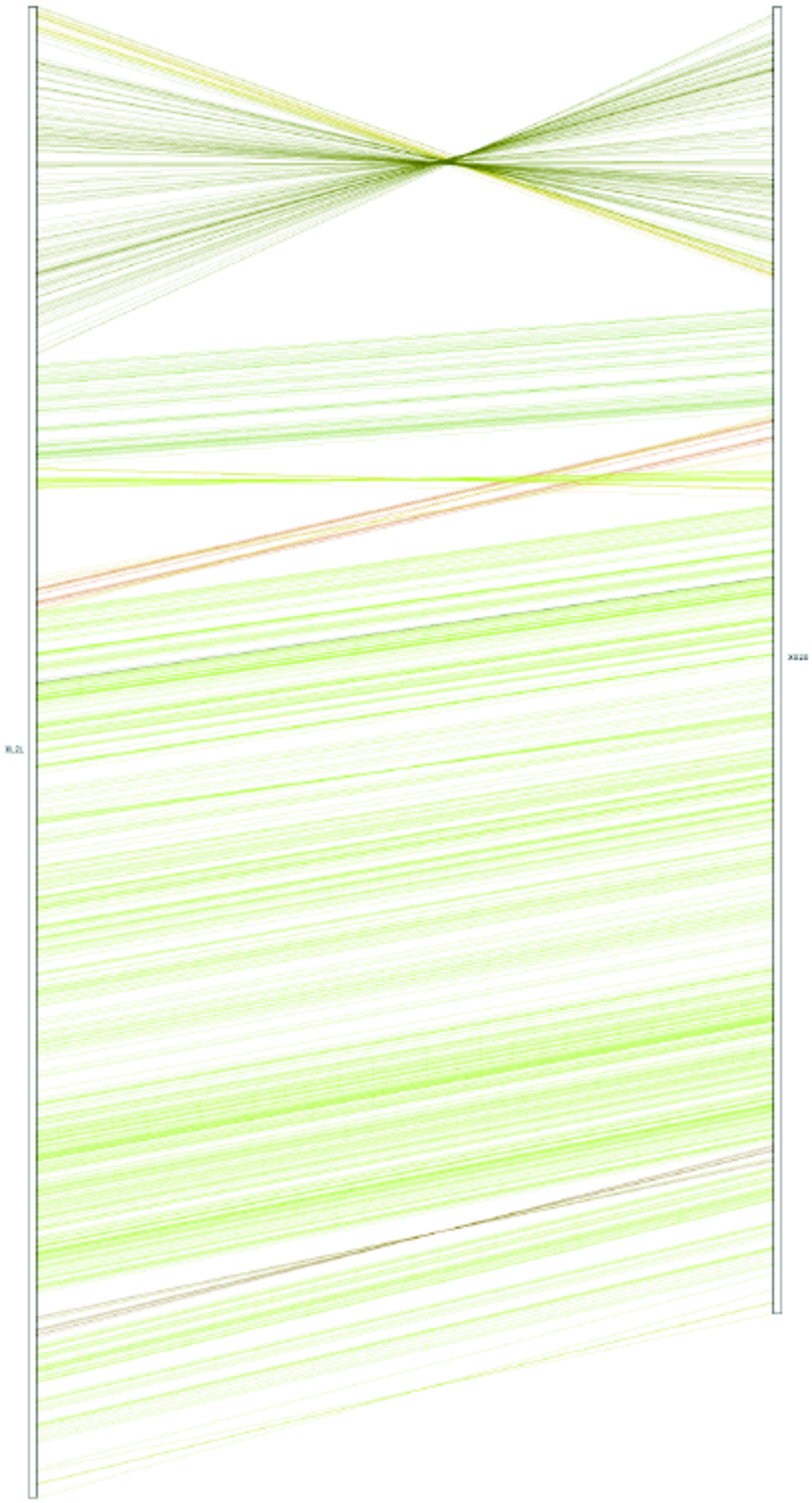
A.





B.

Figure 2.7: Genomic rearrangements captured by early *X. laevis* assemblies
Examples of early synteny maps showing rearrangements on *X. laevis* chromosomes 2S and 8S. The HiRise superscaffolds are shown on the left, *X. tropicalis* chromosomes on the right. **A.** 2S synteny, the inversion in the p arm of chromosome 2 is captured by the HiRise scaffold, no special scaffolding necessary. **B.** 8S synteny, the rearrangements of the 8S chromosome were not captured by the HiRise super scaffolds. These scaffolds were manually ordered/oriented to recapitulate the rearrangements identified by BAC-FISH in Figure 2.6. Visualized using MCScanX (Wang, 2012)



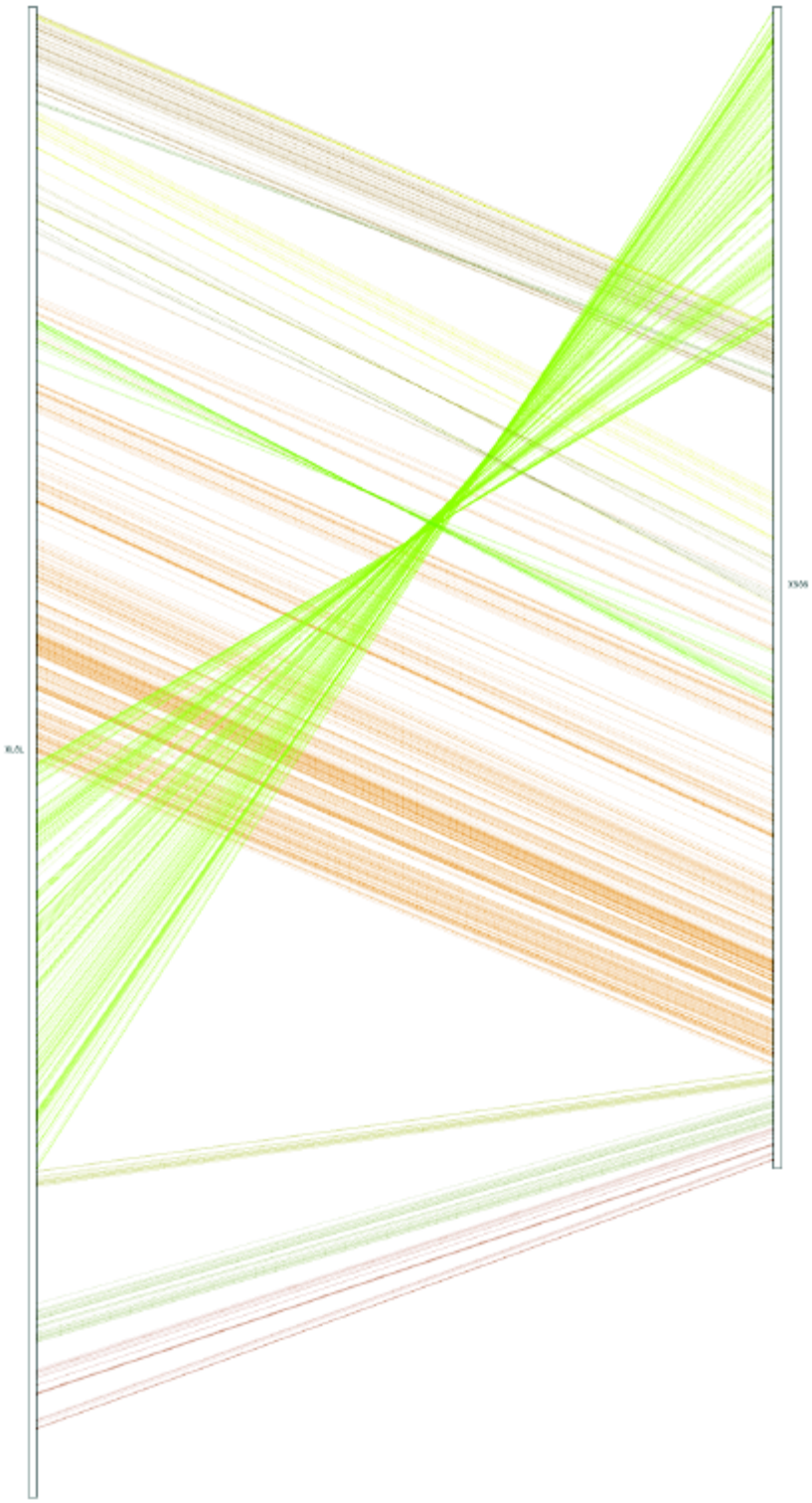


Figure 2.8: *X. laevis* v8 synteny maps
Visualizations of chromosome-scaled synteny between L and S chromosomes of *X. laevis*. The BAC-FISH identified rearrangements seen in Figure 2.6 are recapitulated by protein-coding gene synteny. **A.** Chromosome 2L (left) compared to chromosome 2S (right). **B.** Chromosome 8L (left) compared to Chromosome 8S (right)

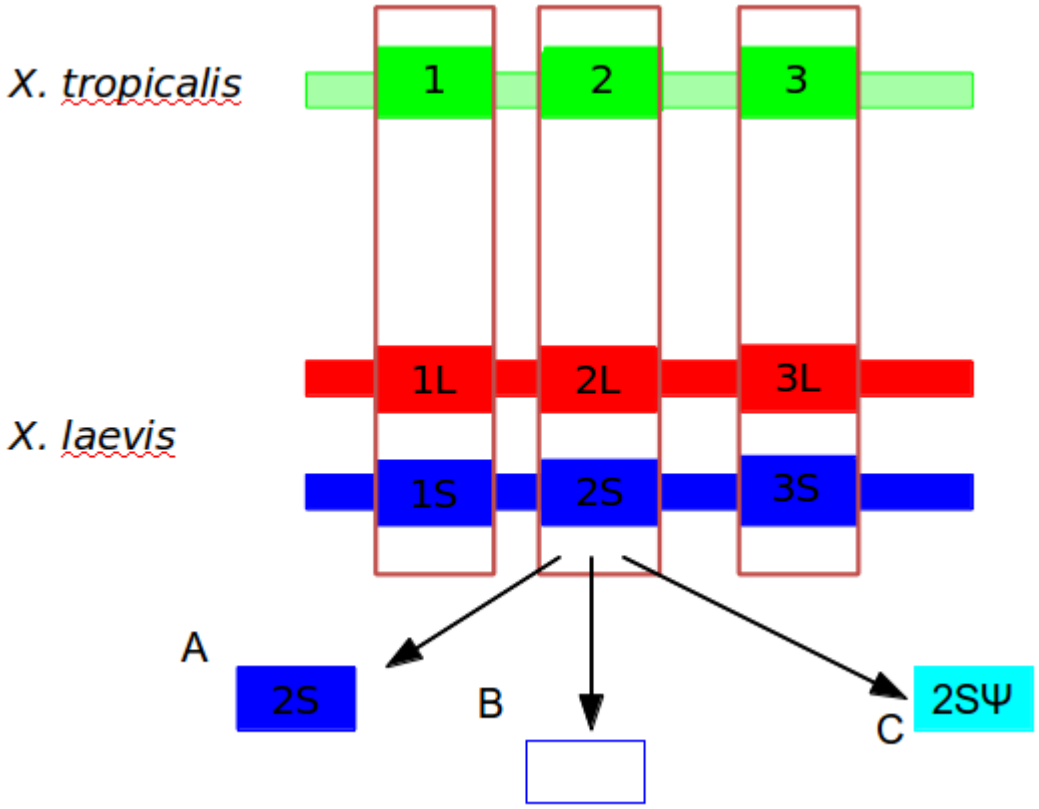


Figure 2.9: Pseudogene identification
Syntenic triplets of genes between *X. tropicalis* and both sub-genomes of *X. laevis*. The 2nd gene of either *X. laevis* copy has 3 possible fates: (A) The gene is retained in both sub-genomes, (B) The gene is deleted from one of the 2 genomes, and there is no remnant, or (C) The gene accumulates the mutations necessary to be nonfunctionalized, but the gene sequence itself is left to mutate in place as a unitary pseudogene.

Using this synteny-based algorithm allows us to be confident that we are measuring a unitary pseudogene decaying following the hybridization event.

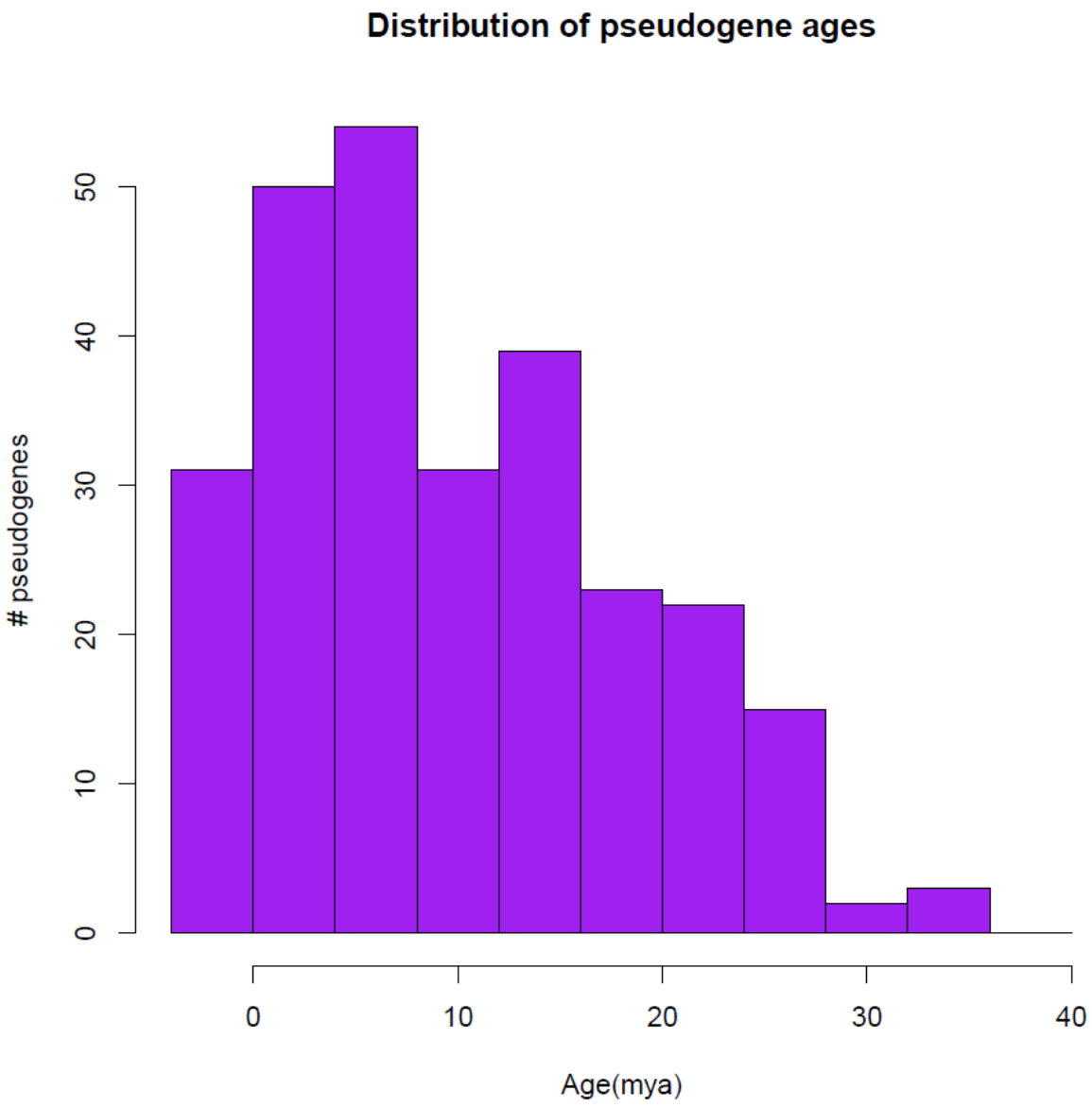


Figure 2.10: Distribution of pseudogene ages
Calculation of pseudogene ages is discussed in the text. Pseudogenes of negative age are expected to be newly formed pseudogenes whose extant homeolog was rapidly evolving prior to the nonfunctionalization of the pseudogene.

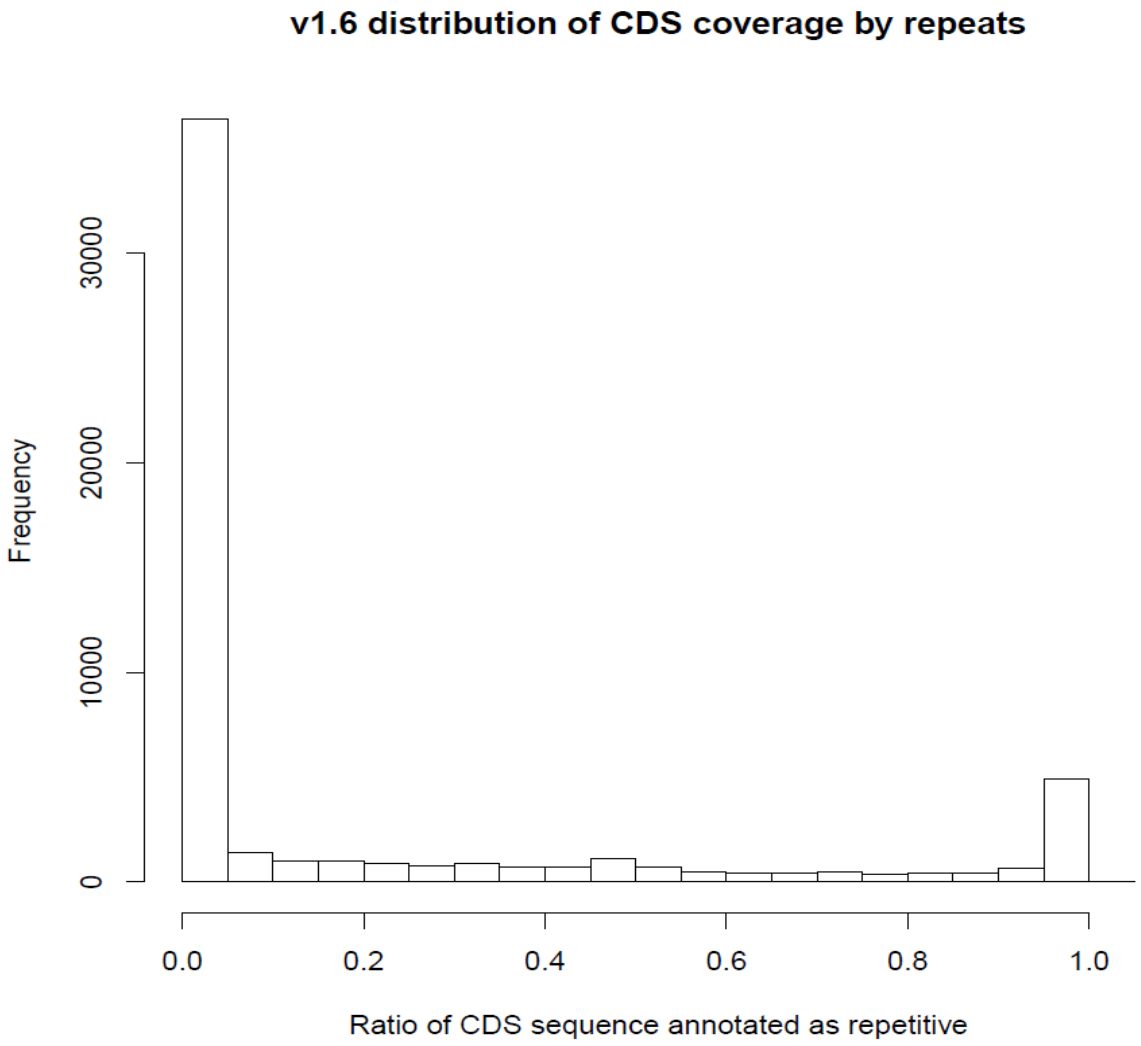


Figure 2.11: Coverage of previous transcripts by *de novo* RepeatModeler-identified transposons
Locations of novel transposons identified by RepeatModeler were compared to previous annotations using bedtools. The fraction of the CDS sequence of the previous annotation covered by newly identified transposons is shown on the x-axis. The large number of transcripts at 0 and 1 represent protein-coding genes, and full-length transposons respectively.

Chapter 3
Structural evolution of *X. laevis* genome

Chromosomes are remarkably dynamic molecules, changing shape and structure throughout the lifespan of a single cell, or between the cells of a single organism, to allow for coordinated expression of different loci. Since it is important to replicate this process with fidelity, it is unsurprising that comparisons between closely related species have shown that both micro-collinearity (that between tightly linked loci) and macro-collinearity (that at a chromosome scale) are often maintained despite a number of chromosomal rearrangements, such as translocations, inversions, and duplications. Still, there are closely related species which change their genomic structure, even differing in the types of chromosomal rearrangements observed (Zhao, 2004). These differences in the rate of chromosomal rearrangements between phyla make it difficult to develop a null hypothesis for how we expect genome collinearity to evolve following polyploidy, but it is undeniable that genomic rearrangements have an effect on gene expression. The sequence elements that are the units of collinearity are subject to point mutations as well. All types of structural changes are subject to selection over time; the retention of specific structural differences between sub-genomes informs us of the molecular history of *Xenopus*.

Point mutations come in different types. Transitions are mutations that turn a purine into a purine (A<->G) or pyrimidine into a pyrimidine (C<->T). Transversions are mutations that turn a purine into a pyrimidine, or *vice versa* (A/G <-> C/T). Despite there being more possible transversions, transitions occur almost twice as often, due to the underlying chemistry behind them. Tracking the rate at which these changes occur between different homologous sequences allows us to assess the evolutionary plasticity of DNA sequences between species, or between chromosomes within a species.

Protein-coding gene molecular evolution has more variables to consider. Protein-coding gene DNA is transcribed into mRNA, which is then translated into protein. Triplets of nucleotides in DNA/RNA code for single amino acids in protein sequence. There are 64 possible codons and only 20 amino acids (plus a “stop” signal), meaning the genetic code is degenerate (multiple codons code for the same amino acid). This degeneracy means that point mutations to protein-coding genes, whether transitions or transversions, can be split into two primary classes: (1) those that change a nucleotide but not the resulting amino acid (synonymous mutations), or (2) those that change the nucleotide and the amino acid (nonsynonymous mutations). When normalized against the degeneracy in the genetic code (Li, 1993), the rate of synonymous mutations is known as K_S, and the rate of nonsynonymous mutations is known as K_A. The K_A of a gene is influenced by its protein identity and function, and whether that protein is under purifying selection, or can tolerate amino acid substitutions without significant loss of function. The K_S of a gene is influenced by the accessibility of its DNA to different types of mutations. The K_S is thought to represent the background mutation rate of a gene, so that if its protein product were under no purifying selection, its K_A would equal its K_S.

In addition to K_A and K_S, protein-coding gene point mutations can be classified by the type of nucleotide change (transition *vs*. transversion) and the amount of degeneracy of a position in a codon (4-fold synonymous means all changes at that position produce the same amino acid, 2-fold synonymous means 2 changes produce the same amino acid, and 0-fold synonymous means that all nucleotide changes change the amino acid sequence). Since 4-fold degenerate sites are largely free from selection (aside from codon bias, Lynn, 2002), and transversions happen at a much slower rate than transitions, 4-fold synonymous transversions (4DTv) are useful for comparing sequence change across large evolutionary distances (Hellsten, 2007).

Comparative genomics suggests that the genome architecture is not a straightforward result of continuous adaptation but rather is determined by the balance between the selection pressure (which is itself dependent on the effective population size and mutation rate), the level of recombination, and the activity of selfish elements. Although genes and, in many cases, multigene regions of genomes, possess elaborate architectures that ensure regulation of expression, these arrangements are evolutionarily dynamic and typically change substantially even on short evolutionary scales when gene sequences diverge minimally. Thus, the observed genome architectures are, mostly, products of neutral processes. The following chapter outlines the structural changes between the sub-genomes of *X. laevis*, and outlines what these largely neutral processes can tell us about the molecular history of *Xenopus*.

3.1 Chromosome differences between sub-genomes

One of the initial findings after discovering the "A" and "B" sub-genomes in 2.2 was that the A sub-genome had a longer assembled sequence length (included in Table 2.2). At the same time, Masanori Taira's group performed the cDNA-FISH and BAC-FISH shown in Figure 2.8, and found that for each pair of *X. laevis* chromosomes, one was consistently shorter than the other (called "S" for short, while the homeologous chromosome was called "L" for long). While this is not a surprising result, as chromosome length is plastic over time, we hypothesized that our "B" scaffolds were the same as their "S" scaffolds. The TE-FISH in Figure 2.3 shows that this hypothesis was correct, however it does not answer why the S sub-genome is shorter. The Taira group has shown that the S sub-genome has also undergone more large-scale rearrangements than L (Figure 2.8). In order to complement their analysis, I sought to understand whether gene loss was contributing to this chromosome length difference.

3.1.1 Gene loss increased in S

I partitioned the *X. laevis*-*X. tropicalis* orthologs identified in section 2.4 into those retained on the L sub-genome and S sub-genome; the results are included in Table 2.2. The L sub-genome retains 10,342 orthologous sequences, while S contains 8,442 orthologous sequences, a significant difference against a null model of equal gene loss (Fisher's exact, $p=1.05e^{-22}$). If we assume that the L and S sub-genomes started with an equal number of orthologs to *X. tropicalis*, the sub-genomes have retained 80% and 65% of their original gene content respectively. Interestingly the regions of *X. laevis* 9_10 chromosomes orthologous to *X. tropicalis* chromosome 10 do not fit this trend and retain genes at similar rates (80% and 76%, p=0.42), whereas those regions orthologous to Xtr-9 do show a significant difference (85% vs 63%, $p=1.1e^{-3}$). The unique evolutionary signatures of those genes orthologous to Xtr-10 are discussed more in section 3.4.

This measurement of gene loss refers to the presence or absence of a complete open reading frame (ORF) in the assembly, so does not directly contribute to understanding the difference in chromosome lengths between sub-genomes. Gene loss could occur through deletion, or through mutation and pseudogenization. In the latter case the sequence would be retained until a deletion removed it from the genome or extensive substitutions made it unrecognizable as a pseudogene. So on a long time scale, increased gene loss could contribute to a decrease in chromosome size. In order to test this model we must identify loci in both sub-genomes where we expect a deletion has occurred, and compare those deletion lengths to the change in chromosome length between L and S.

3.1.2 Deletions larger in S

The unitary pseudogenes discussed in section 2.6 offer a unique opportunity to isolate genomic deletions. In the "2-1-2" loci where no pseudogene is found, we expect a deletion must have removed the gene sequence, either before or after pseudogenization. If a pseudogene is found, or if the middle gene is retained (2-2-2 locus), we have no evidence for a deletion. By comparing the difference in intergenic DNA lengths (between the end of gene 1 of the triplet, and the beginning of gene 3) of the deletion set between sub-genomes, we can assess whether known deletions in S are larger than L. There is a large variation in intergenic sequence range

(Figure 3.1), therefore we normalize the *laevis* distances by their orthologous distances in *X. tropicalis*. The results of this analysis for the different classes are shown in Table 3.1. If the middle gene was present in both copies, either functional or as a pseudogene, the length of the sequence is close to that of *X. tropicalis* (92%-107%). When a pseudogene is not found, the length of the extant gene is again close to *X. tropicalis* but the missing gene has a shorter length than *X. tropicalis* (83% for L, 51.3% for S). Without normal distributions for our data it is difficult to discern if this difference in deletion lengths between sub-genomes is statistically significant, however we are currently working on identifying more loci to study in a similar way.

3.1.3 Gene loss by location is largely random

One possible explanation for the increased gene loss in S are longer deletions. If the deletions in the S sub-genome are longer, they would remove consecutive sets of genes more often than L deletions. The distribution of consecutive gene losses/retentions is shown in Figure 3.2. The red line represents the expected exponential decrease in number of consecutive genes deleted. Both sub-genomes appear to be experiencing consecutive gene loss randomly, with a few exceptions, an olfactory gene cluster, and the type II genes in the MHC cluster. These loci are discussed more in chapter 4.

3.2 Differing rates of protein-coding sequence change between sub-genomes

To better understand the difference between sub-genomes I also compared their rates of sequence change. An increased rate of sequence change, coupled with relaxation of purifying selection, could lead to the increased rate of gene loss we see on the S sub-genome. A possible factor that could increase the rate of sequence change on S following the hybridization is increased methylation, similar to the polyploid-induced “genomic shock” observed in recent *Arabidopsis* tetraploids (Comai, 2005). Increased methylation could lead to increased C->T transitions at sites preceding G’s, which in the wake of relaxed purifying selection could lead to increased gene loss.

To compare the rate of sequence change at orthologous sites, I chose the longest transcript from each protein-coding locus for alignments. CDS alignments between *Xenopus* homologs were done using the Dialign-TX package using default parameters on the longest ORF of each sequence (Amarendran, 2008). The CDS sequence content and evolutionary rates were calculated using the seqinR package (Charif and Lobry, 2007). The calculation of sub-genome-specific rates is explained in Figure 3.3. We used R's default two-tailed Wilcoxon-Rank sum test (Wilcoxon, 1945) to determine statistically significant differences between mutation rates. *X. tropicalis* chromosomal locations were determined by the placement of the *X. tropicalis* ortholog on the v8 map.

The K_S of the homeologous gene pairs should be uncorrelated with one another, since K_S is a measurement of sequence change that is not influenced by protein identity. In contrast, K_A is influenced by the protein identity. These definitions predict that while the K_S of homeologous genes should be uncorrelated, the K_A of homeologous genes that are both under selection should be correlated. Figure 3.4 shows that homeologous K_S values are uncorrelated ($r^2 = 0.05$), and the K_A values show a weak correlation ($r^2 = 0.45$). We are currently investigating if those genes under similar selection in both sub-genomes are under similarly tight selection across tetrapods.

The 4DTv distributions of the aligned orthologs is shown in Figure 3.5. These are comparable to previously published measurements (Hellsten, 2007). Directly comparing the raw K_A and K_S measurements between *X. laevis* L/S and *X. tropicalis* does not reveal a difference in rate of sequence change between sub-genomes, however we are interested in comparing the sequence change specific to the L/S sub-genomes, regardless of how much change happened for a gene on the *X. tropicalis* lineage (Figure 3.6). For those loci that retain both homeologs in *X. laevis*, we can estimate the sequence change that occurred only after L/S diverged from one another (Figure 3.3).

3.2.1 Differing rates of sub-genome-specific sequence change may be due to speciation

The sub-genome-specific rates of protein-coding sequence change reveal that the S sub-genome has undergone slightly more sequence change than L for both synonymous and nonsynonymous change, as well as 4DTv (Figure 3.7, S/L K_S shift=9%, S/L K_A shift=18%). Without a divergent *Xenopus* polyploid for comparison we cannot tell if this accelerated sequence change happened before or after the hybridization of the *X. laevis* progenitors. I performed similar analysis on a set of well annotated mammals with no recent history of polyploidy (mouse, rat, human), to ask if speciation alone could cause the difference we see in genome-specific mutation rates (Figures 3.8, 3.9). The rat genome shows a higher rate of sequence change than mouse since their speciation. This is across all types of sequence change, similar to the S and L relationship. These experiments support a model that the accelerated rate of sequence change is due to differences that occurred prior to the hybridization of the progenitor species. We are currently working on obtaining data from *Xenopus borealis*, a divergent allotetraploid, to confirm this.

3.2.2 Using *Hymenochirus* to determine rate of sequence change in *Xenopus* ancestors

While the sub-genome-specific rates of sequence change allow me to compare the L/S sub-genomes to one another, I needed to do more partitioning of the sequence change to ask if the sub-genomes of *X. laevis* have experienced more sequence change than *X. tropicalis*. I took all available mRNA data for *Hymenochirus boettgeri*, a related frog, from NCBI, and identified orthologs in *X. tropicalis*. Alignments were performed as previously described for *Xenopus*, and Figure 3.10 shows the expanded amphibian phylogenetic tree and the equations needed to separate the sequence change in the *X. laevis* ancestor (c), from the sequence change in the *X. tropicalis* lineage (t). Those loci with a *Hymenochirus* ortholog, and two copies in *X. laevis* were used to calculate the total sequence change in the *tropicalis* lineage (t) and both *X. laevis* lineages (a+c for L, b+c for S). For both K_S and K_A, both lineages of *X. laevis* have experienced more sequence change than *X. tropicalis* (Figure 3.11). This supports the hypothesis of increased mutation rates following polyploidy.

I am currently working with Kelly Miller in Rebecca Heald's lab to assemble the complete transcriptome of *H. boettgeri*. Kelly generated 6 RNA-seq libraries from different developmental stages. I used Trinity (Grabherr, 2011) to assemble them into transcripts (normalizing read counts per experiment). I used BLASTX to align them to the *X. tropicalis* proteome, and used the BLAST bit score to identify the longest best hit for each *X. tropicalis* protein in the Trinity output. 13,520 *X. tropicalis* genes have an ortholog in *H. boettgeri*. The heatmap in Figure 3.12 shows that many of the *Hymenochirus* orthologs are fragmented. Manual inspection of many of these has shown that shorter Trinity transcripts can complete the *Hymenochirus* ortholog, and we are currently working on "scaffolding" these contigs into a complete transcriptome to use in comparative analysis with *Xenopus*. A complete *Hymenochirus* transcriptome will give us better resolution in comparing the sequence change between *Xenopus* ancestors, as well as help us understand the differences in sequence change observed between chromosomes discussed in section 3.4.

3.3 Non-coding sequences show differences in rates of sequence change and predict gene retention

The work above outlines the differences in mutation of protein-coding genes between sub-genomes, but does not include analysis of the non-coding elements. Noncoding sequences can be difficult to predict without a wealth of epigenetic data, so I started by studying the pvCNEs discussed in section 2.4.

The alignments of pvCNEs were done using MUSCLE (Edgar, 2004). The alignments of all elements were concatenated. Gaps were removed from the alignment using Gblocks (Talavera, 2007) and a neighbor-joining tree was generated using MEGA6 (Tamura, 2013) with

1,000 bootstraps, the Kimura 2-parameter model, and uniform rates among sites. The evolutionary rates of pan-vertebrate aCNEs were compared for every pair of tetrapods using Tajima's relative rate test with elephant shark as an outgroup (Table 3.2). The neighbor-joining tree was built using MEGA6 and is described in Figure 3.13. The tree shows that the S sub-genome pvCNEs are subject to a higher rate of sequence change than the L pvCNEs, and that rat/mouse show a similar relationship. Similar to protein-coding genes, the L sequences also have experienced more mutations than *X. tropicalis*.

Next we wanted to understand if all CNEs showed this difference between sub-genomes. Whole-genome alignments were done using CACTUS (Paten, 2011). Prior to running the program all annotated genomic sequences were masked, including repetitive elements, protein-coding genes, and microRNA genes. The *X. tropicalis*, *X. laevis-L*, and *X. laevis-S* genomes were analyzed as distinct species, using default parameters. Each set of masked orthologous chromosomes placed by BAC-FISH was fed to CACTUS to reduce the computational load of aligning non-homologous chromosomes. We filtered alignments for those >50 bp in length, present once and only once in *X. tropicalis*, and at most once in either or both sub-genomes of *X. laevis*. This analysis is ongoing; the results discussed below concentrate on the CACTUS alignments between *X. tropicalis* Chr01, *X. laevis* Chr01L, and *X. laevis* Chr01S.

To determine the best inter-element distance to combine conserved sequences into putative "enhancers" we computed an ROC curve to test how different merging lengths best replicate the lengths of experimentally-confirmed functional non-coding elements (Figure 3.13, data kindly provided by Rachel Kjølby in Richard Harland's lab). Merging elements within 650 bp maximizes the True Positive Rate while minimizing the False Positive Rate. The alignments computed by CACTUS within these regions are used for our non-coding evolution analysis.

I concatenated CACTUS alignments and removed gaps using Gblocks. Trees were built using the R *ape* package (Paradis, 2004), and significance of branch lengths computed by a Tajima's relative rate test on the final concatenated/ungapped alignments. This analysis reveals that the S sub-genome CNEs are mutating faster than L (Figure 3.14). Conserved non-coding elements (CNEs) within +/- 100kb of a gene are assigned to that gene as its "regulatory landscape". If two protein-coding genes are within 200 kb of one another, the intergenic distance is halved and CNEs are assigned to the nearest gene. One model of evolution following polyploidy predicts that genes with more enhancer sequences would be more likely to be retained due to a higher likelihood of subfunctionalization (Wertheim, 2013). I found that *X. tropicalis* genes with more flanking CNEs are more likely to be retained in two copies (Figure 3.15). CNEs near a gene are not necessarily enhancers for that gene. We need more functional data to be confident that the CNEs we identify are enhancer, and then to assign those enhancers to their relevant genes. One interpretation of the results in Figure 3.15 is that those genes with more flanking CNEs are buffered against large deletions. The flanking CNEs, whether they regulate the nearest gene or not, are important sequences that are maintained. Thus, the increased retention rate of these genes is not support for a model of subfunctionalization-driven gene retention, but instead supports a model for sequence density driving gene retention. We are currently investigating if this is true for all protein-coding genes, as well as working with collaborators who have done the functional experiments necessary to assign enhancers to genes and ask if the potential for subfunctionalization is one of the driving forces of gene retention in the *X. laevis* genome.

3.4 Differences between *X. laevis* regions orthologous to Xtr-9 and Xtr-10 reveal insights into chromosome fusion effects on sequence change

Table 2.2 shows that the gene retention on the regions orthologous to Xtr-10 fails to reject the null hypothesis of equal gene retention in each sub-genome. It is possible this difference could be due to the small number of genes on Xtr-10, however cataloging all of the ways those regions differ in their structural evolution could reveal mechanisms behind the structural changes seen in the rest of the genome. Figure 3.16 shows boxplots of K_S and K_A

distributions by *X. laevis* chromosome. Interestingly, the regions of chromosomes 15 and 18 (9_10 L and 9_10 S) orthologous to Xtr-10 have an elevated K_S, but not K_A when compared to the rest of the genome. This acceleration of sequence change is distinct from those discussed above, as it only increases the synonymous substitution rate.

Further investigation of Xtr-10 evolution reveals that the 3rd codon GC% is elevated on Xtr-10 compared to other chromosomes (Figure 3.17). This increase is shared by *X. laevis* orthologs for this region (Figure 3.18). Figure 3.19 shows the difference in 3rd codon GC% between *X. laevis* sub-genomes and *X. tropicalis* for chromosomes 1–9 and 10. Interestingly the *X. laevis* sub-genomes have similar 3rd codon GC% to *X. tropicalis* on the first 9 chromosomes, but less 3rd codon GC% for those regions orthologous to Xtr-10. Additionally the S sub-genome appears to have a wider variation of 3rd codon GC% when compared to L on chromosomes 1-9, but a lower 3rd codon GC% on the region orthologous to Xtr-10.

These differences between the species are likely due to differences in chromosome length, similar to results seen in humans (Duret, 2009). Xtr-10 is the smallest chromosome, about half the length of Xtr-9. If we assume the rate of recombination by chromosome arm is equal (i.e., 1 crossover/chromosome arm/gamete), then Xtr-10 experiences a higher rate of recombinations/nucleotide. Gene conversion occurs during recombination, and is known to be GC-biased (Duret, 2009). The increased GC% of Xtr-10 may be due to gene conversion converting alleles to the GC-rich allele much more often than the other chromosomes over thousands of generations. When the chromosome fusion happened in the *X. laevis* ancestor, the rate of recombinations/nucleotide were lowered to be similar to the rest of the genome so afterwards the GC% would evolve neutrally. Under this model, we hypothesize that the 3rd codon GC% difference between L/S for these genes is due to the accelerated rate of sequence change in the S sub-genome alone.

These experiments characterize the structural differences between sub-genomes in *X. laevis.* A number of structural variables are involved in determining the evolution of DNA sequences following allopolyploidy, however an understanding of the functional evolution following polyploidy is needed to fully discuss the molecular history of *X. laevis*.

	L		S	
2-2-2 N=1,401	1.05		1.03	
2-1-2 (pseudogene) L=216 S=497	Missing: L 1.02	Extant: S 1.01	Missing: S .92	Extant: L 1.07
2-1-2 (no pseudogene) L=81 S=416	.83	.998	.513	1.11

Table 3.1: Difference in sub-genome deletion lengths
Identification of triplet loci is described in Figure 2.9. Loci were classified into groups based on the presence of gene 2 in both *X. laevis* sub-genomes (row 1), versus those that had a pseudogene in the middle (row 2) or no remnant of the middle gene at all (row 3). The length of sequence between *X. laevis* genes 1&3 was divided by the length between *X. tropicalis* A&C to normalize the intergenic lengths. The median of the normalized ratio distribution is included in the table.

If gene B was present in both copies, either functional or as a pseudogene, the length of the sequence is close to that of *X. tropicalis* (92%-107%). When a pseudogene is not found the length of the extant gene is again close to *X. tropicalis* but the missing gene has a shorter length than *X. tropicalis* (83% for L, 51.3% for S).

Species A	SpeciesB	Identical sites	Divergent sites	Independent Outgroup Substitutions	Independent A Substitutions	Independent B Substitutions	Corrected p-value	Significant?
Human	Mouse	101918	425	10334	933	2387	2.3E-139	A<B
Human	Rat	101743	433	10309	954	2580	1.4E-163	A<B
Human	Chicken	103040	432	8976	2286	1271	8.4E-64	A>B
Human	Lizard	101851	667	8629	2400	2501	2.7377	NA
Human	Tropicalis	99375	1005	7955	2736	4977	1.79E-142	A<B
Human	Laevis L	99209	1041	7932	2723	5143	8.74E-163	A<B
Human	Laevis S	98389	1082	7840	2774	5963	5.76E-254	A<B
Mouse	Rat	102090	204	12357	584	755	4.32E-05	A<B
Mouse	Chicken	101447	575	8712	3856	1386	6.14E-254	A>B
Mouse	Lizard	100337	801	8477	3868	2514	2.76E-63	A>B
Mouse	Tropicalis	97953	1166	7866	4114	4898	2.1E-15	A<B
Mouse	Laevis L	97768	1198	7827	4121	5083	1.6E-22	A<B
Mouse	Laevis S	96964	1249	7741	4156	5887	1.0E-65	A<B
Rat	Chicken	101290	575	8712	4030	1391	3.3E-280	A>B
Rat	Lizard	100176	816	8457	4049	2521	4.0E-78	A>B
Rat	Tropicalis	97770	1167	7839	4316	4927	2.9E-09	A<B
Rat	Laevis L	97594	1199	7810	4313	5103	5.5E-15	A<B
Rat	Laevis S	96784	1248	7719	4355	5913	3.3E-52	A<B
Chicken	Lizard	103030	478	9018	1183	2296	2.8E-78	A<B
Chicken	Tropicalis	100390	849	8145	1685	4936	0	A<B
Chicken	Laevis L	100202	870	8116	1693	5124	0	A<B
Chicken	Laevis S	99381	325	8009	1745	5945	0	A<B
Lizard	Tropicalis	99308	1044	7950	2803	4943	1.9E-129	A<B
Lizard	Laevis L	99133	1070	7928	2799	5118	1.3E-148	A<B
Lizard	Laevis S	98349	1137	7846	2814	5902	9.1E-239	A<B
Tropicalis	Laevis L	101301	261	13045	631	810	3.5E-05	A<B
Tropicalis	Laevis S	100415	390	12799	748	1696	8.2E-81	A<B
Laevis_L	Laevis S	100331	363	12921	832	1601	1.1E-53	A<B

Table 3.2: Evolutionary comparison of pan-vertebrate CNEs
pvCNEs were identified in each genome as described in the main text. For those pvCNEs retained in both *X. laevis* sub-genomes, we concatenated all alignments for each species, aligned them via MUSCLE (Edgar, 2004), and removed gaps using Gblocks (Talavera, 2007). The ungapped, concatenated alignments were analyzed using Tajima's relative rate test from the *ape* package in R (Paradis, 2004). A total of 116,048 sites were analyzed, and elephant shark was used as the outgroup for all comparisons. To correct for multiple sampling of the 116,048 nucleotides across the 28 tests, we multiplied the chi-squared p-value by 28.

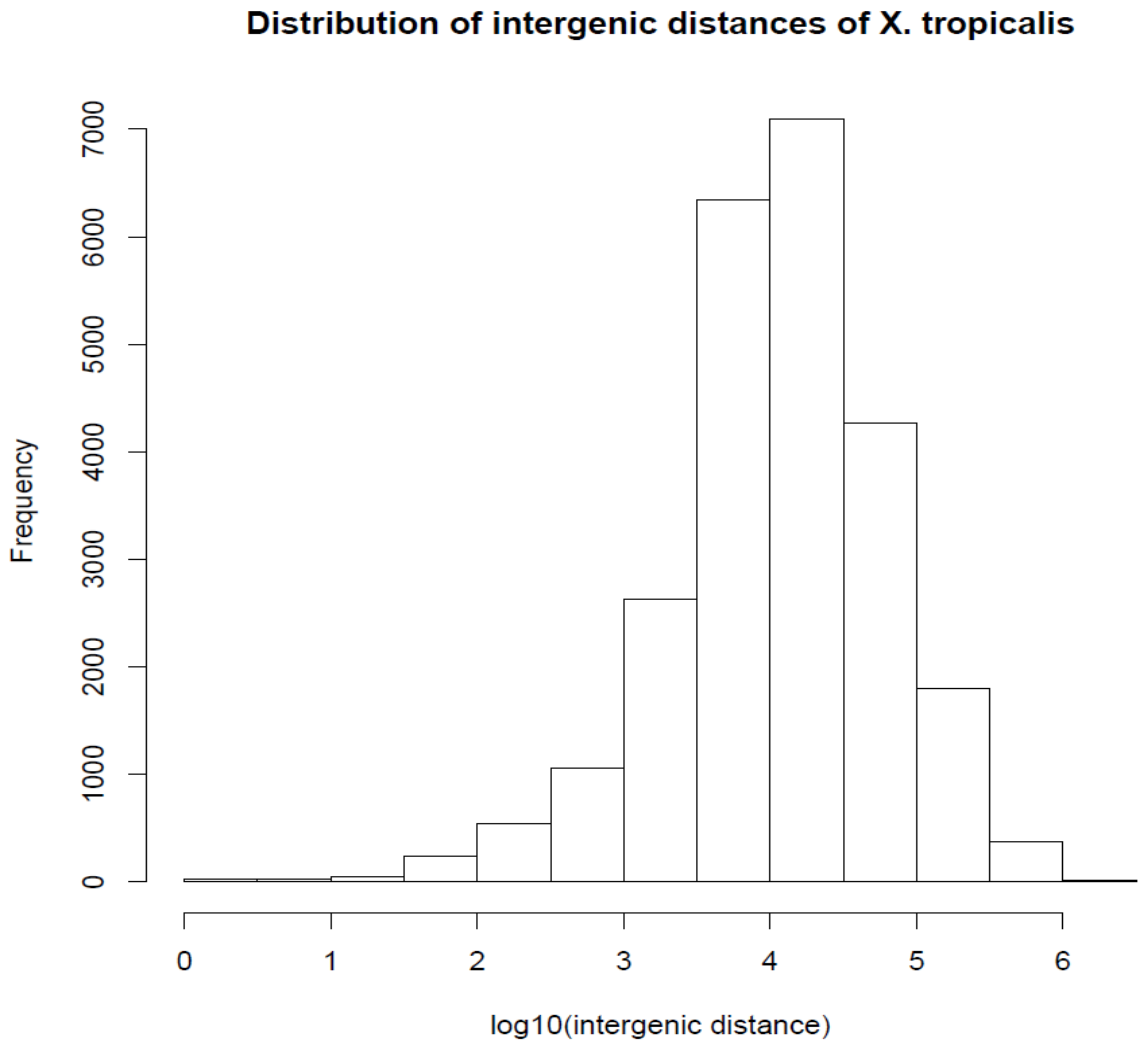


Figure 3.1: Distribution of intergenic distances of *X. tropicalis*
The locations of the protein-coding genes of *X. tropicalis* were extracted from the gff file of the annotation. All non-transposable element loci were used. Intergenic length was defined as the nucleotide distance between the end of the 5' gene of a pair, to the start of the 3' gene. The mean intergenic distance is 36.96 kb ± 77.7 kb.

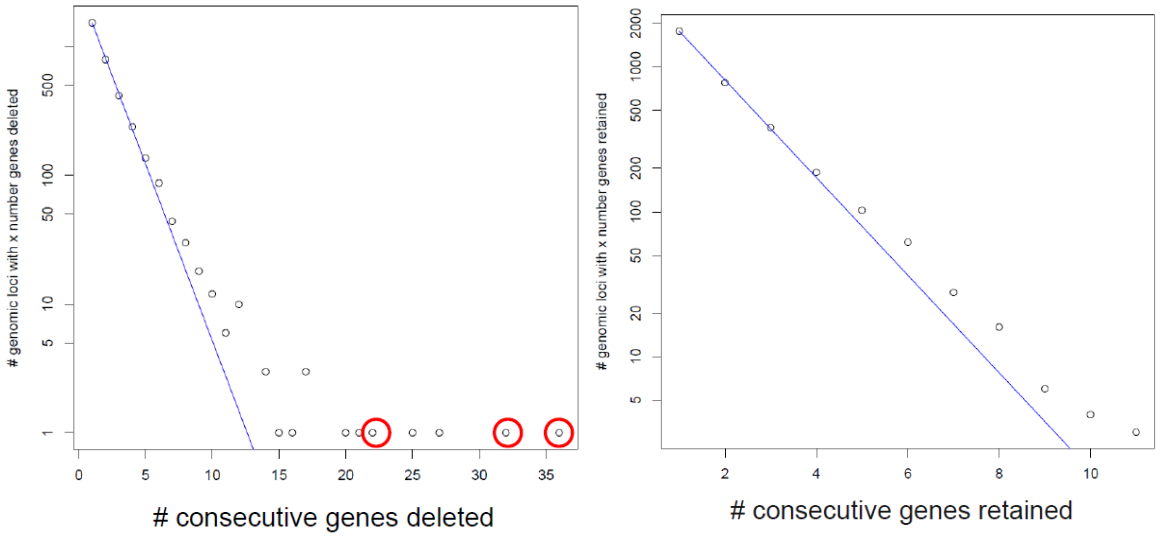


Figure 3.2: Distribution of consecutive gene loss and gene retention
X. laevis protein-coding gene sequences were aligned to *X. tropicalis* in protein space, using BLASTP, evalue cutoff of 1e-10, Smith-Waterman alignment. Protein names were mapped back to each assembly via the gff files output by the annotation pipeline. The number of consecutive *X. tropicalis* loci with a single ortholog in *X. laevis* represent the deletion distribution. The number of consecutive *X. tropicalis* loci with two orthologs in *X. laevis* represent the retention distribution. The blue lines represent the best fit exponential line to the data. The y-axis of both plots is on a log scale, so that the exponential line appears linear. The circled points in the deletion distribution represent the MHC locus, a cluster of cadherin genes, and a cluster of olfactory genes.

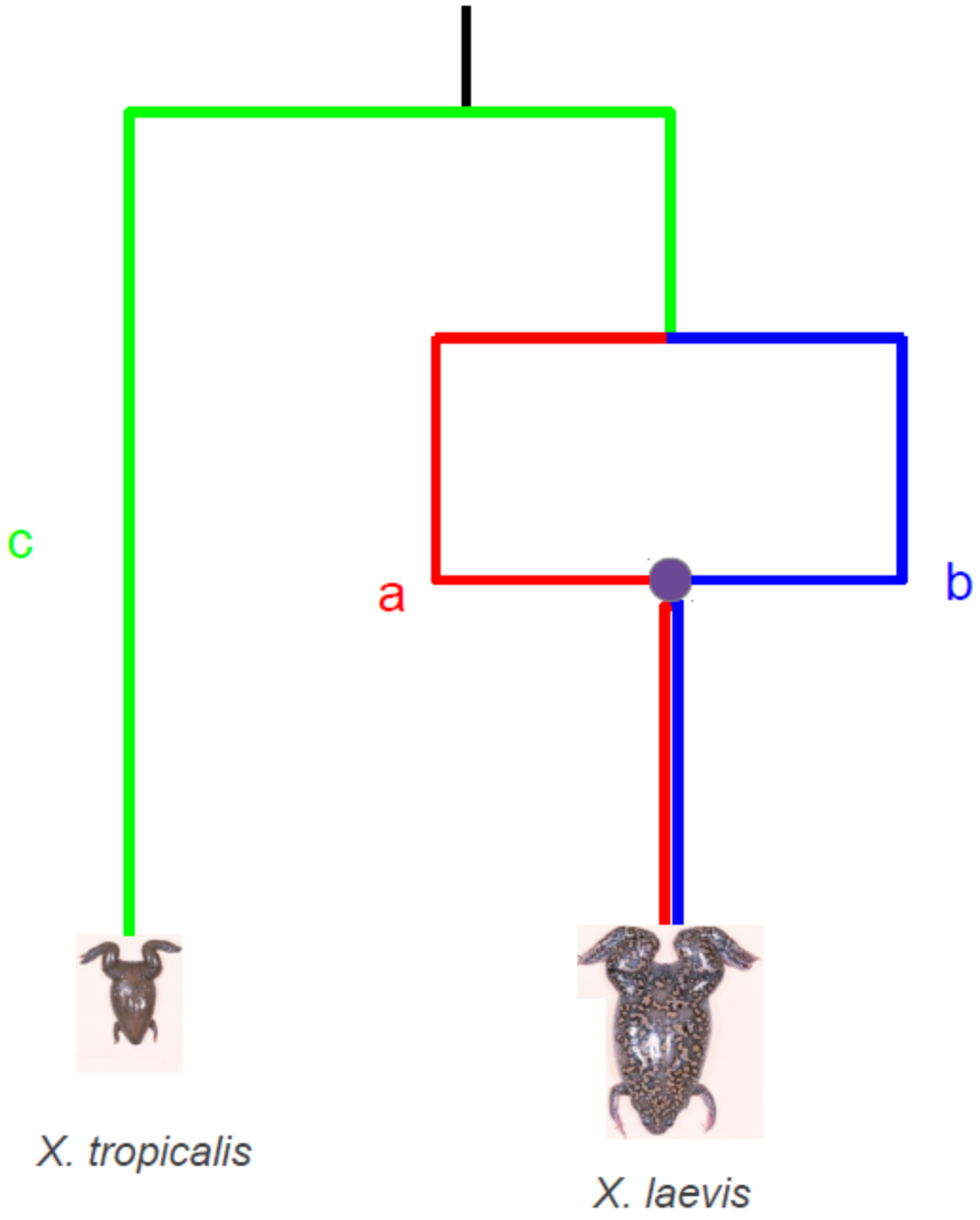


Figure 3.3: Calculation of *Xenopus* sub-genome-specific rates of evolution
Phylogenetic tree of *Xenopus* color-coded to indicate the different epochs of sequence change we can isolate by comparing the K_S or K_A rates between *X. laevis* homeologs and X. *tropicalis* orthologs. When both homeologs of *X. laevis* are retained, we can compare the sequence change measurements with the following equations to isolate the a,b,c variable illustrated above. Tropicalis<->L measurements (TL) measure sequence change along the a+c lineages. Tropicalis<->S measurements (TS) measure sequence change along the b+c lineages. L<->S measurements (LS) measure sequence change along the a+b lineages. From these

measurements we can extrapolate a= $\left(\frac{TL+LS-TS}{2}\right)$ b= $\left(\frac{TS+LS-TL}{2}\right)$ c= $\left(\frac{TL+TS-LS}{2}\right)$

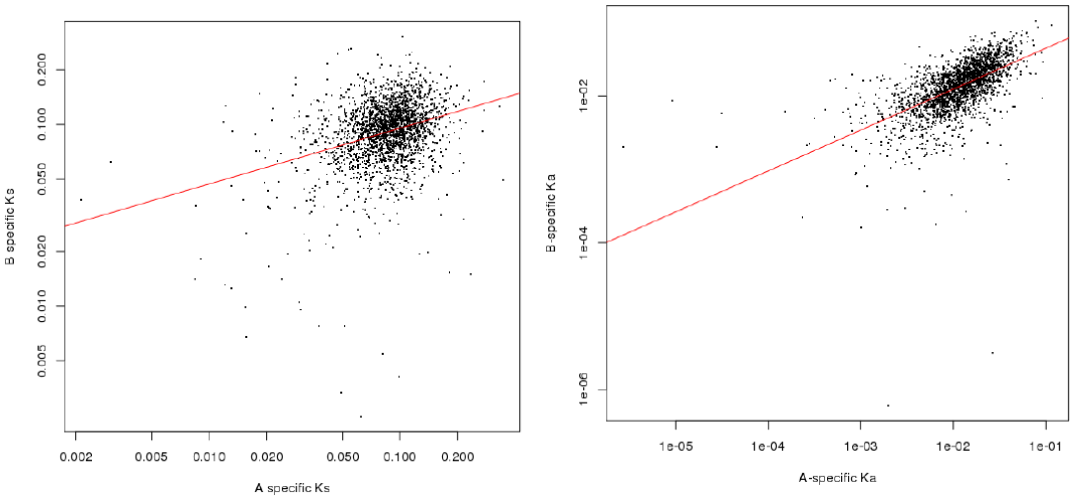


Figure 3.4: Scatterplots of the sub-genome-specific K_S and K_A distributions between homeologs

Scatterplots of sub-genome-specifc rates of K_S (left) and K_A (right). The x-axis is the L-specific rate, the y-axis is the S-specific rate. The red line represents the best fit linear line to each distribution. K_S r^2 = 0.05 (p value 0.76). K_A r^2= 0.42 (p value < $2.2e^{-12}$).

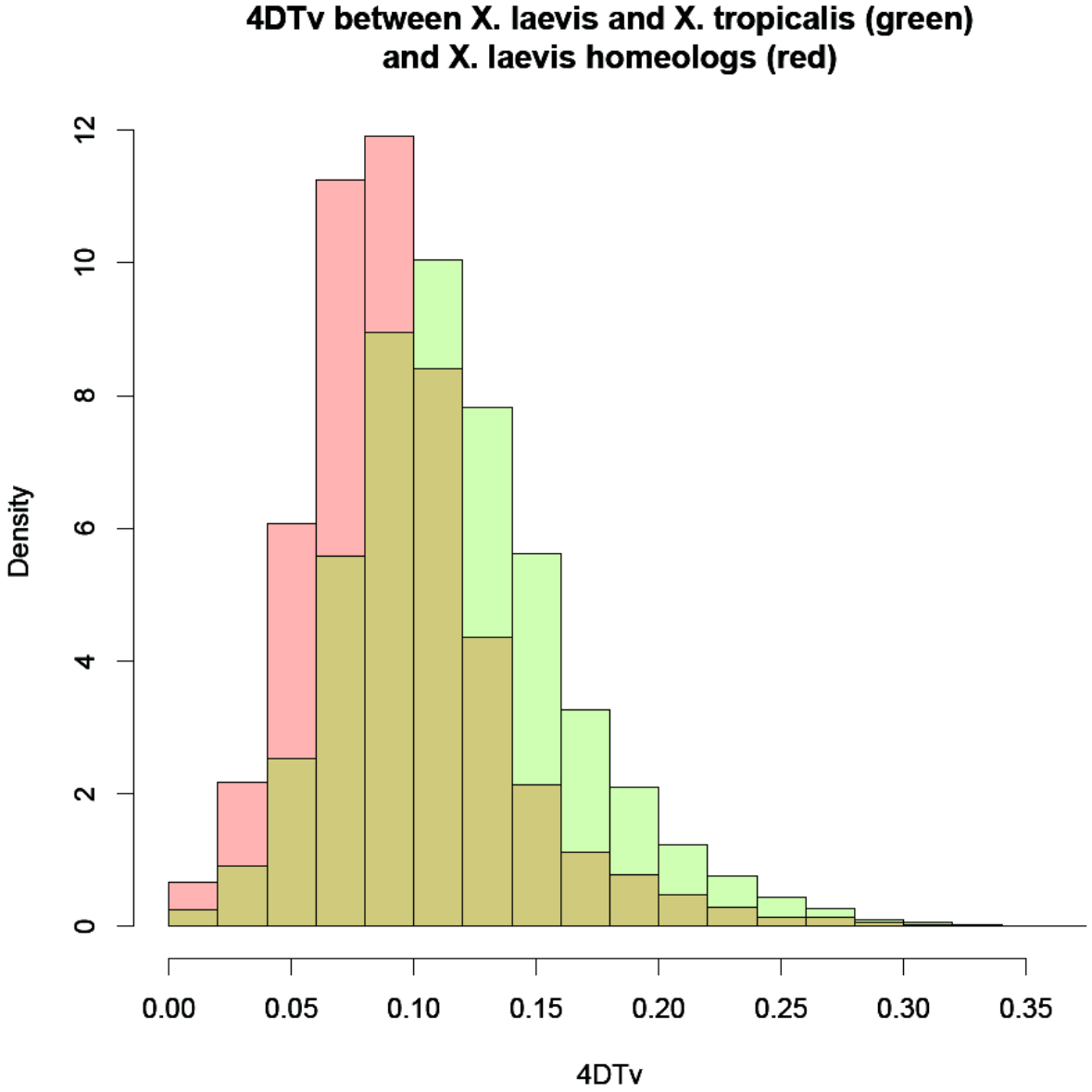


Figure 3.5: 4DTv distributions between *Xenopus* orthologs and homeologs
Four-fold degenerate transversion distributions between X. laevis homeologs (red) and Xenopus homeologs (green). The x-axis is the raw 4DTv distribution. The y-axis is the probability density. We use density for this plot because there are twice as many comparisons for the ortholog comparison as the homeolog comparison. These distributions agree with previously published results by Hellsten et al. 2007 (who had a fraction of the data).

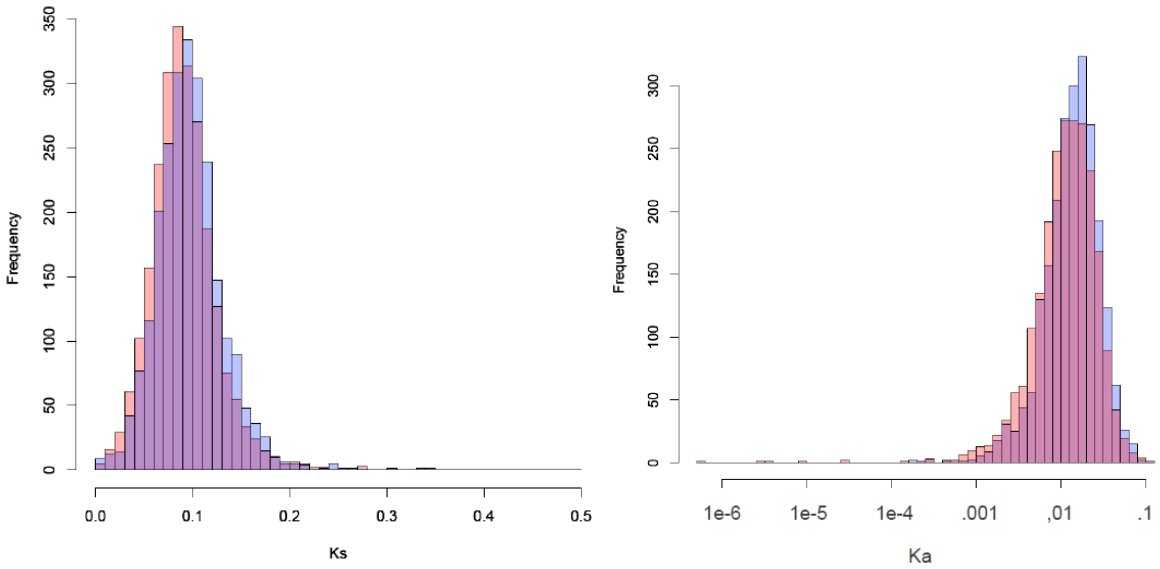


Figure 3.6: sub-genome-specific rates of sequence change between sub-genomes
Histograms of sub-genome-specific rates of sequence change between L (red) and S (blue) for K_S (left) and K_A(right). We used a Wilcoxon-test to test for differences between the distributions and estimate the percentage increase in sequence change in the S sub-genome. K_S pvalue = $2.01e^{-58}$; estimated shift at 99% confidence interval = 9.5% acceleration. K_A pvalue = $1.41e^{-37}$; estimated shift at 99% confidence interval = 18% acceleration.

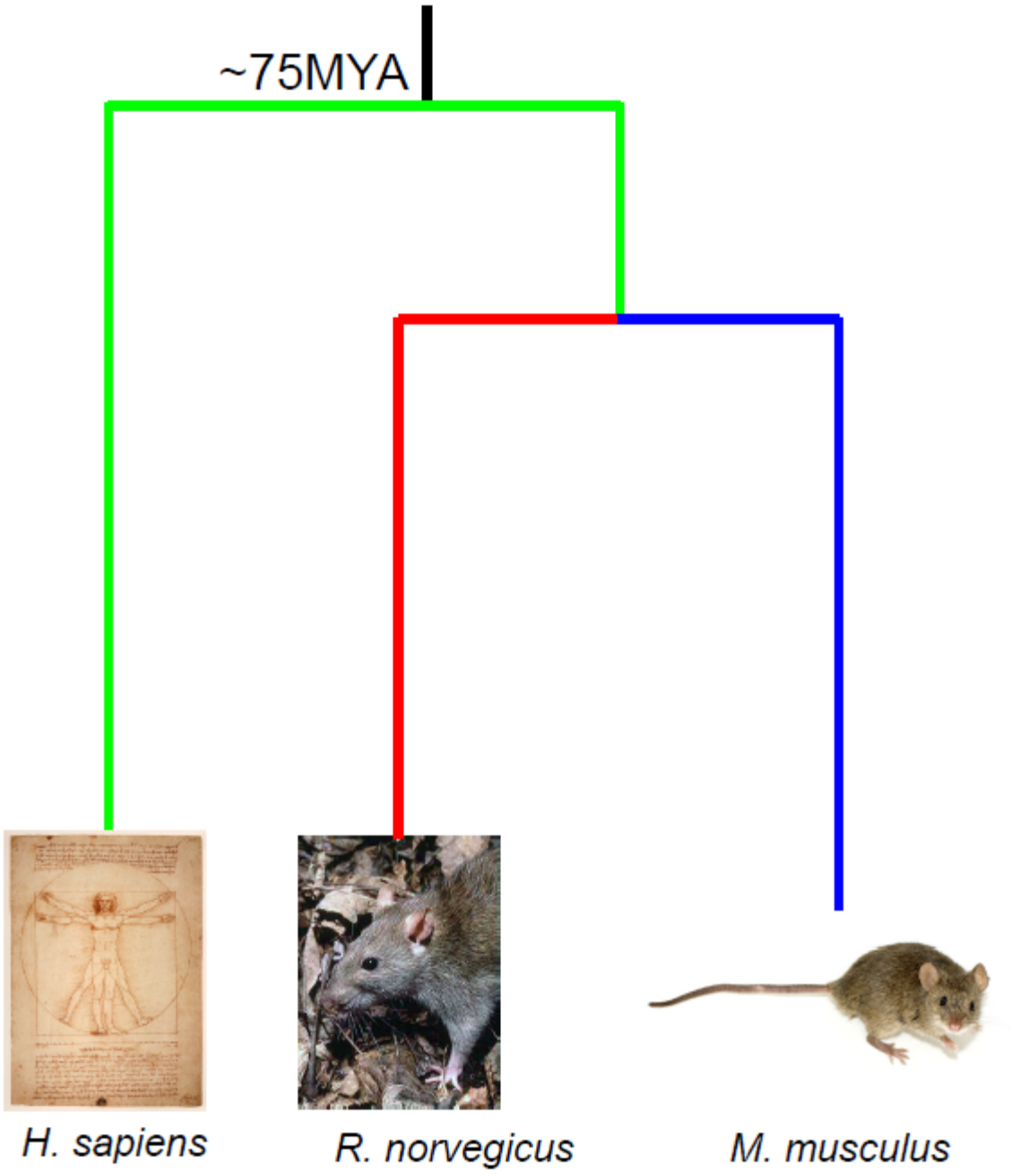


Figure 3.7: Calculation of mammalian genome-specific rates of evolution
Phylogenetic tree of mammals, color-coded to show the epochs of sequence change we isolate with the following equations. Human<->Mouse measurements (HM) measure sequence change along the green and blue lineages. Human<->Rat measurements (HR) measure sequence change along the green and red lineages. Rat<->Mouse measurements (RM) measure sequence change along the a+b lineages. From these measurements we can extrapolate

green= $\left(\frac{HM+HR-RM}{2}\right)$ blue= $\left(\frac{RM+HM-HR}{2}\right)$ red= $\left(\frac{RM+HR-HM}{2}\right)$

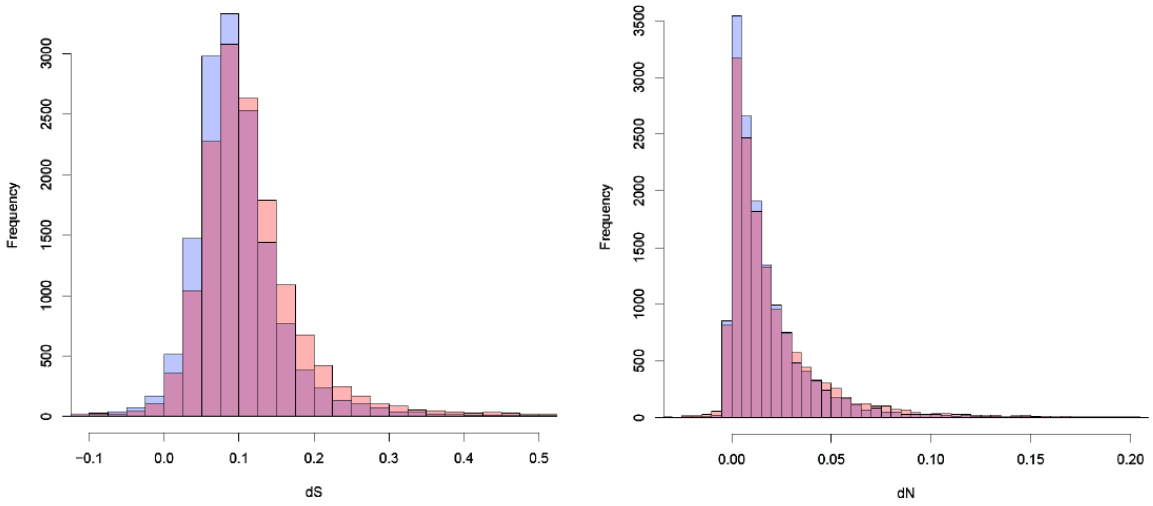


Figure 3.8: Genome-specific rates of evolution of mammals
Histogrames of genome-specific evolutionary rates between murines. We took the ortholog lists and dN/dS measurements from Ensembl v77. The dN/dS tables from Ensembl were rounded to two significant digits, which caused the equations in 3.7 to produce a few negative values. We used the Wilcoxon-Mann-Whitney test to determine significant differences and estimate the acceleration in the rat genome. K_S pvalue = 1e-131; estimated shift at 99% confidence interval = 14% acceleration. K_A pvalue= 3.3e-16;estimated shift at 99% confidence interval = 13% acceleration.

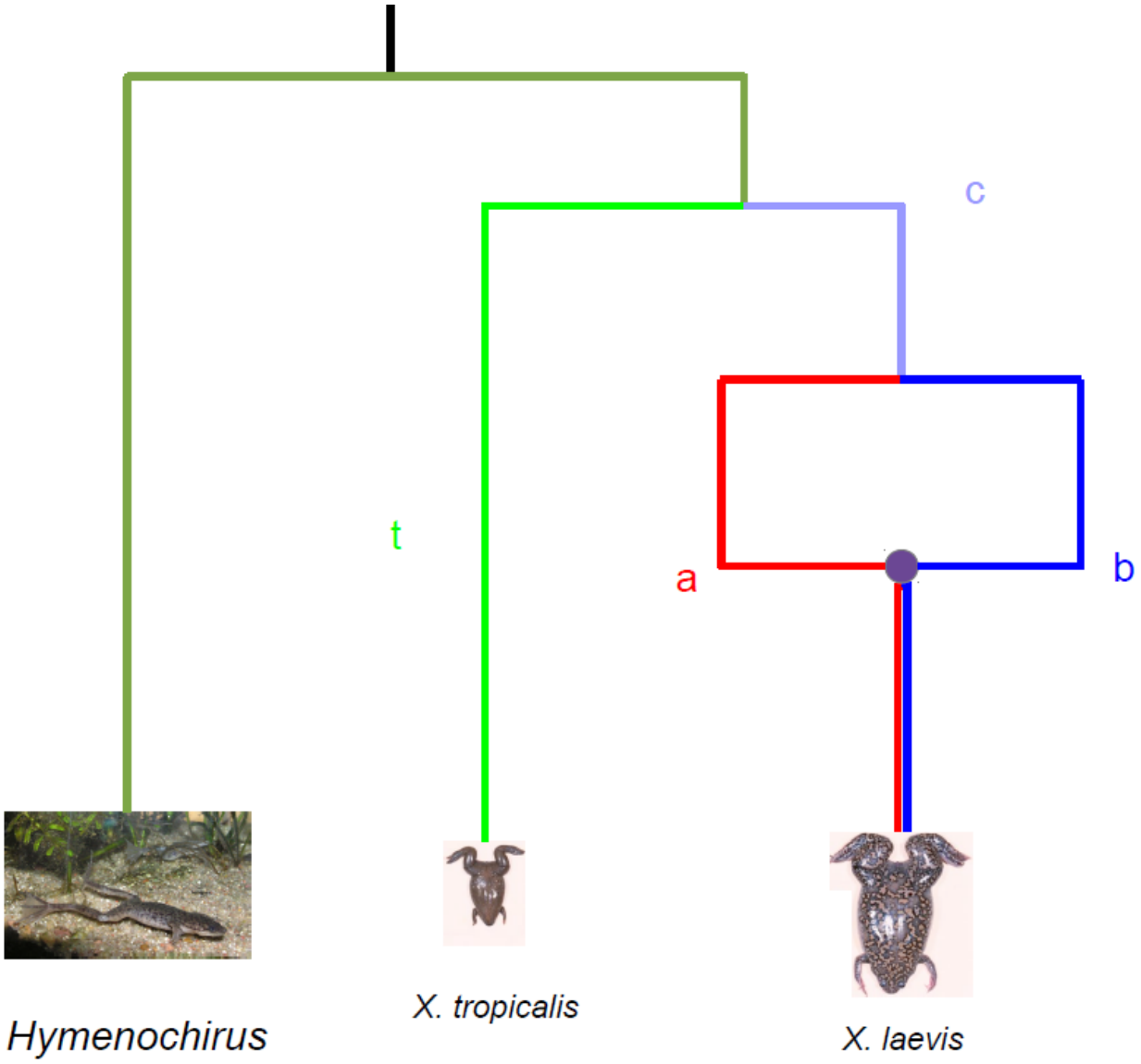


Figure 3.9: Expanded amphibian phylogenetic tree to parse *Xenopus* rates of sequence change

Phylogenetic tree of pipidae, including the dwarf frog *Hymenochirus*. We isolated the a,b,c,t variables above through the following equations: *Tropicalis*<->L measurements (TL) measure sequence change along the a+c+t lineages. *Tropicalis*<->S measurements (TS) measure sequence change along the b+c+t lineages. L<->S measurements (LS) measure sequence change along the a+b lineages. *Hymenochirus*<->L measurements (HL) measure sequence change along the h+c+a lineages. Hymenochirus<->S measurements (HS) measure sequence change along the h+c+b lineages. Hymenochirus<->*Tropicalis* measurements (HT) measure sequence change along the h+t lineages. From these measurements we can extrapolate a= $\left(\frac{TL+LS-TS}{2}\right)$ b= $\left(\frac{TS+LS-TL}{2}\right)$ c= $\left(\frac{TL-HT+HS-LS}{2}\right)$ t= $\left(\frac{TL-HL+HT}{2}\right)$

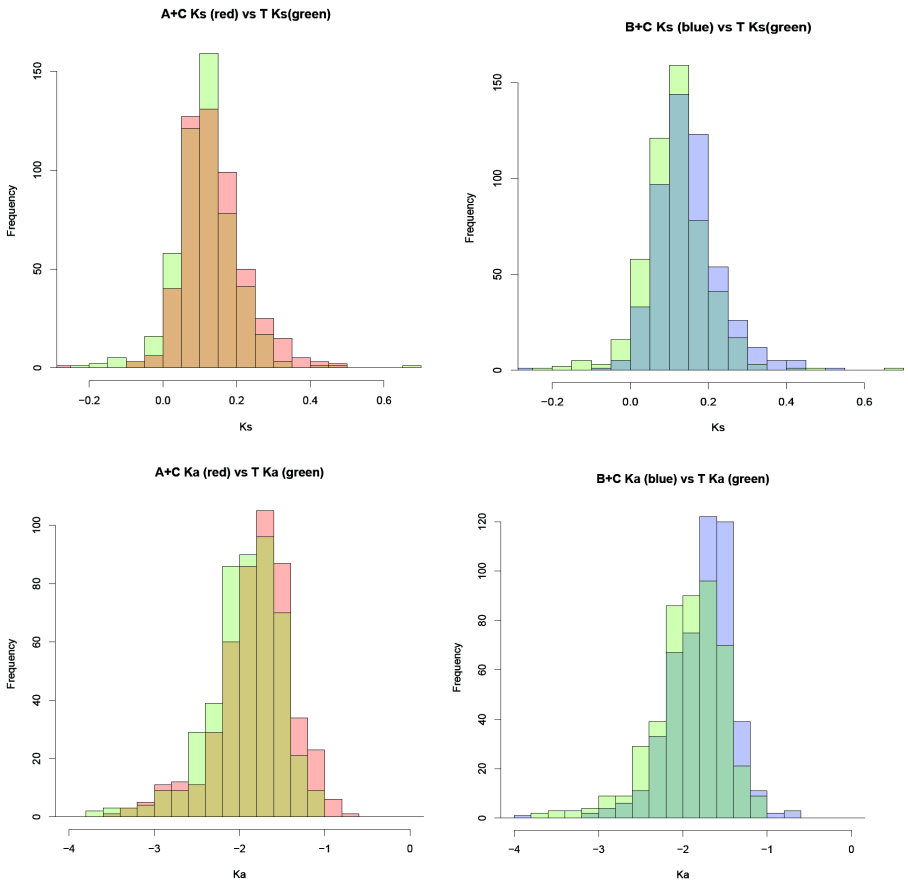


Figure 3.10: Comparison of lineage-specific evolutionary rates between *X. laevis* and X. tropicalis

Histograms of lineage-specific sequence change between *X. tropicalis* (green), and *X. laevis* (L=red, S=blue). *Hymenochirus* sequences were obtained from NCBI and aligned to *X. tropicalis* proteins via BLASTX ($1e^{-10}$, Smith-Waterman refinement) to determine orthology. Wilcoxon-Mann-Whitney tests were used to test for significant differences and estimate acceleration. The y-axis on all plots is frequency, the x-axis on the top plots is K_S, the x-axis on the bottom plots is $\log_{10}(K_A)$. Wilcoxon p-value for all distributions $\leq 2.2e^{-16}$.

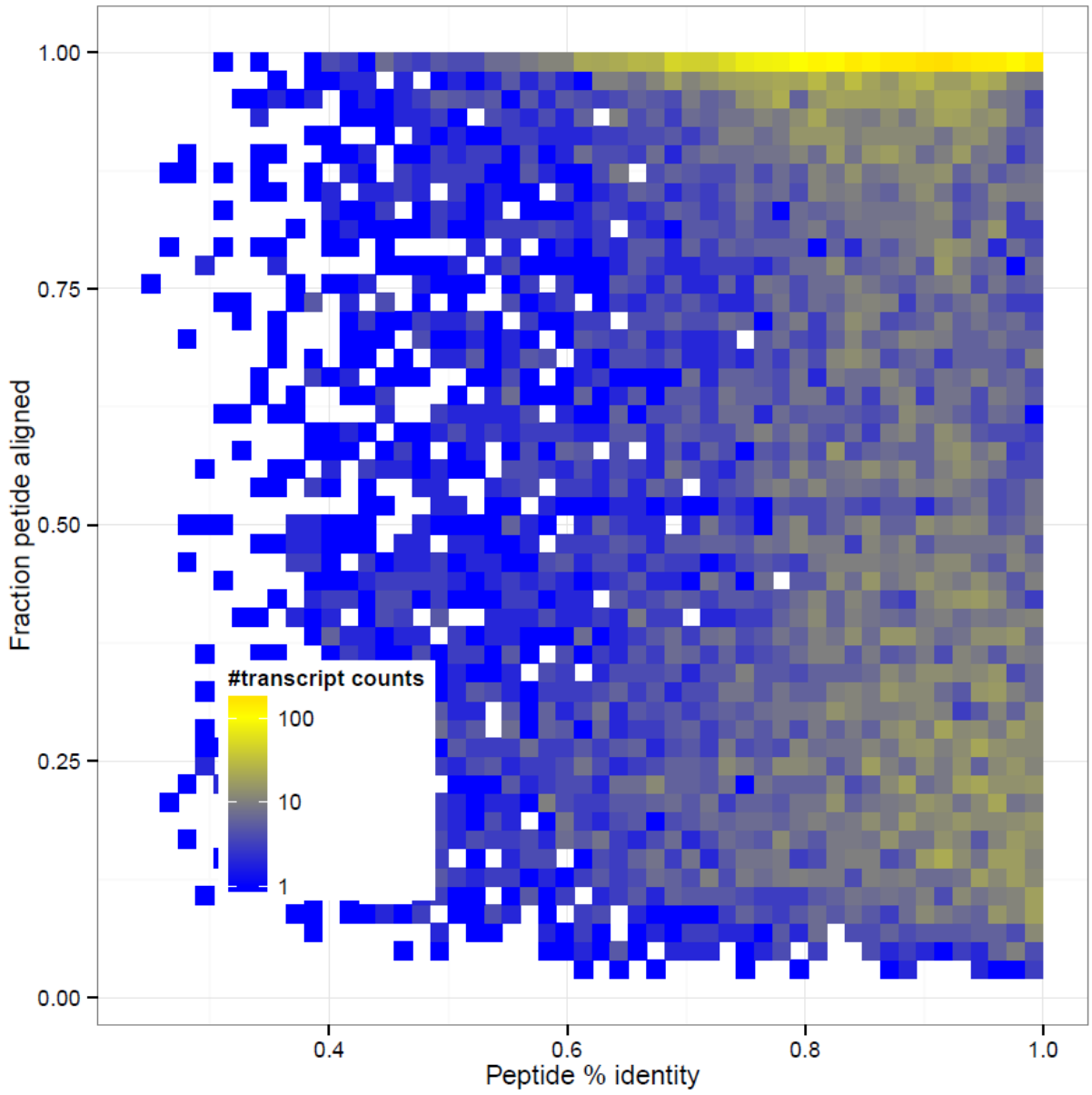


Figure 3.11: Heatmap of assembled *Hymenochirus* transcripts compared to X. tropicalis orthologs

Hymenochirus RNAseq data was kindly provided by Kelly Miller in the Heald lab. We assembled RNAseq reads into transcripts by Trinity Grabherr, 2011). Hymenochirus transcripts were aligned to the *X. tropicalis* proteome by BLASTX (1e-10, Smith-Waterman refinement). The best hit for each *X. tropicalis* protein in the *Hymenochirus* transcriptome was chosen by BLAST bit score (a combination of percent identity and alignment length). The x-axis is the peptide percent identity between species, and the y-axis is the fraction of the *X. tropicalis* protein covered by the longest Hymenochirus transcript.

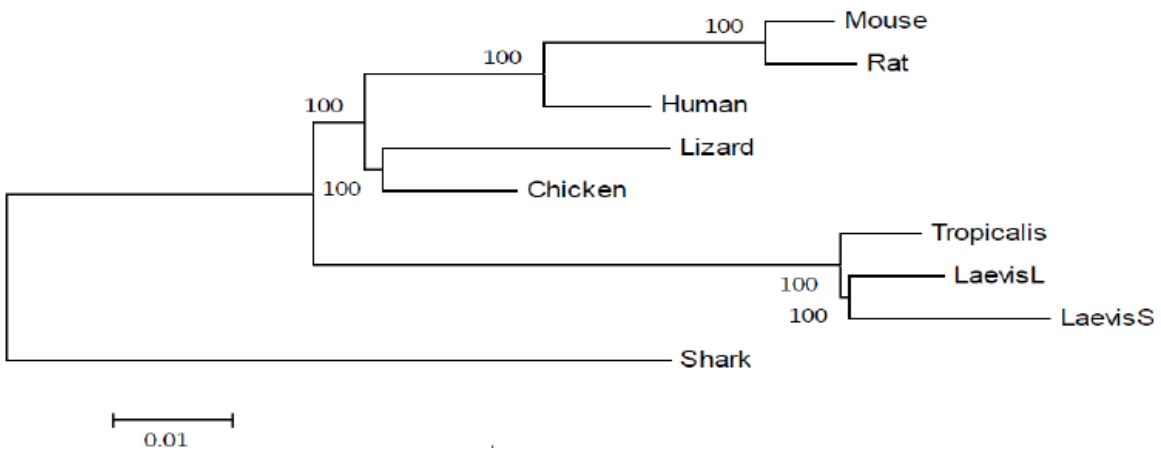


Figure 3.12: Neighbor-joining tree of vertebrate pvCNEs
Phylogenetic tree built from pan-vertebrate CNEs that retain both copies in *X. laevis*; identification is outlined in the main text. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 0.25323906 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Kimura 2-parameter method and are in the units of the number of base substitutions per site. The analysis involved 9 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 115,969 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura, 2013). Stastical investigation of branch lengths is in Table 3.2.

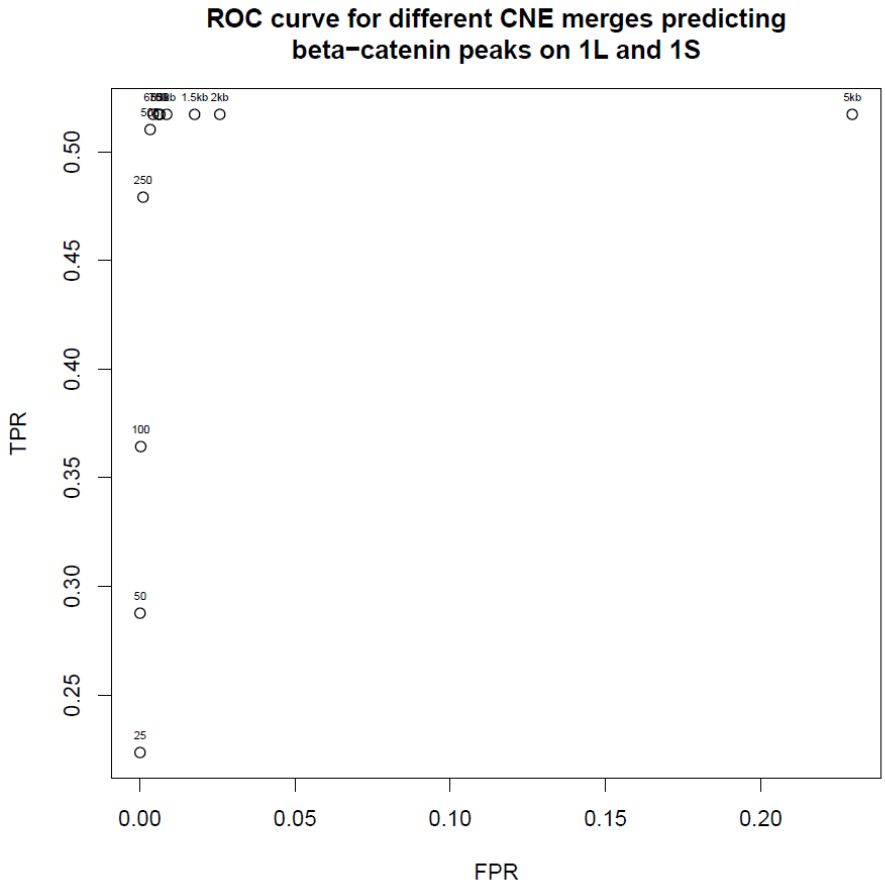


Figure 3.13: ROC curve of CACTUS alignments predicting experimentally-validated enhancers

ROC curve to determine best merging distance for CACTUS CNEs by comparing to experimentally-validated beta-catenin peaks on *X. laevis* chromosome 1L and 1S (kindly provided by Rachel Kjølby, processed by MACS). CACTUS alignments were merged at different distances and compared to beta-catenin peaks by bedtools (Quinlan, 2010). True positive rate (TPR) and false positive rate (FPR) were estimated by comparing the number of unmerged CNEs that overlapped with a beta-catenin peak to those in each of the merged data sets. We selected a 650 bp merging to assess the number of elements flanking gene sequences discussed in Figure 3.15.

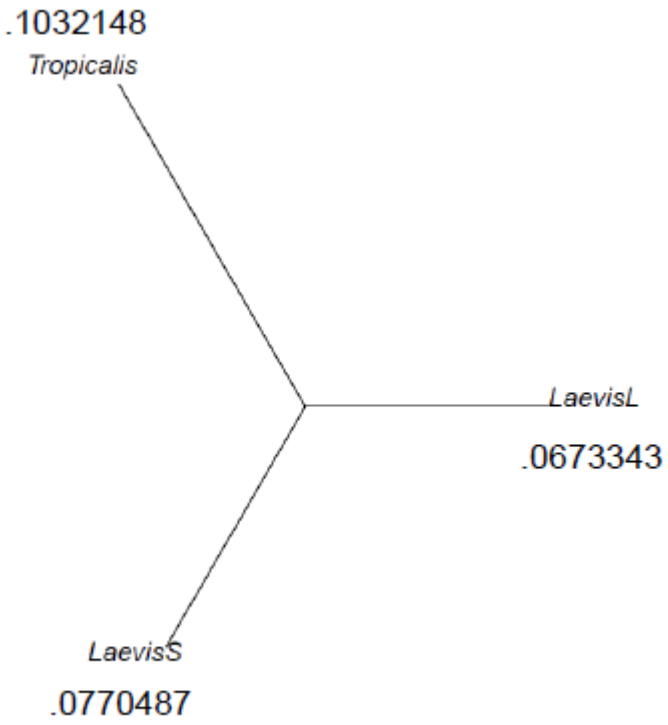


Figure 3.14: Phylogenetic tree of Xenopus CNEs on chromosome 1
CACTUS CNEs longer than 50 bp, and retained in both *X. laevis* lineages. The alignment was extracted directly from CACTUS, concatenated for each species, and analyzed by the *ape* package in R (Paradis, 2004). Tajima's relative rate tests confirms stastically significant difference in mutation rates between L and S CNE sequences (p-value < $2.2e^{-12}$).

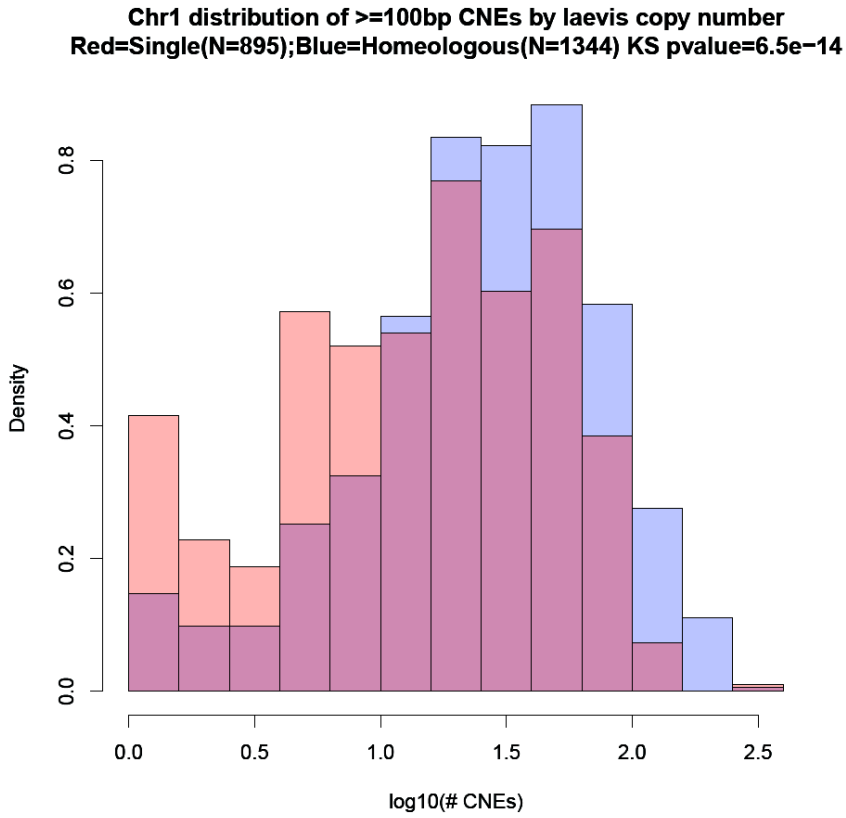


Figure 3.15: Distribution of the number of CNEs flanking *X. tropicalis* loci with one or two *X. laevis* co-orthologs

X. tropicalis CNEs were assigned to genes as described in the main text. The x-axis is $\log_{10}$(#CNEs). The y-axis is the probability density for each distribution. *X. tropicalis* genes with a single *X. laevis* ortholog are shown in red, *X. tropicalis* genes with two *X. laevis* orthologs are in blue. We used a Kolmogorov–Smirnov test to determine significant difference in the # CNEs retained by the two groups (p-value = $6.5e^{-14}$).

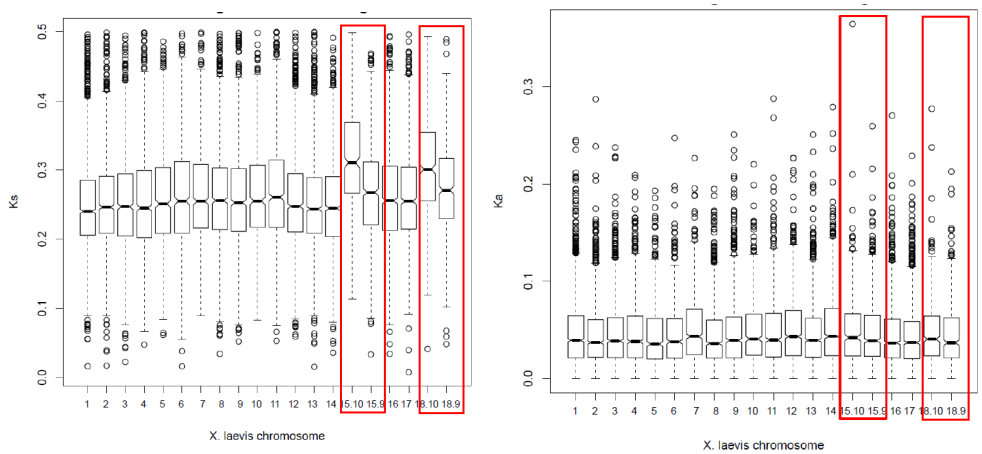


Figure 3.16: Boxplots of K_S and K_A by *X. laevis* chromosome

Alignments of orthologs were done by Dialign-TX [ref]. K_S and K_A calculations were done using the seqinR package (Charif and Lobry 2007). Chromosomes are numbers by their original *X. laevis* karyotype assignment, with the exception of 15 and 18, the 9_10L and 9_10S chromosomes respectively. [15/18].10 refers to those regions of the chromosomes orthologous to Xtr-10; [15/18].9 refers to those regions of the chromosomes orthologous to Xtr-9. The red boxes illustrate the increase K_S rate only on the regions orthologous to Xtr-10. The K_A does not accelerate in these regions, however.

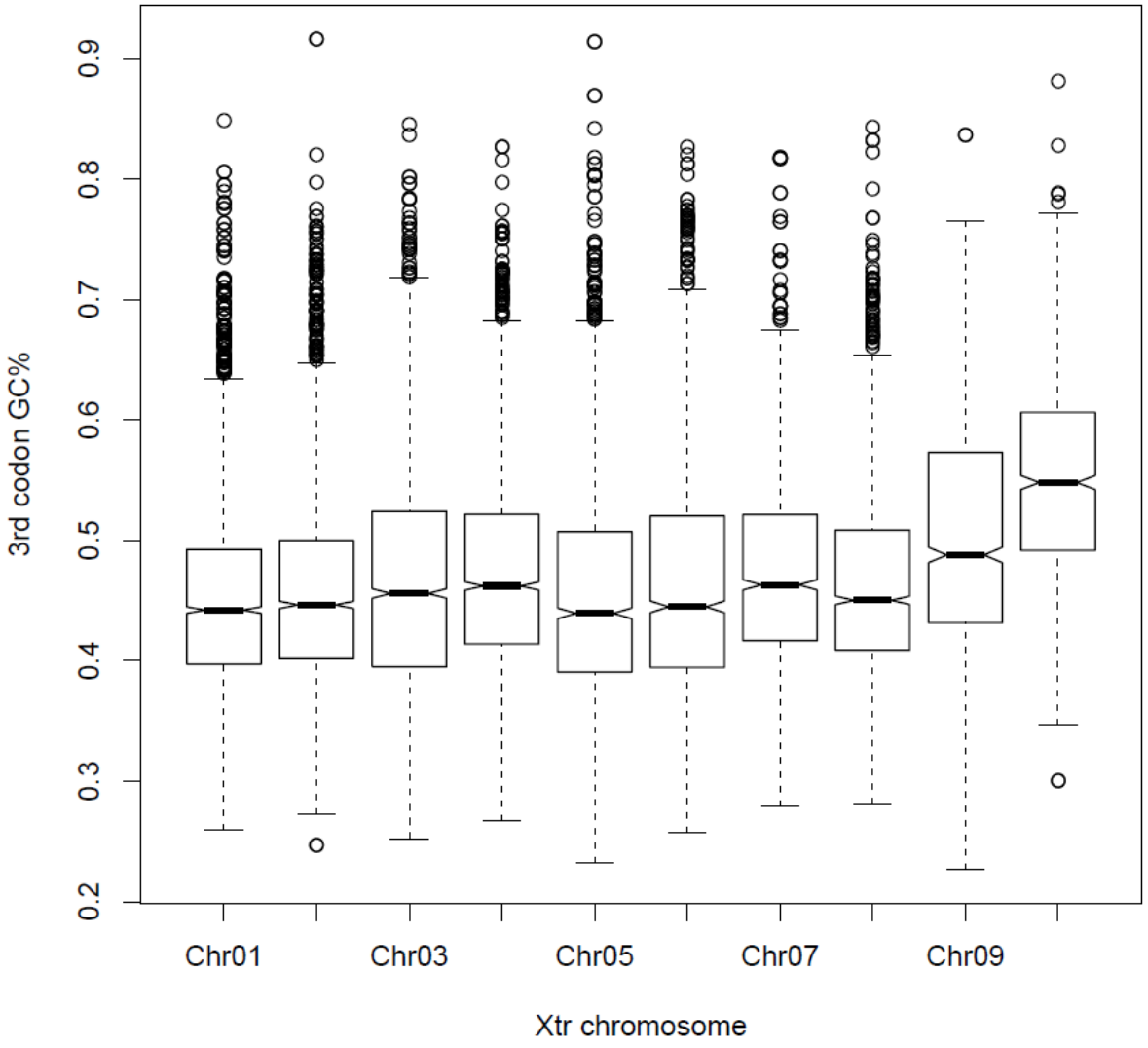


Figure 3.17: *X. tropicalis* 3rd codon GC%
X. tropicalis 3rd codon GC% was computed by the seqinR packages (Charif and Lobry 2007). The genes on Xtr-10 have an elevated amount of GC% at the 3rd codon positions relative to the other chromosomes.Wilcoxon p-value < 2.2 e^{-12}

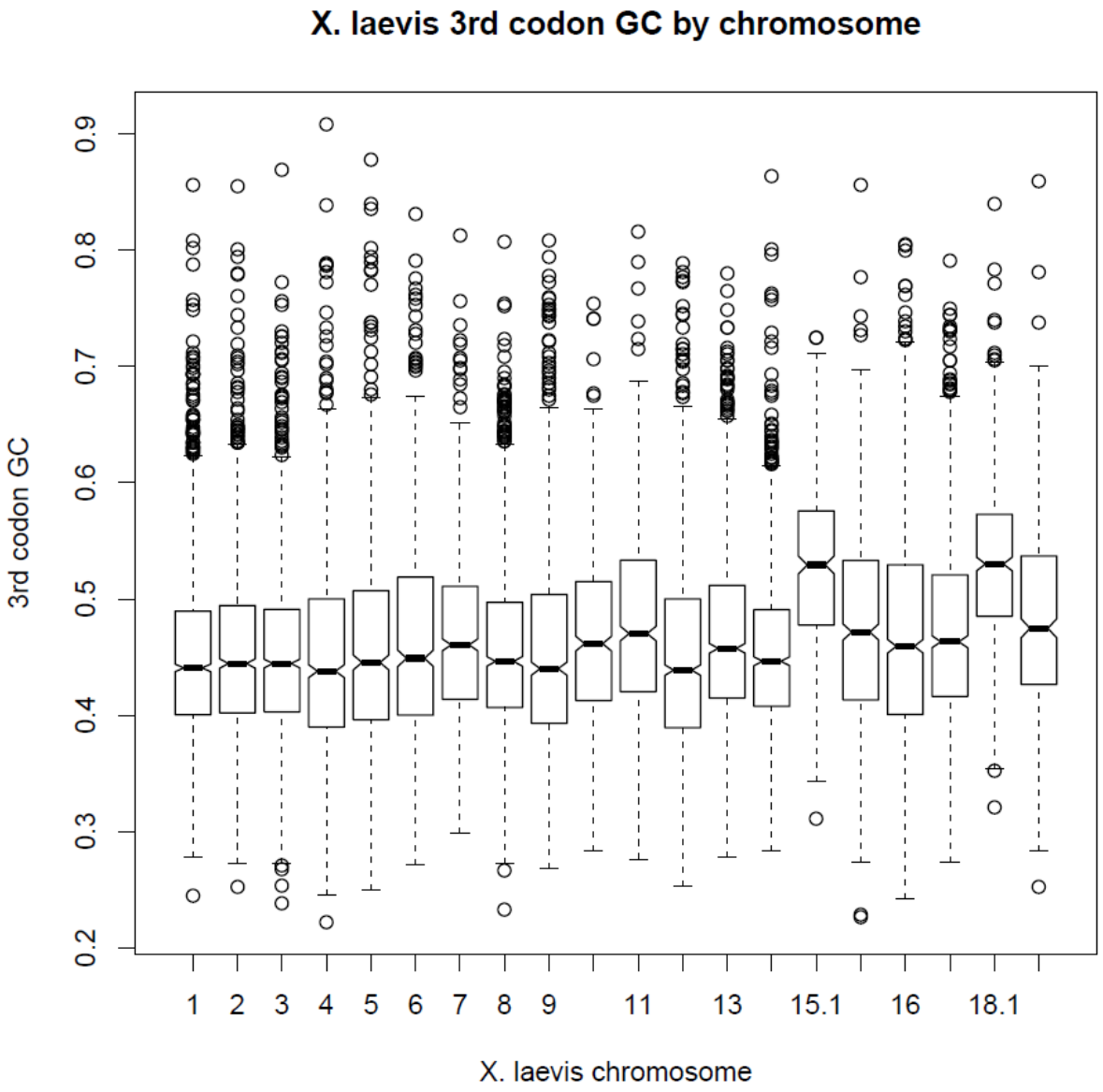


Figure 3.18: *X. laevis* 3rd codon GC%
X. laevis 3rd codon GC% was computed by the seqinR packages (Charif and Lobry 2007). The genes orthologous to Xtr-10 have an elevated amount of GC% at the 3rd codon positions relative to the other chromosomes. Wilcoxon p-value < $2.2e^{-12}$.

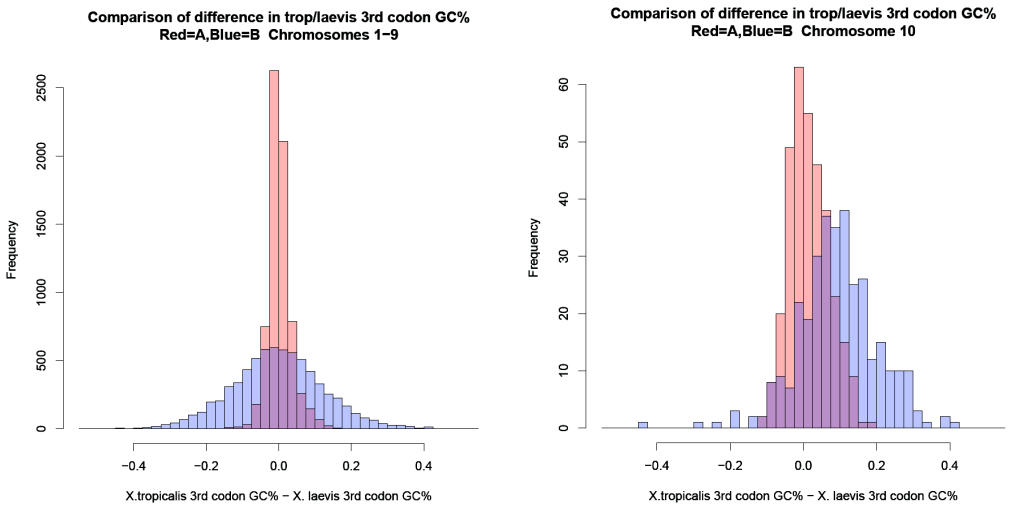


Figure 3.19: Comparison of *X. tropicalis* and *X. laevis* 3rd codon GC%
We subtracted the *X. laevis* 3rd codon GC% from the *X. tropicalis* ortholog for all ortholog pairs. (Left) Histogram of 3rd codon GC% difference between species for L genes (red, mean=0.0018) vs S genes (blue, mean=0.0087) on chromosomes 1-9. (Right) Same distributions for chromosome 10. Lmean = -0.009 Smean = 0.232.

Chapter 4
Functional evolution of *X. laevis* genome

The function of a gene is a complex trait determined by a number of variables: its protein sequence, expression domains, expression levels, direct and indirect interaction partners, as well as other factors. During evolution genetic pathways may develop independent mutations that help to differentiate species. Yeast geneticists have used cybrids, yeast hybrids that exchange mitochondria between species, to investigate pathways which are incompatible between the nuclear-encoded mitochondrial genomes of yeast (Spirek, 2015). Allopolyploidy is a natural experiment with a similar effect, since two nuclear-encoded genomes are brought together, but only a single mitochondrial genome can be inherited from the initial mother of the hybridization event (Figure 1.6). If any mutations occurred while the two species were apart to make a nuclear-encoded mitochondrial protein incompatible with the mitochondrially-encoded protein set from the opposite species then we would see a sub-genome-bias in genes associated with the mitochondria. This interspecific incompatibility may not be restricted to mitochondria, but may have an effect on the evolution of all genetic pathways, where genes acting in the same pathway may have developed compensatory mutations while the progenitor species were apart.

While gene loss is one mechanism to deal with the redundancy following allopolyploidy, sub-functionalization is another potential resolution. Partitioning of biochemical function and/or expression domains between homeologs could lead to novel, beneficial mutations. Separating the natural decay of a gene sequence from subfunctionalization is especially daunting since genes can pass through a point of non-functionalization to gain novel phenotypes (Bridgham, 2009). The expression patterns of genes prior to duplication may also affect the rate of gene retention, or potential for subfunctionalization. Cataloging the rate at which different types of gene expression differ adds to our knowledge of the plasticity of gene expression evolution. The following chapter seeks to outline how the function of a gene affects its chance at being retained or subfunctionalized following allopolyploidy.

4.1 Gene retention biases by functional categories

Biologists have developed a number of databases to classify genes. Pfam (Finn, 2014) is a database of protein domain sequences (such as 7-pass transmembrane receptors or globin), and Gene Ontology (GO) (Carbon, 2009) is a database of gene functions based on experimental evidence (such as “G-protein coupled receptor activity” or “heme binding”). These classifications are useful for understanding whether certain types of protein sequences are more likely to be retained or lost following polyploidy. The paralogs from the ancient vertebrate duplications are enriched for genes with regulatory function, such as DNA/RNA-binding (Putnam, 2008). It is hypothesized that regulatory genes need to be retained or else they will not regulate all of their targets effectively. The high retention rate in *X. laevis* allows us to test this hypothesis, as well as explore if there are other types of protein sequences retained at significantly higher levels.

Since the mitochondrial-localization sequence has flexibility that can be difficult to predict, we utilize MitoCarta (Pagliarini, 2008), a database of experimentally validated mitochondrial associated proteins in mammals, to investigate whether there is a mitochondrial bias following hybridization of the *X. laevis* progenitors. Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2014) is a group of databases which include cataloging genetic pathways. Genes are not grouped by shared sequence or function, but instead by experimental evidence of involvement in a shared process. Using this database we are able to investigate which pathways are more likely to be retained, and whether any experience a sub-genome-bias due to interspecific incompatibilities.

For each classification discussed in this chapter, we determined whether the homeolog retention rate was higher than expected by a Fisher's exact test (p value < 0.01). Significant differences (or lack thereof) between L/S retention rates were also determined by a Fisher's exact test.

4.1.1 GO, KEGG, Pfam retention

PfamScan (Pfam v27.0) was used to assign Pfam domains to gene loci (Finn, 2014). InterPro2GO was used to map pfam assignments to GO terms (Mitchell, 2015). *X. tropicalis* KEGG assignments were extracted from the KEGG database via the REST API, and mapped onto *X. laevis* loci via orthology. Figure 4.1 contains scatterplots of the L retention rate vs. S retention rate for each group in the different types of classifications. A sample of the groups with significantly higher/lower retention rates is included. As is found with other whole genome duplications, DNA repair and RNA polymerase pathways are reduced to single copy more often than other loci, while homeobox, DNA-binding, and major developmental signaling pathways are retained at significantly higher rates. The best fit line is linear, with a slope that varies from 1.6-2, indicating that the S sub-genome is losing genes at a 60–100% increased rate than L. There was no L/S enrichment of any genetic pathway or functional category, suggesting that interspecific incompatibility has not played a measurable role in the gene loss of *X. laevis*.

4.1.2 MitoCarta retention rates

Mouse loci identified to be associated with the mitochondria by GFP localization were mapped onto the *X. tropicalis* annotation via BLASTP ($1e^{-10}$, Smith-Waterman refinement) and mapped onto *X. laevis* via orthology. I identified 713 nuclear-encoded mitochondrial genes in *X. laevis*, 395 retain both homeologs, 354 are single-copy (retention rate = 55.3%, $p=2.49e^{-3}$). The mitochondrial genes are reduced to single-copy more often than others. 467 loci are from the L sub-genome, 454 from S. This small difference is not statistically significant, so we cannot argue that there are significant selective pressures from cytonuclear incompatibility between the progenitor species of *X. laevis* contributing to mitochondrial gene loss.

4.1.3 WGCNA retention

To classify the expression patterns of *X. laevis* genes, we analyzed expression variation among homeologous genes. TPM values were calculated by Taejoon Kwon, who wrote code specifically to compute TPM from BWA-mem alignments (unpublished data). Prior to analysis, all TPM values <0.5 were reduced to 0. Genes with no expression values > 0.5 TPM across all experiments were removed from analysis. For developmental expression data we restricted ourselves to 3,797/7,137 homeolog pairs with both genes expressed that showed differential expression in at least one experiment (10x expression difference). For adult data, all 8,374 homeolog pairs with both genes expressed were used to extract module eigengenes. Eigengenes are example “genes” whose expression pattern reflects that of a given group. The observed expression values ($\log_{10}(TPM+0.1)$) for each gene in a homeolog pair were summed in a homeolog expression matrix. We then inferred a weighted undirected co-expression network using the WGCNA method (Langfelder, 2008) with a soft thresholding power of 12 for stage expression data, and 14 for adult data. Next, groups of closely connected genes, or modules, were identified by clustering genes based on the topological overlap matrix and cutting the resulting dendrogram with the cutreeDynamic method in R (parameters: deepSplit=2, pamRespectsDendro=FALSE, minModuleSize=30). Non-module genes were summarized by an artificial “grey” module. Initial modules whose expression profiles were very similar (eigengene correlation >=0.85) were merged. Eigengene expression profiles are visualized in the Appendix Figure 1. For the heatmap visualization in Figure 4.2 the genes were organized by group, and expression patterns were visualized by the heatmap function in R.

Single copy genes, and homeolog pairs that were originally not used in the WGCNA analysis based on expression, were assigned to WGCNA modules by computing a correlation matrix between each gene and the eigengene expression patterns. We then utilized the corPvalueStudent function (with a p-value cutoff of 0.01) of the WGCNA package to test for

significant correlations between genes and eigengenes. If the smallest p-value > 0.01, the gene was assigned to the artificial “grey” module. Table 4.1 contains summaries of the eigengene expression profiles. Co-expression modules revealed a number of unique expression patterns, including identifying a set of neural crest markers in a “brain/kidney” group. There was some overlap with this group and the genes expressed in the mammalian adrenal gland (Lin, 2014), and after this was brought to the attention of collaborators who obtained the RNA samples from different tissues, they reported that the adrenal gland was left in the kidney dissection.

Scatterplots illustrating the retention rates of co-expression modules on the L and S sub-genomes is included in Figure 4.3. Similar to Pfam, GO, and KEGG, the L sub-genome retains genes at a higher rate, and the linear slope of retention reflects that L-bias.

In the developmental stage data, genes whose expression peaks at the maternal-zygotic transition of transcription (MZT, Stages 8-9), and those whose expression peaks at Stage 12 (onset of neuralization) are retained at higher levels ($p < 0.01$). Conversely, those genes whose expression peaks at Stage 40, pre-MZT/Stage 40, and in the oocyte/MZT stages are retained at lower levels.

In the adult, genes expressed across all tissues, and those whose expression peaks in the brain/eye (neural), are retained at higher levels. Conversely, genes whose expression is most important in the adrenal gland, kidney, or in the eye/skin are retained at lower levels. There is a good amount of overlap in the two-copy genes retained in the neuralization and brain groups, which is expected since neural differentiation likely deploys similar loci to an adult brain. We are still working on understanding how the retention rates of different co-expression modules overlap with retention rates of specific functional categories.

4.2 Expression bias between sub-genomes depends on tissue/timepoint

The WGCNA work above classifies genes into groups based on their expression patterns and assess sub-genome dominance by gene retention. Alternatively we can study the L/S expression ratios to ask whether there is a global bias of gene expression, and if so, whether all stages show the same bias. Prior to analysis, all TPM values <0.5 were reduced to 0. Any gene with no expression value > 0.5 was removed from analysis. For each homeolog pair at each tissue and timepoint, L/S expression ratio was calculated and log transformed according to ($\log_{10}(L_{tpm}+0.1/S_{tpm}+0.1)$). The boxplot of expression ratios between sub-genomes is included in Figure 4.4. On average the L sub-genome is expressed higher than the S in all tissues and timepoints, however the magnitude of that differential expression varies. Prior to MZT, and in the adult ovary, genes of the the L sub-genome are expressed 5-7% higher than genes of S, on average. Post-MZT, and in somatic adult tissues, L is expressed 15-17% higher than S on average. These results imply that maternal expression may have a different set of selective pressures than zygotic expression.

Maternal gene expression can be controlled by different promoters than zygotic expression, and if purifying selection is relaxed on both promoters of both homeologs, a potential path to subfunctionalization is one gene becoming maternal-specific, while the other becomes zygotic-specific. I scanned the homeolog pairs for a pattern of one gene being on prior to MZT, and the other completely shut off, while both are on after MZT (example in Figure 4.5). The results for this subfunctionalization analysis are included in Table 4.1. There are 140 homeologs where L is expressed early, and S is not, and 157 where S is expressed early, and L is not. Conversely, there are only 19 homeolog pairs which partition their expression between the embryo and somatic adult tissues (*i.e.*, they have no sub-genome bias). We are currently investigating whether the increased plasticity of maternal expression is due to more rapid turnover of maternal promoters.

4.3 Subfunctionalization of gene expression

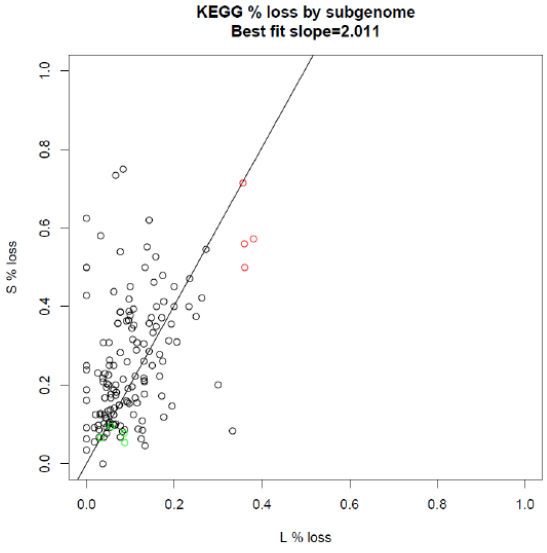
There have been a number of instances of subfunctionalization categorized in *X. laevis* (Hellsten, 2007). Now that the entire genome is sequenced, we can ask questions about the rate at which genes subfunctionalize instead of reporting case studies. Figure 4.6 compares the

correlation distribution for all genes to homeologous gene pairs. Because prior to allotetraploid hybridization the homeologous gene pairs were the orthologous loci in the diploid progenitors, their average correlation is close to 1, as opposed to 0 for any random pairwise correlation. This indicates that most of the homeologous gene pairs have not diverged in their patterns of expression. Some homeologous gene pairs have negative correlations, and are candidates for subfunctionalization (615/7,147 8.6% of homeologs in stage data, 96/8,374 1.1% in tissue data). If subfunctionalization is a mark of fixation of duplication in a genome, then these low rates of subfunctionalization of expression are expected, as the eventual fate of most duplicated loci is to be lost. We are working on studying these divergences in combination with comparing sequences to wild-type *X. laevis* populations to assess whether those genes that show evidence of subfunctionalization of expression domains exhibit increased purifying selection as measured by a McDonald-Kreitman test (Charlesworth, 2008).

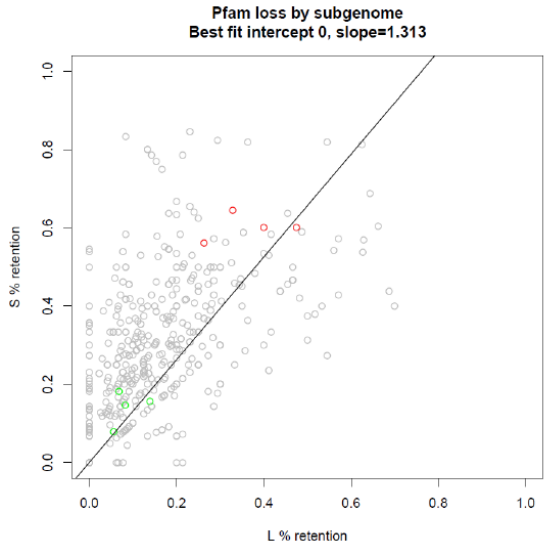
Adult										
Class	Total	2	1	L	S	Fraction Retained	Chi^2	Fraction L retained	Fraction S retained	Chi^2
Brain	348	202	146	271	246	5.8E-01	5.4E-01	7.7E-01	7.1E-01	4.3E-01
Brain_Eye	1359	986	373	1181	1076	7.2E-01	2.9E-05	8.6E-01	7.9E-01	5.8E-02
Muscle_Pancreas	1142	522	620	869	643	4.5E-01	5.6E-09	7.6E-01	5.6E-01	9.1E-03
Eye_Skin	191	66	125	148	84	3.4E-01	5.1E-05	7.7E-01	4.4E-01	4.1E-03
NA	355	243	112	310	266	6.8E-01	2.0E-01	8.7E-01	7.4E-01	8.7E-01
NA	374	94	280	242	155	2.5E-01	8.2E-16	6.4E-01	4.1E-01	7.6E-03
Intestine_Stomach	112	56	56	92	63	5.0E-01	2.3E-01	8.2E-01	5.6E-01	2.2E-01
Brain_Kidney	289	58	231	184	108	2.0E-01	2.3E-16	6.3E-01	3.7E-01	2.9E-03
Heart_Muscle	193	118	75	166	130	6.1E-01	1.0E+00	8.6E-01	6.7E-01	5.5E-01
Intestine	156	62	94	110	79	3.9E-01	4.1E-03	7.1E-01	5.1E-01	3.0E-01
Eye	226	119	107	175	141	5.2E-01	1.8E-01	7.7E-01	6.2E-01	7.1E-01
Intestine_Kidney_Liver	264	114	150	212	137	4.3E-01	1.6E-03	8.1E-01	5.1E-01	1.5E-02
Heart_Muscle_nP	563	228	335	412	304	4.1E-01	6.1E-08	7.3E-01	5.4E-01	7.5E-02
Muscle	132	75	57	109	87	5.6E-01	6.3E-01	8.2E-01	6.5E-01	7.4E-01
Kidney	263	80	183	189	107	3.1E-01	2.0E-08	7.1E-01	4.0E-01	9.9E-04
Ubiquitous	6978	4991	1987	6063	5469	7.1E-01	6.1E-34	8.6E-01	7.8E-01	1.1E-07
Spleen	618	322	296	481	373	5.2E-01	1.5E-02	7.7E-01	6.0E-01	2.2E-01
Embryo										
Class	Total	2	1	L	S	Fraction Retained	Chi^2	Fraction L retained	Fraction S retained	Chi^2
Oocytes	648	375	273	517	441	5.7E-01	9.2E-01	7.9E-01	6.8E-01	6.2E-01
St25	1485	858	627	1215	985	5.7E-01	8.3E-01	8.1E-01	6.6E-01	6.7E-01
Egg_MZT	1384	855	529	1140	966	6.1E-01	1.6E-01	8.2E-01	6.9E-01	5.3E-01
St10	143	99	44	119	113	6.9E-01	2.1E-	8.3E-01	7.9E-01	3.1E-

							01			01
St15	584	308	276	465	353	5.2E-01	1.5E-01	7.9E-01	6.0E-01	2.3E-01
St8_St9_St15_to_St25	360	226	134	309	251	6.2E-01	4.1E-01	8.5E-01	6.9E-01	8.8E-01
St40	226	90	136	169	117	3.9E-01	2.3E-03	7.4E-01	5.1E-01	1.5E-01
St35	541	286	255	421	338	5.2E-01	1.8E-01	7.7E-01	6.2E-01	7.2E-01
St8_St9_St25	284	153	131	225	174	5.3E-01	4.5E-01	7.9E-01	6.1E-01	5.4E-01
St9	877	529	348	714	614	6.1E-01	5.4E-01	8.1E-01	7.0E-01	4.5E-01
Oocyte_MZT	882	401	481	674	514	4.5E-01	1.9E-05	7.6E-01	5.8E-01	1.7E-01
St12	539	386	153	468	425	7.1E-01	1.7E-03	8.6E-01	7.8E-01	1.5E-01
Egg_to_St15	648	348	300	507	418	5.3E-01	2.1E-01	7.8E-01	6.4E-01	1.0
PreMZT_St40	1006	478	528	790	588	4.7E-01	1.3E-04	7.8E-01	5.8E-01	5.1E-02
St8_St9	1326	1026	300	1193	1087	7.7E-01	3.6E-13	9.0E-01	8.2E-01	1.2E-02
St30_35	913	492	421	726	582	5.3E-01	1.4E-01	7.9E-01	6.3E-01	6.1E-01

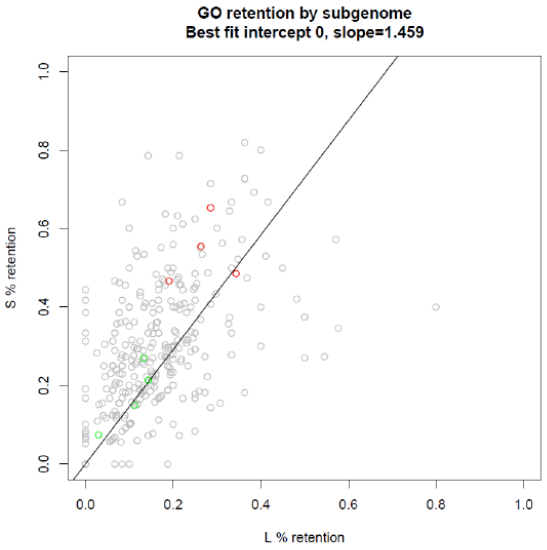
Table 4.1: Summary of eigengene expression profiles
Identification of eigenegene groups is discussed in the main text. Plots of eigengene profiles are included in the appendix, the interpretation of their domains is in column 2. For each group in embryo or adult, we computed the number of *X. laevis* single-copy genes (singletons) vs homeolog pairs and computes a fraction retained. To determine statistical significance we used a Chi-squared test to compare the ratio of singletons:homeolog pairs in each group to other groups in the same data set (embryo vs adult).



Name	# singletons	# homeologous pairs	Fraction loss	p-value
TGF-β signaling	4	57	.066	2.68e-06
Adherens junction	5	52	.088	4.46e-05
Wnt signaling	13	94	.122	1.81e-06
FoxO signaling	16	89	.153	8.4e-05
Whole genome	6,948	9,102	.433	NA
Base excision repair	23	5	.822	1.05e-07
Metabolism of xenobiotics by cytochrome P450	16	5	.762	5.59e-05
RNA polymerase	18	7	.72	1.14e-04
Steroid hormone biosynthesis	25	11	.695	7.78e-06



Name	# singletons	# homeologous pairs	Fraction loss	p-value
Homeobox	22	157	0.123	3.49e-16
PDZ domain	23	86	0.212	3.12e-05
SH2 domain	18	70	.205	1.1e-04
Ras family	29	86	0.253	0.001
Whole genome	6,948	9,102	.433	NA
Cytochrome P450	53	23	.698	2.41e-07
Short chain dehydrogenase	39	18	.685	1.95e-05
Immunoglobulin C1-set domain	29	11	0.725	6.04e-05
50S ribosome-binding GTPase	14	1	0.933	2.73e-05



Name	# singletons	# homeologous pairs	Fraction retained	p-value
nucleus	150	359	.30	1.17e-06
K^+ transport	7	61	.103	7.71e-08
protein-binding	685	1415	.327	1.77e-11
Regulation of transcription, DNA-template	101	388	.207	1.81e-19
Whole genome	6,948	9,102	.433	NA
DNA repair	38	11	.776	1.75e-07
Oxidation-reduction process	202	165	.551	7.89e-09
Methyl transferase	44	20	.6875	4.45e-06
Heme binding	66	44	.6	3.35e-05

Figure 4.1: Gene retention by protein classification
Assignment of protein classifications is discussed in the main text. For each classification (KEGG, Pfam, GO), we computed the number of *X. laevis* single-copy genes (singletons) vs homeolog pairs and computes a fraction retained. To determine statistical significance we used a Chi-squared test to compare the ratio of singletons:homeolog pairs in each group to all other groups (so that those genes without classifications would not be considered in the statistical test). The L% retention is shown on the x-axis of the scatter plots, the S% retention is shown on the y-axis of the scatterplots. The best fit line is forced to go through {0,0}, the theoretical starting point of gene loss. For each classification, a few of the groups that are statistically loss at higher (red) and lower (green) rates are colored and shown in the tables to the right.

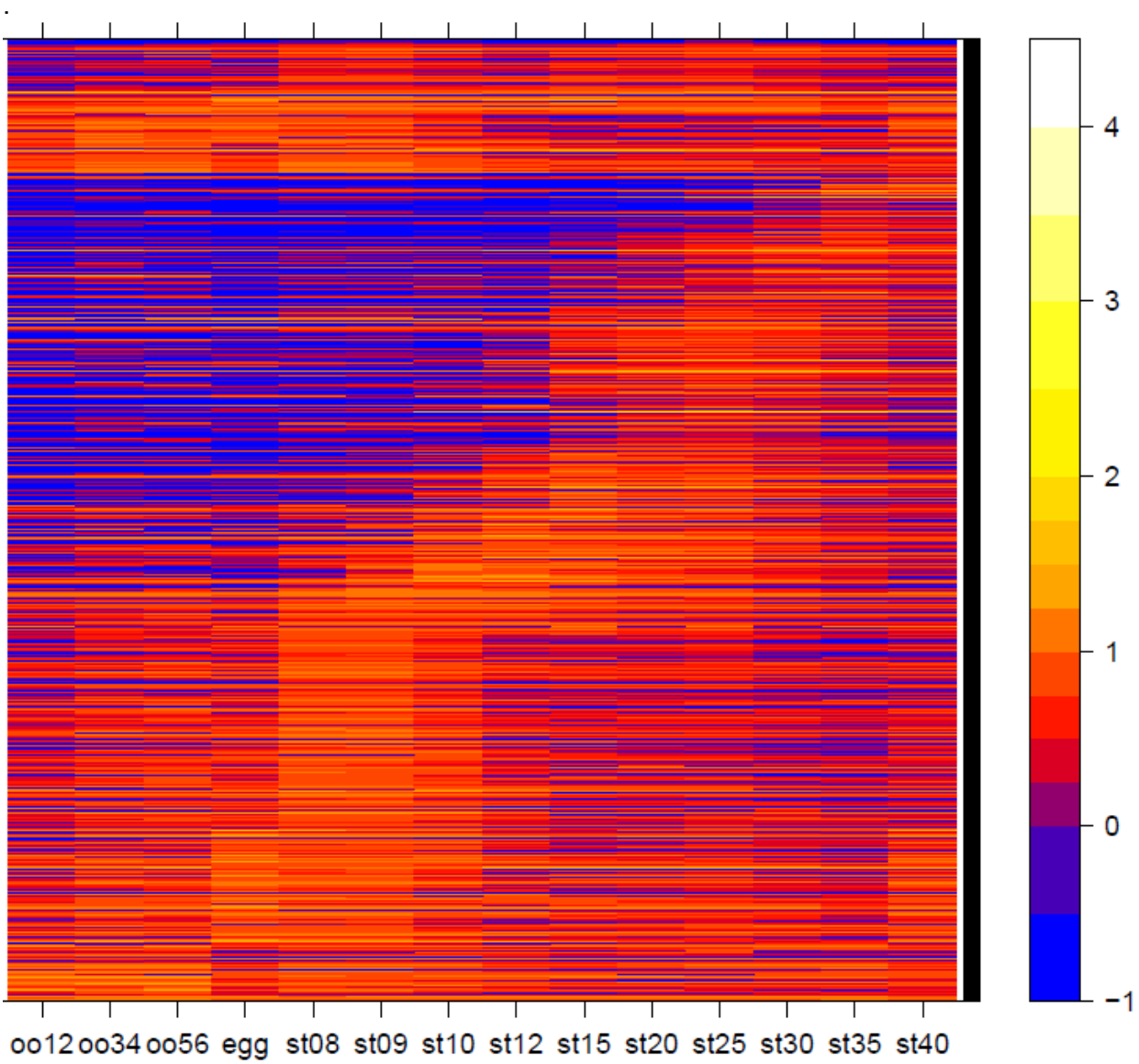


Figure 4.2: Heatmap of developmental WGCNA group expression patterns
Assignment of genes to WGCNA groups is discussed in the main text. Genes were organized by developmental stage, and a heatmap of their $\log_{10}$(TPM+.1) was computed. The red/orange diagonal line represents the progression in expression of each of the groups.

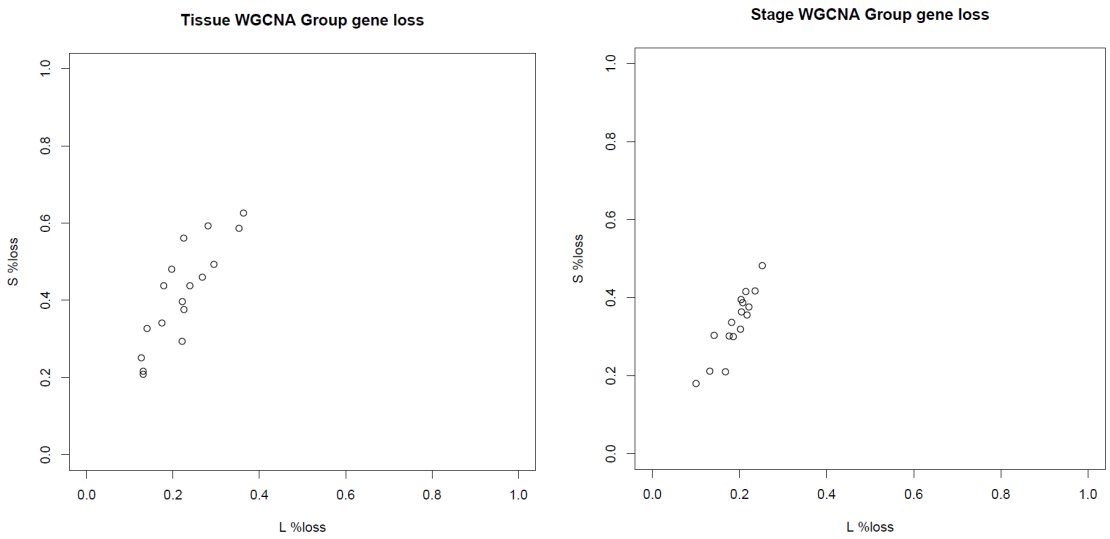


Figure 4.3: Gene retention of WGCNA groups
WGCNA groups were treated similarly to the protein-classifications in Figure 4.1. The L percent-retention is shown on the x-axis of the scatter plots, the S percent-retention is shown on the y-axis of the scatterplots.

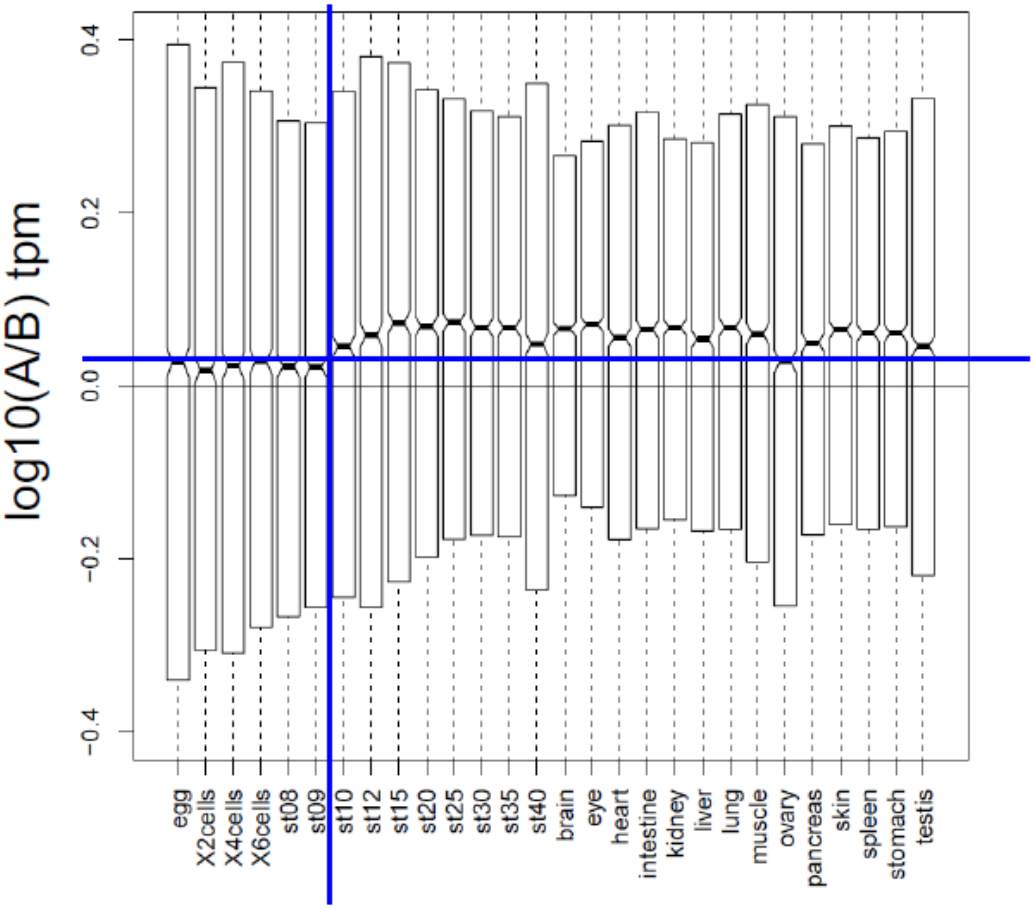


Figure 4.4: Boxplot of homeolog expression ratios

The L/S ratio for TPM values for homeolog pairs that were both expressed at a given stage were computed and log-transformed. The notch in the boxplot represents the median of each distribution. The black line is represents the expected median if there was no difference between sub-genomes. The horizontal blue line is the expression ratio for maternal tissues and time points (prior to MZT) and the vertical blue line separates pre-MZT and MZT from latter expression profiles.

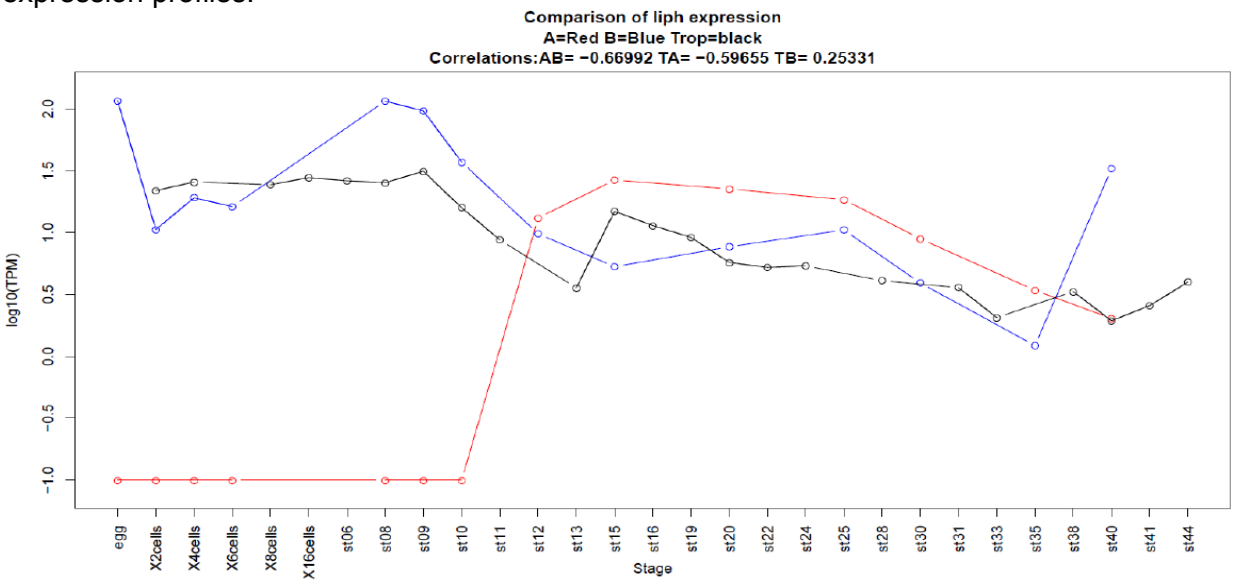


Figure 4.5: Example of maternal differential expression, *liph*
TPM values were calculated by Taejoon Kwon. *X. tropicalis* expression data was taken from Tan et al. 2013 and remapped to the v8 genome. The black line is *X. tropicalis* expression of the *liph* gene, the blue line is S, and the red line is L. Despite similar expression profiles after MZT, the maternal expression has been lost in the L copy.

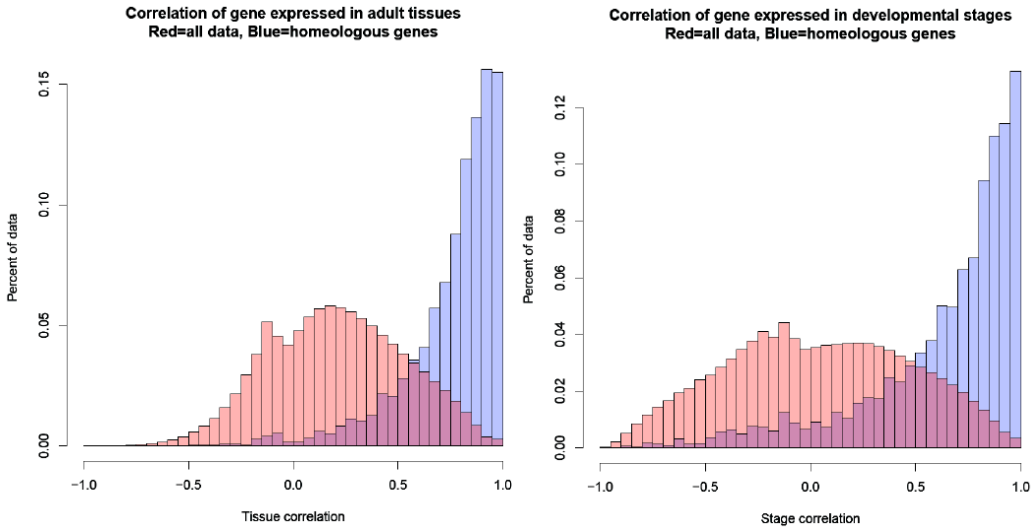


Figure 4.6: Gene expression correlation distributions of all protein-coding genes vs. homeologs in developmental and tissue data
X. laevis TPM values $\leq$ 0.5 were lowered to 0. Any gene with no TPM > 0 was removed from analysis. We then added 0.1 to all TPM values and log transformed ($\log_{10}$). Pairwise pearson correlation values were computed between all genes (red). The correlations of the homeologs were extracted from the matrix, and plotted in blue. The left histogram is for tissue data; right is for developmental data. The x-axis is the correlation; the y-axis is the percent of data. The

homeologous genes have a correlation distribution closer to one due to their being the same locus so recently.

Chapter 5
Analysis of the allohexaploid wheat, *Triticum aestivum*

The *Xenopus* allopolyploidy is of interest to biologists primarily because it serves as a developmental model for vertebrates, and understanding how genes are regulated in *Xenopus* informs us of our own development. In the previous chapters I took advantage of the wealth of experiments performed by *Xenopus* scientists to explore the molecular history of the *X. laevis* genome. Testing that *X. laevis* is an allopolyploid instead of autopolyploid was, in part, inspired by work in plant duplications.

Polyploidy has contributed substantially to the rapid diversification of flowering plants, which are widespread on Earth. Among seed plant species, 35% are polyploids (Comai, 2005). Plants have more mechanisms to form polyploids than vertebrates (Ramsey, 1998), and so autopolyploids are possible. Autopolyploids are unlikely in animals because the reduced fertility and low genetic diversity observed for autopolyploids in plants would be unlikely to form a new radiation to outcompete diploids, without the aid of vegetative expansion. Domesticated plant species that are polyploid may escape these problems by relying on humans. As such humans have domesticated both autopolyploid crops (such as potato, sugarcane and banana), and allopolyploid crops (including wheat, cotton, tobacco, strawberry, and oilseed rape). Applying similar methods done for *Xenopus* above to these plants will allow us to study their molecular history. Heterosis, or hybrid vigor, is defined as the improved function of a hybrid over its progenitors. While we can assume the *X. laevis* tetraploid ancestor outcompeted its diploid progenitors—at least, there are no surviving diploid Xenopus with 9 pairs of chromosomes—there are no annotated cases of heterosis in *Xenopus* involving higher ploidy. Studying the genomes of polyploid crops might reveal shared molecular mechanisms of heterosis between different domesticated crops.

Hexaploid bread wheat (*Triticum aestivum* L., 1C = 16 Gbp, 2n = 6x = 42) is one of the most important agricultural crops, whose genomic history is summarized in Figure 5.1. There are two wheat genomes analyzed in this chapter: Chinese Spring is a naturally occurring strain of hexaploid wheat sequenced by the IWGSC. ‘Synthetic W7984’ was generated by crossing a tetraploid wheat AABB genome with the diploid DD genome, followed by chromosome doubling, resulting in a contemporary reconstitution of hexaploid wheat. [REF TO CHAPMAN et al. GENOME BIOLOGY] This chapter details the beginning of my work to understand the molecular evolution of the *T. aestivum* genome.

5.1 Identification of triplet genes in a draft assembly of *T. aestivum*

To assess the completeness of the genome assembly with respect to known transcribed sequence, we used a collection of 6,137 flcDNAs in the ‘Triticeae full length cDNA database’ from *T. aestivum* cultivar ‘Chinese Spring’ generated by Mochida et al. (Mochida, 2009). These flcDNAs (“full length cDNAs”) are from hexaploid bread wheat and are expected to match the W7984 assembly with the exception of intra-specific polymorphisms and presence/absence or copy number variation. In contrast, they are expected to match the Chinese Spring assembly identically. We used flcDNA rather than short-read RNAseq because the cDNA data are longer, of higher quality, and as clones are not subject to confounding effects arising from attempting to assemble homeologs in distinct scaffolds. We cleaned the flcDNAs by (1) trimming polyA tails with BioPerl ‘TrimEST’; (2) identifying non-wheat contaminations, using BLAST (Atschul, 1997); and (3) identifying putative transposable elements by comparison with RepBase (Jurka, 2005).

We identified three *T. aestivum* flcDNAs in GenBank as being in fact human sequences (RFL_Contig2039, 3209, and 5006) showing near 100% identity to human genes. These are presumably low-level contaminants of the wheat cDNA libraries. These sequences were

excluded from further consideration. We found 99 *T. aestivum* flcDNAs from the Mochida et al. set (99/6,137 = 1.6%) with substantial BLAST alignments (BLASTN default word size, e-10, no DUST filter; >90% identity over >50% of their length) to RepBase entries. These were considered to be transposable elements and not considered in subsequent analyses. To identify other likely non-wheat contaminations in Mochida et al. (Mochida, 2009), we used BLASTN (e-10, no DUST filter; >90%) versus the GenBank non-redundant nucleotide database, and excluded from further consideration flcDNA sequences that (a) had no alignment to both assemblies (>80% length, 1e-10) and (b) did not hit grass sequences in GenBank (>90% identity, >10% length). We found 52 flcDNA sequences that did not align to either assembly. Of these, 17 had alignments to grasses and were kept in further analyses; 32 had no GenBank hits to plants; 3 had only weak hits to non-grasses. These last two categories were not considered further.

Thus, after filtering for contaminants and transposons we consider 6,000 known, non-transposon *T. aestivum* flcDNAs = (6,137 initial flcDNA from Mochida *et al.*) - (99 RepBase transposon-related) - (3 human contamination) - (35 likely non-grass contamination not found in either assembly). We also identified flcDNAs that have 10 or more alignments (>80% identity, >50% length) to one or both of the hexaploid wheat assemblies (126 to W7984, 198 to ‘Chinese Spring’). These are also likely to be repetitive elements, but may include recently diverged large gene families. These are included in all analyses. Non-transposon, non-contaminant cDNA sequences were aligned to both the Meraculous W7984 WGS assembly database and to the IWGSC chromosome sorted ‘Chinese Spring’ assembly database with BLAST (BLASTN default word size, e-10, no DUST filter), initially requiring >80% identity over >50% of the cDNA or mRNA length. The high-scoring pairs (HSPs) of cDNAs aligned to genomic sequence correspond to exons, and minimally overlapping HSPs to a given scaffold were combined to produce a single percentage coverage (Total bases aligned/Total bases in cDNA) and percentage identity (Total positions matched/Total aligned positions excluding gaps). The percent identity distributions of cDNAs between sub-genomes is shown in Figures 5.2 and 5.3.

The distributions in 5.2 and 5.3 illustrate that the wheat genome has been properly assembled by multiple methods (whole-genome shotgun and chromosome capture), and that we can identify homeologous genes between sub-genomes. We currently lack the statistical power to test if the subtle shifts in percent identity between sub-genomes can tell us which are most similar. A whole-genome annotation would be useful in identifying all homeolog pairs, and performing similar analysis as was done for *X. laevis* in chapters 2–4.

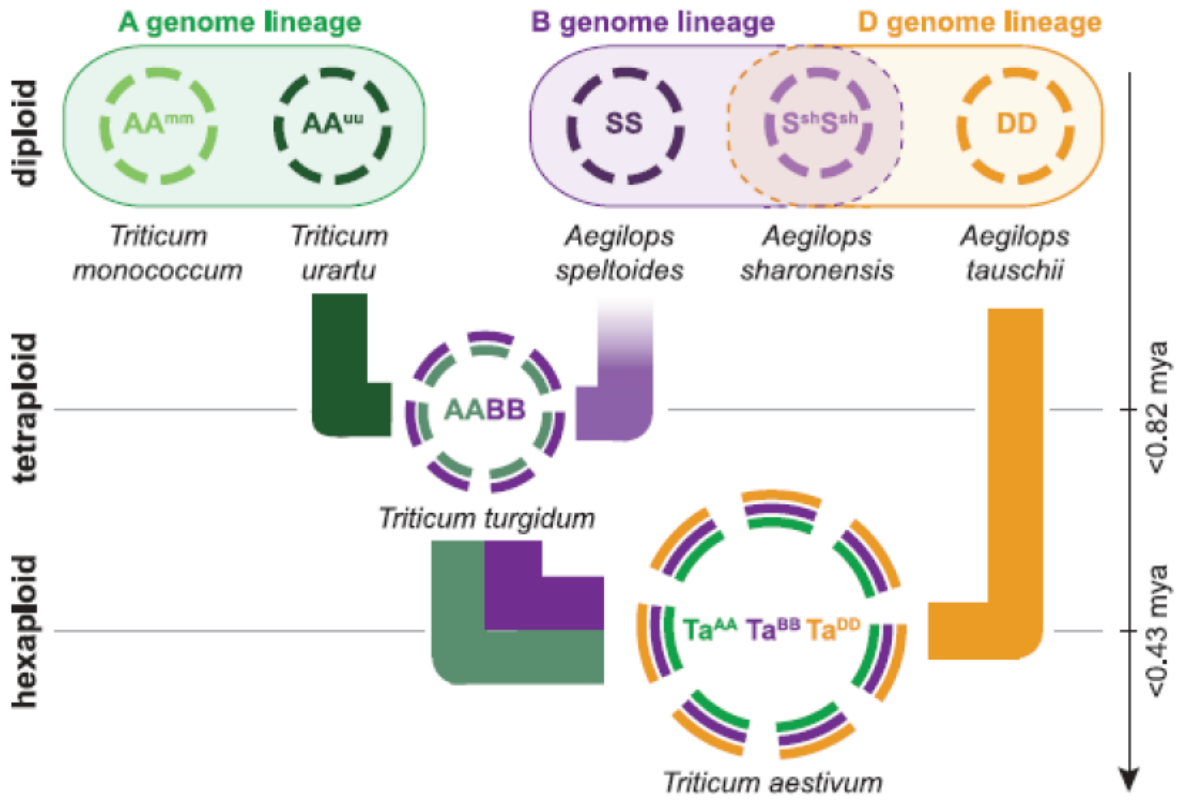


Figure 5.1: Diagram of T. aestivum polyploid evolution history.
Taken from Mayer et al. 2014 (IWGSC). Two diploid grasses, hypothesized to be *Triticum urartu* and *Aegilops speltoides* underwent allopolyploidy ~800,000 years ago to form *Triticum turgidum*. *T. turgidum* then underwent an allopolyploidy event by mating with *Aegilops tauschii* ~430,000 years ago.

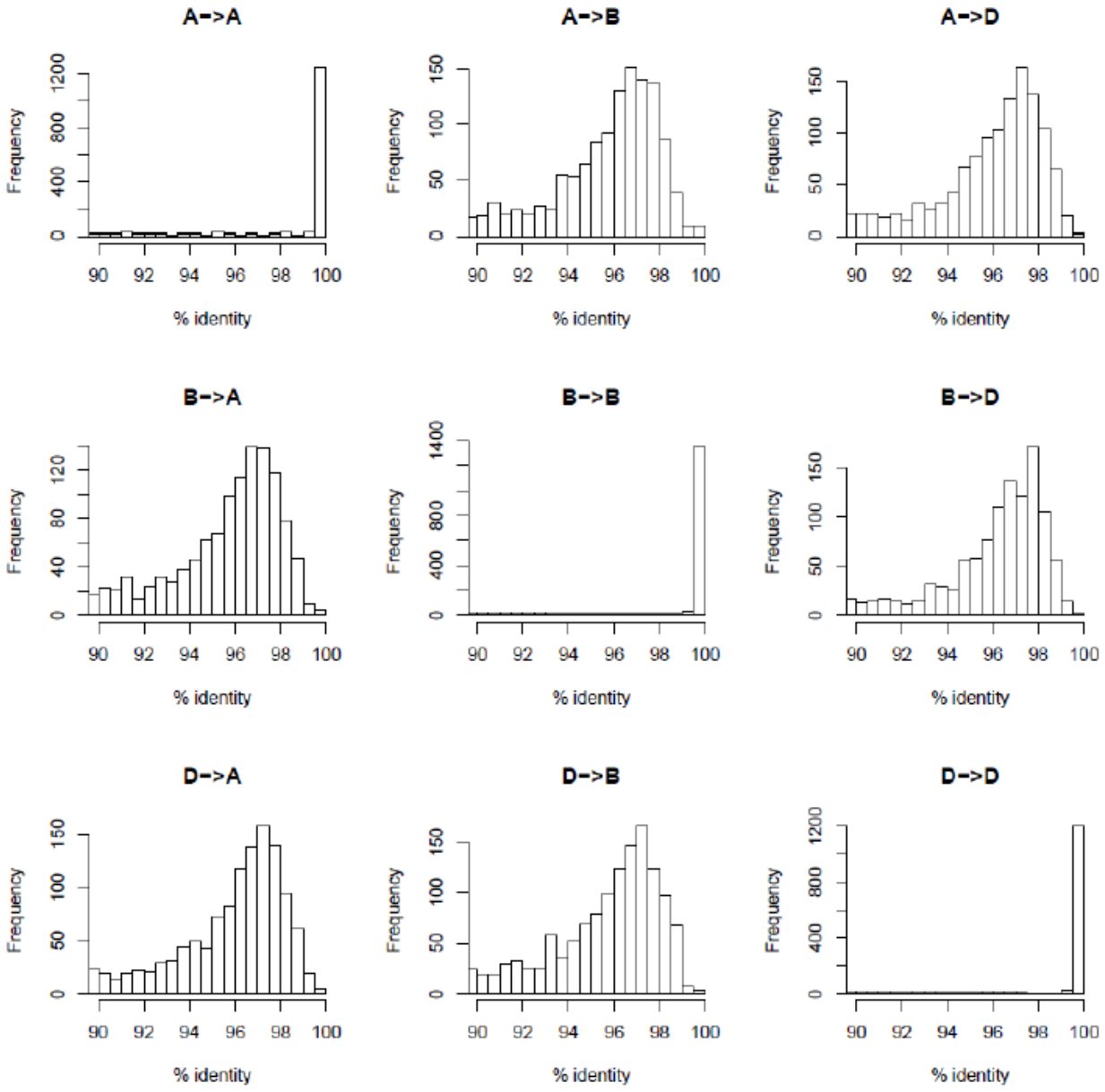


Figure 5.2: Chinese Spring cDNA percent identity between sub-genomes
Previously published in Chapman et al. 2015. Full-length Chinese spring cDNAs were aligned to the Chinese Spring assembly (BLASTN default word size, e-10, no DUST filter; >90% identity over >50% of their length). We assigned loci to one of the sub-genomes using the genetic anchoring of the assembly. This plot shows the distribution of nucleotide identity between cDNAs assigned to the A, B and D sub-genomes and their best BLAST hit in the other two sub-genomes (that is, to their putative homeologous loci).

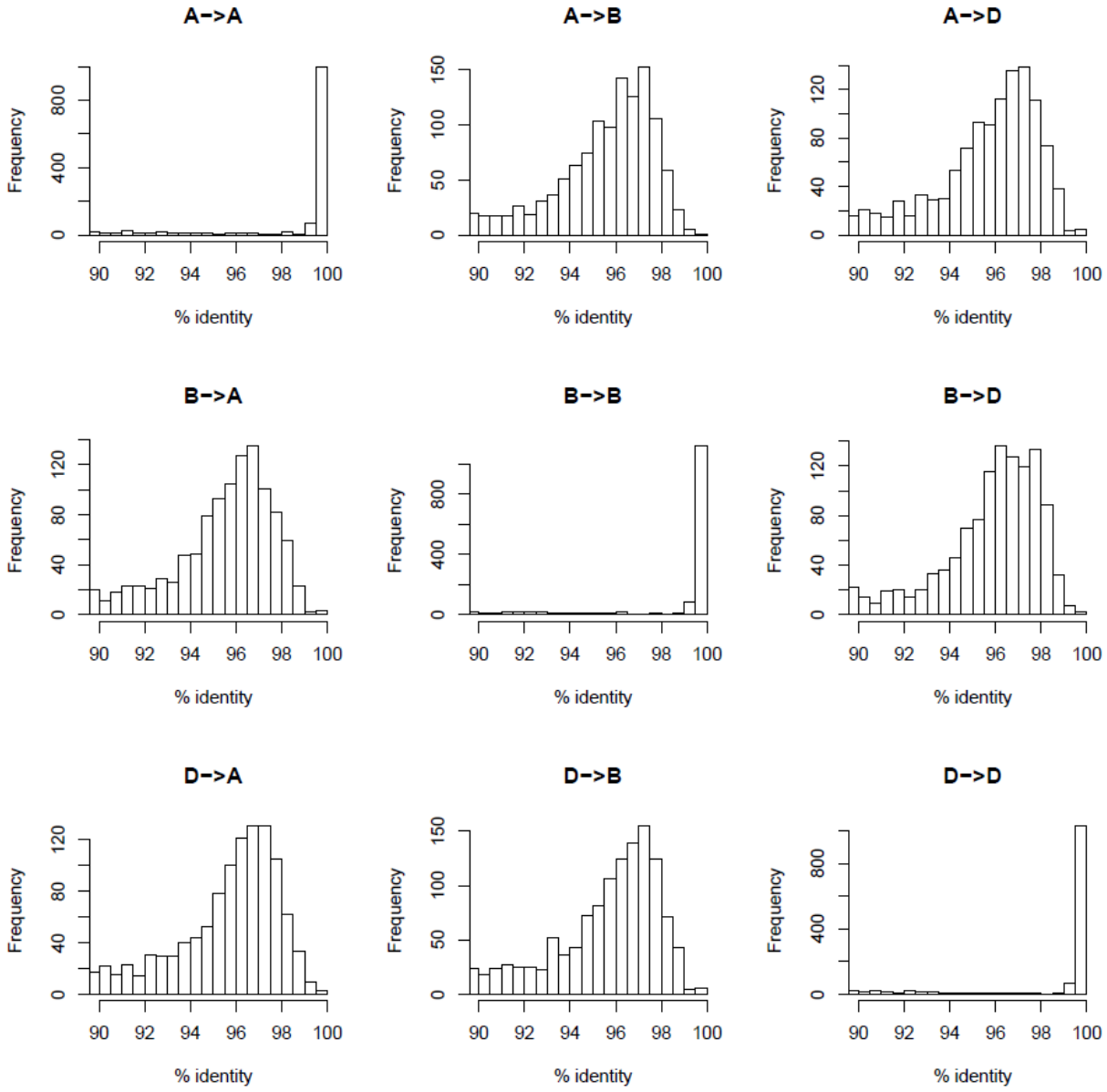


Figure 5.3: w7984 cDNA percent identity by sub-genome
Previously published in Chapman et al. 2015. Full-length Chinese spring cDNAs were aligned to the w7984 assembly (BLASTN default word size, e-10, no DUST filter; >90% identity over >50% of their length). We assigned loci to one of the sub-genomes using the genetic anchoring of the assembly. This plot shows the distribution of nucleotide identity between cDNAs assigned to the A, B and D sub-genomes and their best BLAST hit in the other two sub-genomes (that is, to their putative homeologous loci).

Chapter 6
Conclusion

The goal of my dissertation research has been to annotate and analyze the genome of *X. laevis*, and the explore to explore the nature of its allotetraploidy. Prior this analysis, extant diploid species were required to understand the speciation history of polyploid organisms (Gill, 2009). While we are still working to make sure the time measurements between repeats and pseudogenes are calibrated to the same units, the fact that molecular analysis can be used to predict the distribution of transposable elements by FISH is an amazing illustration of the molecular history recorded within chromosome sequences. As C.D. Darlington wrote, "The chromosomes provide us with a record of the past, a living record, significant in a surprisingly similar way to the dead record which fossils provide for the paleontologist." (Chromosome Botany and the Origins of Cultivated Plants 1963). In the case of *Xenopus*, where the high level of similarity between skeletons makes the fossil record less useful for determining divergence times, the molecular history recorded by genomes gives us a view into their recent activity.

The duplication of an entire genome is a spectacular natural experiment in which tens of thousands of genes are effectively duplicated synchronously, so that each gene has a matched homeologous partner with a highly similar or identical sequence, expression domains, and chromosomal context. Subsequent divergence, loss, and rearrangement then gradually erode the signs of duplication. Polyploidy can be a powerful evolutionary force, but the polyploidies that occurred early in the vertebrate and teleost lineages are so ancient (~500 Mya and ~350 Mya, respectively) that the immediate evolutionary response is obscured in modern genomes. The allopolyploid genome of *X. laevis* underwent hybridization so recently that it allows us to study the early genomic response to polyploidy. The early gene loss seems to follow similar patterns as other animal duplications, however we have yet to study what genes are currently under selection in *X. laevis*. Unfortunately the nature of genome assembly makes this difficult, choosing to sequence an inbred population of a lab animal means any variation at the population level of wild *X. laevis* has been lost over the 30+ generations of inbreeding. In Chapter 1 I discussed balance of gene dosage as a possible driving force of gene retention. The two gene “copies” obtained from each species are actually four “alleles”, one from each progenitor species per parent. It possible that many loci are under selection to keep three or four alleles active, and in the wildtype population there are nonfunctionalized loci segregating at a low rate. Inbreeding of laboratory animals has been shown to cause unpredictable gene loss (Warringer, 2011). It is possible that the signal of selection for many genes was lost by inbreeding of the J strain. We are currently crossing wild frogs to build a genetic map for *Xenopus.* The extra advantage of wild frogs is that we will be able to assess what genes are currently under selection in wild *Xenopus* to predict what will be loss, and study the effects of inbreeding on an allopolyploid genome.

While we briefly discussed shared gene loss trends across polyploid groups, it is beyond the scope of this thesis to discuss what drives the differences in gene loss between polyploid phyla or what shared phenotypes allow for polyploidy. Flowering plants, *Paramecium*, yeast, and vertebrates have all had ancient and recent polyploidy events, but what about these groups makes polyploidy so prevalent, when it seems so rare elsewhere? Even within vertebrates polyploidy appears more common in amphibians and fish than in mammals and reptiles. Is there a feature of amniotes that restricts the possibility of polyploidy? These questions about the nature of polyploidy are related to the biological mysteries surrounding meiosis and speciation. It is generally accepted that one positively selected aspect of sexual reproduction is that it prevents the accumulation of deleterious mutations across generations (Muller, 1964). Sexual reproduction, in conjunction with speciation, allows for whole genome duplications through

allotetraploidy. This hybridization of DNA from different species is similar to horizontal gene transfer seen in prokaryotes. Polyploidization could be a similar process to horizontal gene transfer, allowing phenotypes of closely related species to be mixed to create a novel organism with potentially beneficial phenotypes.

The differences in gene loss between phyla are interesting as well, and could be attributed to their organismal differences. The plant and animal lineages diverged about 1.5 billion years ago (Douzery, 2004). They have evolved their multicellular organization independently but using the same initial tool kit—the set of genes inherited from their common unicellular eukaryotic ancestor. Most of the contrasts in their development come from photosynthesis and semi-rigid cell walls in plants. This dictates a body plan different from that of animals, which typically ingest other organisms and have more cell movements, such as cortical rotation, that are essential to proper development (Gerhart, 1989). In addition, animal development is largely buffered against environmental changes. Because they cannot interact with their environment by moving, plants adapt instead by opportunistically altering the course of their development (Reeves, 2000). A given type of organ—a leaf, flower, or root—can be produced from the fertilized egg by many different paths according to environmental cues. Although the developmental path of a plant varies, its structure at the organ level does not. A leaf, a flower, or indeed an early plant embryo, is as precisely specified as any organ of an animal, possessing a determinate structure, in contrast with the indeterminate pattern of branching and sprouting of the plant as a whole. Although both plants and animals would be sensitive to changes to interaction networks during their development, the differences in growth and regulation between the groups may lead to different evolutionary outcomes. Indeed, the study of polyploidy may reveal unique signatures of developmental control between phyla. The current explosion of genome sequences of polyploid organisms provides an exciting resource to study the mechanisms of genome evolution.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410.
2. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N et al: Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 2006, 444(7116):171-178.
3. Bailey JA, Liu G, Eichler EE: An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 2003, 73(4):823-834.
4. Bao R, Friedrich M: Molecular evolution of the Drosophila retinome: exceptional gene gain in the higher Diptera. Mol Biol Evol 2009, 26(6):1273-1287.
5. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A et al: The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 2014, 5:3657.
6. Birchler JA, Yao H, Chudalayandi S, Vaiman D, Veitia RA: Heterosis. Plant Cell 2010, 22(7):2105-2112.
7. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y: The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biology 2006, 7(5).
8. Boveri T: Über Mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. Verhandlungen der Physikalische-medizinischen. Gesellschaft zu Würzburg 1902, 35:67-90.
9. Boveri T: In: Die Entwicklung dispermer Seeigeleier Ein Beitrag zur Befruchtungslehre und zur Theorie des Kerns Gustav. Zellenstudien VI, editor Fischer; Jena 1907.
10. Bridgham JT, Ortlund EA, Thornton JW: An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 2009, 461(7263):515-519.
11. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997, 268(1):78-94.
12. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: AmiGO: online access to ontology and annotation data. Bioinformatics 2009, 25(2):288-289.
13. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M et al: Conservation of core gene expression in vertebrate tissues. J Biol 2009, 8(3):33.
14. Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS: Meraculous: de novo genome assembly with short paired-end reads. PLoS One 2011, 6(8):e23501.
15. Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L et al: A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biology 2015, 16(1):26.
16. Charif D, Lobry JR: Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis.. In.; 2007.
17. Charlesworth J, Eyre-Walker A: The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol 2008, 25(6):1007-1015.
18. Chou HH, Hayakawa T, Diaz S, Krings M, Indriati E, Leakey M, Paabo S, Satta Y, Takahata N, Varki A: Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc Natl Acad Sci U S A 2002, 99(18):11736-11741.
19. Comai L: The advantages and disadvantages of being polyploid. 2015.
20. Crick FHC: On Protein Synthesis. Symp Soc Exp Biol XII, 1956:139-163.
21. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G et al: The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 2014, 345(6201):1181-1184.

22. Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H: The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc Natl Acad Sci U S A 2004, 101(43):15386-15391.
23. Duret L, Galtier N: Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 2009, 10:285-311.
24. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32(5):1792-1797.
25. HAY ED, GURDON JB: Fine Structure of the Nucleolus in Normal and Mutant Xenopus Embryos. 1967.
26. Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA: Integrating Diverse Datasets Improves Developmental Enhancer Prediction. PLoS Comput Biol 2014, 10(6).
27. Ferris SD, Whitt GS: Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol 1979, 12(4):267-317.
28. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al: Pfam: the protein families database. Nucleic Acids Res 2014, 42(Database issue):D222-230.
29. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G et al: The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. 2014.
30. Gall JG, Pardue ML: Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc Natl Acad Sci U S A 1969, 63(2):378-383.
31. Gerhart J: The primacy of cell interactions in development. Trends Genet 1989, 5(8):233-236.
32. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004, 428(6982):493-521.
33. Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA: Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol 2009, 151(3):1167-1174.
34. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011, 29(7):644-652.
35. Mendel G: Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver Brünn 1866, 4:3-47.
36. Gurdon JB, Hopwood N: The introduction of Xenopus laevis into developmental biology: of empire, pregnancy testing and ribosomal genes. Int J Dev Biol 2000, 44(1):43-50.
37. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 2008, 9(1):R7.
38. KOBEL HR, PASQUIER LD: Hyperdiploid species hybrids for gene mapping in Xenopus. 1979, 279(5709):157-158.
39. Hartsoeker N: Essay de dioptrique. Par Nicolas Hartsoeke 1694.
40. He X, Zhang J: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 2005, 169(2):1157-1164.
41. Izutsu Y, Yoshizato K: Metamorphosis-dependent recognition of larval skin as non-self by inbred adult frogs (Xenopus laevis). J Exp Zool 1993, 266(2):163-167.
42. WATSON JD, CRICK FHC: Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. 1953, 171(4356):737-738.
43. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al: Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 2004, 431(7011):946-957.

44. Bridgham JT, Ortlund EA, Thornton JW: An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 2009, 461(7263):515-519.
45. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005, 110(1-4):462-467.
46. JR, Schemske DW: PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. http://dxdoiorg/101146/annurevecolsys291467 2003.
47. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014, 42(Database issue):D199-205.
48. Katagiri C: Xenopus laevis as a model for the study of immunology. Dev Comp Immunol 1978, 2(1):5-13.
49. Kobel HaD, L.: Genetics of polyploid Xenopus. Trends in Genetics 1986, 2:310-315.
50. Kondrashov AS, Sunyaev S, Kondrashov FA: Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 2002, 99(23):14878-14883.
51. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.
52. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9:559.
53. Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B: Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Mol Biol Evol 2011, 28(3):1205-1215.
54. Wolpert L, Beddington R, Jessell T, Lawrence P, Meyerowitz E, Smith J: Principles of Development, Second Edition: Oxford University Press.; 2002.
55. Li W: Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution 1993, 36(1):96-99.
56. Lin YC, Boone M, Meuris L, Lemmens I, Van Roy N, Soete A, Reumers J, Moisse M, Plaisance S, Drmanac R et al: Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nat Commun 2014, 5:4767.
57. Lopez-Flores I, Garrido-Ramos MA: The repetitive DNA content of eukaryotic genomes. Genome Dyn 2012, 7:1-28.
58. Lund E, Liu M, Hartley RS, Sheets MD, Dahlberg JE: Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. RNA 2009, 15(12):2351-2363.
59. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science 2000, 290(5494):1151-1155.
60. Lynn DJ, Singer GA, Hickey DA: Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 2002, 30(19):4272-4277.
61. Ming R, Man Wai C: Assembling allopolyploid genomes: no longer formidable. Genome Biology 2015, 16(1):27.
62. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S et al: The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 2015, 43(Database issue):D213-221.
63. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K: TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol 2009, 150(3):1135-1146.
64. Nakamura T: Lethal graft-versus-host reaction induced by parental cells in the clawed frog, Xenopus laevis. Transplantation 1985, 40(4):393-397.
65. Nei M, Gu X, Sitnikova T: Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 1997, 94(15):7799-7806.

66. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K et al: The amphioxus genome and the evolution of the chordate karyotype. Nature 2008, 453(7198):1064-1071.
67. Nieuwkoop P, Faber J: Normal table of Xenopus Laevis. Amsterdam, The Netherlands: North Holland Publishing Co; 1967.
68. OHNO S: Evolution by gene duplication.: London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.
69. Otto SP: The Evolutionary Consequences of Polyploidy. Cell 2007, 131(3):452-462.
70. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK et al: A mitochondrial protein compendium elucidates complex I disease biology. Cell 2008, 134(1):112-123.
71. Paradis E, Claude J, Strimmer K: APE: analyses of phylogenetics and evolution in R language. In. Bioinformatics 20: 289-290; 2004.
72. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D: Cactus: Algorithms for genome multiple sequence alignment. Genome Res 2011, 21(9):1512-1528.
73. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK et al: The amphioxus genome and the evolution of the chordate karyotype. Nature 2008, 453(7198):1064-1071.
74. Putnam NH, O'Connell B, Stites JC, Rice BJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE: Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. 2015.
75. Reeves PH, Coupland G: Response of plant development to environment: control of flowering by daylength and temperature. Curr Opin Plant Biol 2000, 3(1):37-42.
76. Bao R, Friedrich M: Molecular Evolution of the Drosophila Retinome: Exceptional Gene Gain in the Higher Diptera. 2009.
77. Salamov AA, Solovyev VV: Ab initio gene finding in Drosophila genomic DNA. Genome Res 2000, 10(4):516-522.
78. Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH: Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A 2007, 104(20):8397-8402.
79. Carroll SB, Grenier JK, Weatherbee SD: From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design: Wiley-Blackwell; 2004.
80. Slater GS, Birney E: Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 2005, 6:31.
81. Smit A, Hubley R, Green P: RepeatMasker Open-4.0. In.; 2015.
82. Spemann H, Mangold H: Induction of embryonic primordia by implantation of organizers from a different species. Roux's Arch Entw Mech 1924, 100:599-638.
83. Spirek M, Polakova S, Jatzova K, Sulo P: Post-zygotic sterility and cytonuclear compatibility limits in S. cerevisiae xenomitochondrial cybrids. Front Genet 2014, 5:454.
84. Subramanian AR, Kaufmann M, Morgenstern B: DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms for Molecular Biology 2008, 3(1):6.
85. Sutton WS: On the morphology of the chromosome group in Brachystola magna. Biol Bull 1902, 4:24-39.
86. Talavera G, Castresana J: Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007, 56(4):564-577.
87. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013, 30(12):2725-2729.
88. Tan Y, Li WH: Trichromatic vision in prosimians. Nature 1999, 402(6757):36.
89. Tanksley SD: Mapping polygenes. Annu Rev Genet 1993, 27:205-233.

90. Thiebaud CH, Fischberg M: DNA content in the genus Xenopus. Chromosoma 1977, 59(3):253-257.
91. TOCHINAI S, Zoological Institute FoS, Hokkaido University, Sapporo 060, Japan, KATAGIRI C, Zoological Institute FoS, Hokkaido University, Sapporo 060, Japan: COMPLETE ABROGATION OF IMMUNE RESPONSE TO SKIN ALLOGRAFTS AND RABBIT ERYTHROCYTES IN THE EARLY THYMECTOMIZED XENOPUS. Development, Growth & Differentiation 2015, 17(4):383-394.
92. Tymowska J, Fischberg M: A comparison of the karyotype, constitutive heterochromatin, and nucleolar organizer regions of the new tetraploid species Xenopus epitropicalis Fischberg and Picard with those of Xenopus tropicalis Gray (Anura, Pipidae). Cytogenet Cell Genet 1982, 34(1-2):149-157.
93. Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L et al: The Genome of the Western Clawed Frog Xenopus tropicalis. 2010.
94. Uno Y, Nishida C, Takagi C, Ueno N, Matsuda Y: Homoeologous chromosomes of Xenopus laevis are highly conserved after whole-genome duplication. Heredity (Edinb) 2013, 111(5):430-436.
95. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H et al: MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 2012, 40(7):e49.
96. Warringer J, Zorgo E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ et al: Trait variation in yeast is defined by population history. PLoS Genet 2011, 7(6):e1002111.
97. Weismann A: Das Keimplasma. Das Keimplasma 1892:XVIII, 628 S.
98. Wells DE, Gutierrez L, Xu Z, Krylov V, Macha J, Blankenburg KP, Hitchens M, Bellot LJ, Spivey M, Stemple DL et al: A genetic map of Xenopus tropicalis. Dev Biol 2011, 354(1):1-8.
99. Wertheim B, Beukeboom L, van de Zande L: Polyploidy in Animals: Effects of Gene Expression on Sex Determination, Evolution and Ecology. Cytogenetics Genome Research 2013, 140:256-259.
100. Wilcoxon F: Individual comparisons by ranking methods. Biometrics Bulletin 1945, 1(6):80-83.
101. Yao H, Kato A, Mooney B, Birchler JA: Phenotypic and gene expression analyses of a ploidy series of maize inbred Oh43. Plant Mol Biol 2011, 75(3):237-251.
102. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M: Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 2010, 11(3):R26.
103. Zhao S, Shetty J, Hou L, Delcher A, Zhu B, Osoegawa K, de Jong P, Nierman WC, Strausberg RL, Fraser CM: Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. Genome Res 2004, 14(10A):1851-1860.

Appendix

BC077553.1	S. mansoni
BC084798.1	C. sinensis
BC075206.1	S. japonicum
BC080121.1	S. japonicum
BC081209.1	C. sinensis
BC081212.1	C. sinensis
BC081283.1	C. sinensis
BC098177.1	M . musculus
BC091629.1	H. influenza
BC095914.1	M. musculus
BC090151.1	M. musculus
BC090157.1	M. musculus
BC093565.1	S. japonicum
BC092154.1	S. mansoni
BC093564.1	C. sinensis
BC100272.1	P. infestans
BC099272.1	D. rerio
BC106387.1	Drosophila
BC110760.1	Drosophila
BC108599.1	Drosophila
BC123160.1	Ectocarpus siliculosus

BC123254.1	Salmo salar
BC129579.1	Batrachochytrium dendrobatidis
BC129580.1	Monosiga
BC129584.1	Glycine max
BC129592.1	Acromyrmex echinatior
BC130140.1	M. musculus
BC130142.1	M. musculus
BC130143.1	M. musculus
BC130145.1	M. musculus
BC130146.1	M. musculus
BC130149.1	M. musculus
BC130150.1	M. musculus
BC133207.1	M. musculus
BC130205.1	M. musculus
BC154967.1	Capsaspora owczarzaki
BC152711.1	Danio rerio
BC158211.1	Danio rerio
BC153790.1	Styela clava
BC157719.1	Capsaspora owczarzaki
BC155884.1	Schizosaccharomyces japonicus
BC153801.1	Capsaspora owczarzaki
BC154993.1	Thielavia terrestris

Appendix Table 1: List of NCBI full-length cDNA contaminants for *X. laevis*
NCBI identifier of cDNA in column 1. The best nr BLAST hit in column 2.

Series name	Sample name	Total reads	Filtered	good pct
Taira201203_stage	Taira201203_XENLA_egg	38,348,636	37,229,374	97.1
	Taira201203_XENLA_st08	77,099,966	64,253,638	83.3
	Taira201203_XENLA_st09	79,879,478	66,485,248	83.2
	Taira201203_XENLA_st10	65,948,718	54,834,640	83.1
	Taira201203_XENLA_st12	67,699,896	56,479,540	83.4
	Taira201203_XENLA_st15	41,491,890	34,497,488	83.1
	Taira201203_XENLA_st20	68,343,338	67,914,166	99.4
	Taira201203_XENLA_st25	86,811,640	86,280,470	99.4
	Taira201203_XENLA_st30	105,954,110	105,311,686	99.4
	Taira201203_XENLA_st35	84,960,480	84,444,276	99.4
	Taira201203_XENLA_st40	102,731,980	102,113,322	99.4
	Subtotal	81,927,013,200	75,984,384,800	
	Sample name	Total reads	Filtered	good pct
Taira201203_tissue	Taira201203_XENLA_brain	58,181,568	57,392,896	98.6
	Taira201203_XENLA_eye	69,984,646	69,071,260	98.7
	Taira201203_XENLA_heart	62,732,656	61,896,004	98.7
	Taira201203_XENLA_intestine	64,925,678	62,993,848	97
	Taira201203_XENLA_kidney	78,965,634	77,925,086	98.7
	Taira201203_XENLA_liver	77,762,686	71,226,866	91.6
	Taira201203_XENLA_lung	116,208,352	112,416,810	96.7
	Taira201203_XENLA_muscle	62,909,858	58,058,622	92.3
	Taira201203_XENLA_ovary	101,556,142	93,170,430	91.7
	Taira201203_XENLA_pancreas	58,989,614	57,159,656	96.9
	Taira201203_XENLA_skin	57,287,468	55,545,858	97
	Taira201203_XENLA_spleen	68,752,594	63,630,312	92.5
	Taira201203_XENLA_stomach	128,739,570	127,029,110	98.7
	Taira201203_XENLA_testis	77,904,524	71,542,888	91.8
	Subtotal (bp)	108,490,099,000	103,905,964,600	
	Sample name	Total reads	Filtered	good pct
Ueno201210_stage	Ueno201210_XENLA_egg	60,114,334	59,935,632	99.7
	Ueno201210_XENLA_2cells	147,279,386	146,827,618	99.7
	Ueno201210_XENLA_4cells	70,384,748	70,153,484	99.7
	Ueno201210_XENLA_6cells	83,490,188	83,224,086	99.7
	Ueno201210_XENLA_st08	44,961,078	44,804,788	99.7
	Ueno201210_XENLA_st09	61,674,928	61,482,438	99.7
	Ueno201210_XENLA_st10	47,722,386	47,547,154	99.6
	Ueno201210_XENLA_st12	40,733,418	40,583,392	99.6
	Ueno201210_XENLA_st15	47,967,632	47,808,146	99.7
	Ueno201210_XENLA_st20	54,993,210	54,820,790	99.7
	Ueno201210_XENLA_st25	49,103,388	48,945,134	99.7
	Ueno201210_XENLA_st30	83,276,768	82,986,350	99.7
	Ueno201210_XENLA_st35	2,878,474	2,868,310	99.6

	Ueno201210_XENLA_st40	43,564,296	43,428,092	99.7
	Subtotal (bp)	83,814,423,400	83,541,541,400	
	Sample name	Total reads	Filtered	good pct
Ueno201210_tissue	Ueno201210_XENLA_brain	64,427,256	64,194,824	99.6
	Ueno201210_XENLA_eye	61,032,850	60,821,334	99.7
	Ueno201210_XENLA_heart	63,379,128	63,160,712	99.7
	Ueno201210_XENLA_intestine	83,465,934	83,169,164	99.6
	Ueno201210_XENLA_kidney	83,513,834	83,218,198	99.6
	Ueno201210_XENLA_liver	49,653,630	49,489,192	99.7
	Ueno201210_XENLA_lung	44,802,742	44,667,794	99.7
	Ueno201210_XENLA_muscle	68,801,670	68,563,346	99.7
	Ueno201210_XENLA_ovary	61,291,946	61,105,556	99.7
	Ueno201210_XENLA_pancreas	64,541,760	64,346,324	99.7
	Ueno201210_XENLA_skin	70,903,222	70,679,664	99.7
	Ueno201210_XENLA_spleen	66,338,694	66,125,104	99.7
	Ueno201210_XENLA_stomach	67,181,626	66,972,538	99.7
	Ueno201210_XENLA_testis	65,376,770	65,168,412	99.7
	Subtotal (bp)	91,471,106,200	91,168,216,200	
	Sample name	Total reads	Filtered	good pct
Ueno201302_stage	Ueno201302_XENLA_st08	250,876,274	248,464,004	99
	Ueno201302_XENLA_st10	226,682,030	225,279,422	99.4
	Ueno201302_XENLA_st35	261,206,656	259,576,076	99.4
	Subtotal		73,331,950,200	

Appendix Table 2: List of RNA-seq libraries used in annotation and expression analysis

Lib. ID	Library Name	Sequences
animal cap	1 library	
Lib.5323	Wellcome CRC pSK animal cap	3,105
bone	1 library	
Lib.20093	NICHD XGC bone	5,769
brain	3 libraries	
Lib.8910	NICHD XGC Brn1	11,005
Lib.19388	NICHD XGC olfb	4,011
Lib.2550	Xenopus EST library	1,917
digestive	5 libraries	
Lib.20092	NICHD XGC panc	5,898
Lib.5540	NICHD XGC Li1	3,956
Lib.17189	Xenopus liver tumor cDNA library	2
Lib.12208	Xenopus embryonic liver diverticulum plasmid library	1
Lib.12209	Xenopus adult liver ZAP Express phage library	1
dorsal lip	2 libraries	
Lib.15914	Blumberg Cho dorsal blastopore lip	3,973
Lib.7109	Wellcome CRC pRN3 dorsal lip	2,604
ectoderm	1 library	
Lib.15679	Osada Taira anterior neuroectoderm (ANE) pCS105 cDNA library	69,915
endoderm	2 libraries	
Lib.1963	activin-induced ectoderm cDNA library	46
Lib.1962	Xenopus laevis ZAP Express endodermal cDNA library	17

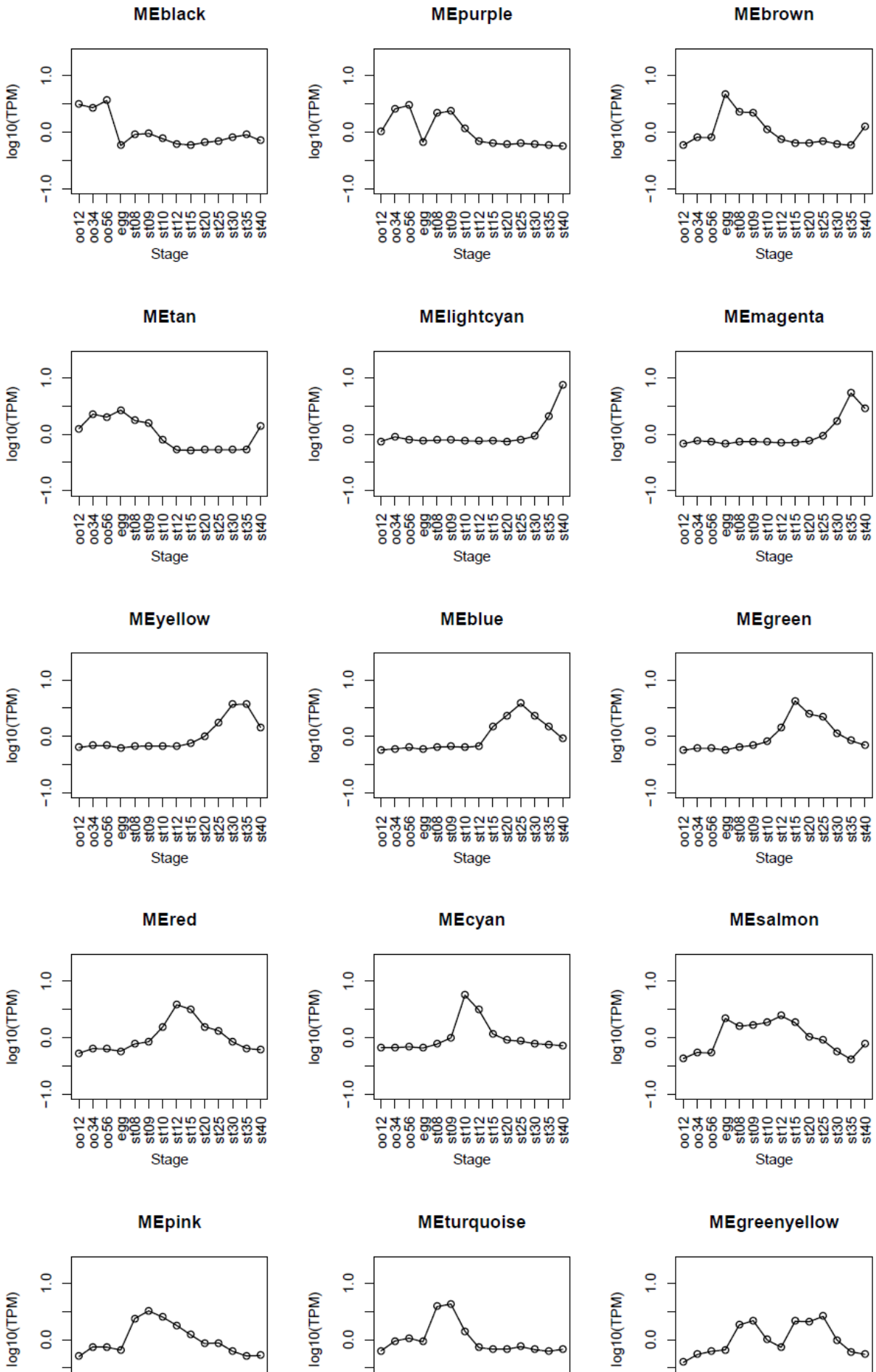
endomesoderm	3 libraries	
Lib.20680	Osada Taira anterior endomesoderm (AEM) pCS105 cDNA library	66,334
Lib.10252	Shibata Xenopus AEM lambda-ZAP II cDNA library	1,043
Lib.10098	AEM cDNA library (lambda-ZAPII)	1
fat body	2 libraries	
Lib.17706	NICHD XGC FaBN	6,245
Lib.17705	NICHD XGC FaB	5,838
head	3 libraries	
Lib.8911	NICHD XGC Eye1	12318
Lib.8603	Wellcome CRC pRN3 head	2,972
Lib.10367	Cornea-lens transdifferentiation library	771
heart	1 library	
Lib.8704	NICHD XGC He1	4,496
kidney	1 library	
Lib.11985	NICHD XGC Kid1	9,662
limb	2 libraries	
Lib.19386	NICHD XGC limb m	5,862
Lib.19644	NICHD XGC limb	5,447
Lib.7211	NICHD XGC Lu1	6,049
ovary	3 libraries	
Lib.7212	NICHD XGC Ov1	17,255
Lib.5329	Harland ovary	105
Lib.893	Xenopus laevis ovary (S.Hirohashi)	7
skin	1 library	
Lib.19645	NICHD XGC skin m	5768
spleen	2 libraries	
Lib.8600	NICHD XGC Sp1	15,807
Lib.19384	NICHD XGC sple PHA	5,465
tail	1 library	
Lib.19387	NICHD XGC tail m	5,605
testis	3 libraries	
Lib.15418	NICHD XGC Te2	12,231
Lib.15412	NICHD XGC Te2N	11,700
Lib.12882	NICHD XGC Te1	2,407
thymus	1 library	
Lib.19565	NICHD XGC thy	5,862
whole body	54 libraries	
Lib.10009	NIBB Mochii normalized Xenopus early gastrula library	40,476
Lib.10008	NIBB Mochii normalized Xenopus tailbud library	35,548
Lib.10005	NIBB Mochii normalized Xenopus neurula library	28,720
Lib.8602	NICHD XGC Emb4	22,270
Lib.4012	Blackshear/Soares normalized Xenopus egg library	19022
Lib.5575	NICHD XGC Emb1	15792
Lib.6801	NICHD XGC OO1	14,764
Lib.12613	NICHD XGC Tad2	13,898
Lib.12612	NICHD XGC Tad1	10,474
Lib.17620	NICHD XGC Emb10	10375
Lib.5324	Wellcome CRC pSK egg	9704
Lib.19385	NICHD XGC int m	8947
Lib.17619	NICHD XGC Emb9	5598
Lib.7111	Wellcome CRC pRN3 St13 17 egg animal cap	3976
Lib.2532	Xenopus laevis oocyte	3846
Lib.5659	Xenopus laevis gastrula non normalized	3659
Lib.5539	NICHD XGC Emb3	3244
Lib.5661	Xenopus laevis unfertilized egg cDNA library	3211
Lib.4114	Harland stage 19-23	3172
Lib.8700	RIKEN Xenopus egg	3023
Lib.8601	Kirschner embryo St10 14	2906
Lib.7258	Wellcome CRC pcDNAI egg	2753
Lib.7108	Wellcome CRC pRN3 St19 26	2687

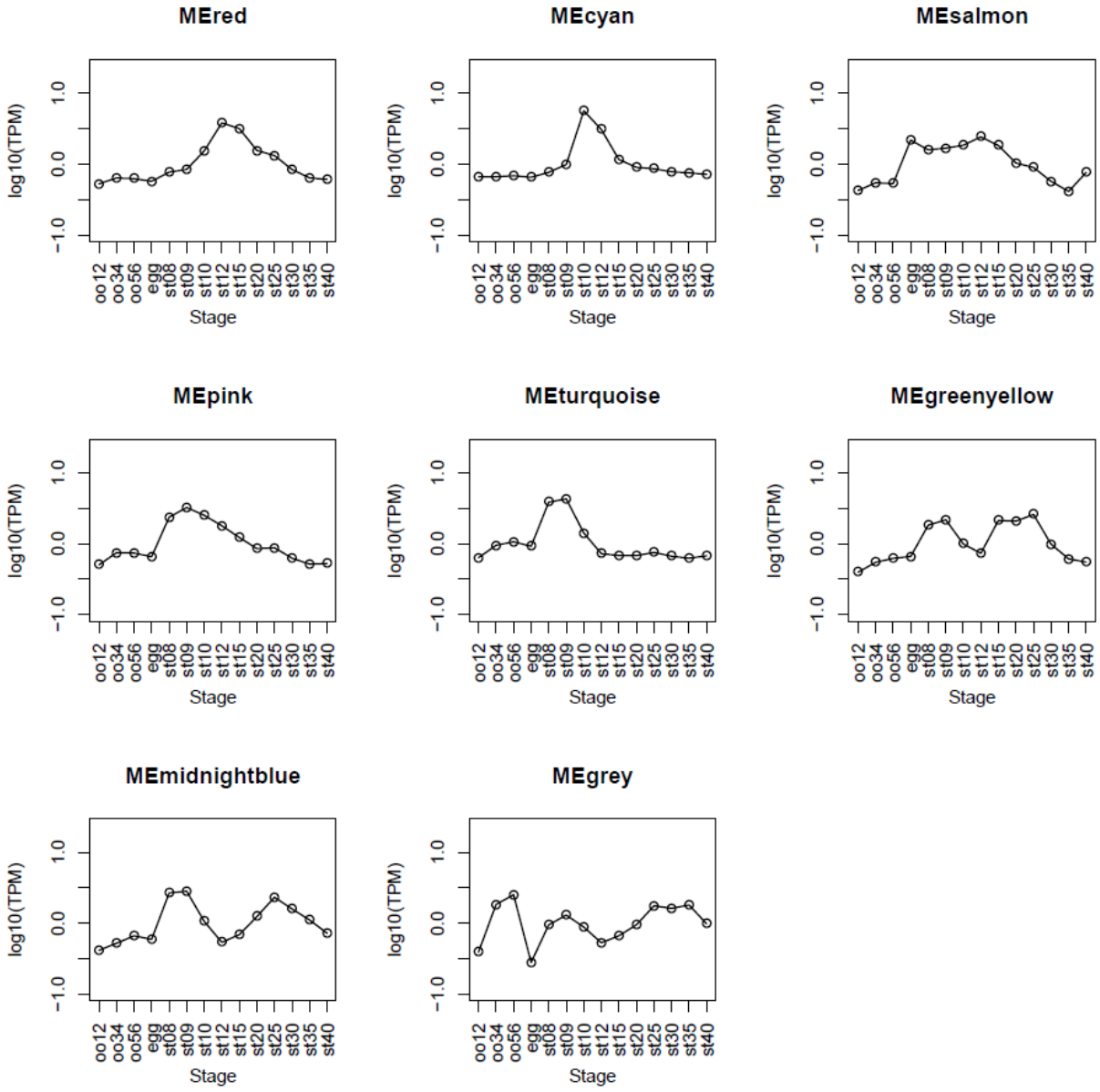
Lib.7110	Wellcome CRC pRN3 oocyte	2668
Lib.5660	Xenopus laevis oocyte non normalized	2583
Lib.4915	Soares NXEG	2322
Lib.4113	Xenopus laevis tadpole stage 24	1852
Lib.2533	normalized Xenopus laevis gastrula	1609
Lib.5325	Wellcome CRC pSK St 10 5	1550
Lib.8599	Cho Li treated gastrula	1272
Lib.2776	Xenla 13LiCl	840
Lib.8754	Stage 10+ Gastrula Library	758
Lib.8598	Wellcome CRC pCS2+ st19-26	662
Lib.9669	Wellcome CRC pcDNAI St10 5	575
Lib.8804	Wellcome CRC pRN3 St10 5	522
Lib.9714	Wellcome CRC pRN3 St19 26 egg animal cap	517
Lib.1326	Xenopus neurula plasmid library	460
Lib.16543	Xenla 13LiCl	146
Lib.8702	Wellcome CRC pcDNAI St24-26	103
Lib.16542	Xenla 13	87
Lib.17018	Xenopus laevis oocyte cDNA subtracted library	84
Lib.8809	Wellcome CRC pRN3 St13 17	77
Lib.16740	Xenopus laevis Lambda TriplEx cDNA Express Library	76
Lib.3800	Xenla 13	72
Lib.8711	Harland stage 19-23 Xenopus laevis cDNA	63
Lib.17257	LiCl-dorsalized gastrula cDNA expression library	46
Lib.17256	UV-ventralized gastrula cDNA expression library	31
Lib.8565	cDNA from differential display on Platinum-treated embryos	19
Lib.17255	32-cell stage cDNA expression library	10
Lib.14127	Xenopus laevis Lambda TripleEx Express Library	6
Lib.10087	cDNA from differential display on mercury-treated embryos	2
Lib.10234	Xenopus laevis tadpole	2
Lib.11059	RT-PCR product from stage 20 RNA	1
Lib.12005	Xenopus Stage 6 cDNA Expression Library	1
uncharacterized tissue	3 libraries	
Lib.1113	Xenopus laevis mitotic phosphoprotein cDNA	13
Lib.1966	Xenopus laevis library (Cao Y)	1
Lib.19662	Xenopus total RNA	1
mixed	1 library	
Lib.14491	Xenopus laevis AGM region stage 46-52	52
not yet classified	2 libraries	
Lib.20683	Yamamoto/Hyodo-Miura NIBB/NBRP Xenopus DMZ pCS2p+ cDNA library	69183
Lib.4650	Xenopus laevis intestine adult	5
Developmental Stage		
oocyte	7 libraries	
Lib.6801	NICHD XGC OO1	14764
Lib.2532	Xenopus laevis oocyte	3846
Lib.7110	Wellcome CRC pRN3 oocyte	2668
Lib.5660	Xenopus laevis oocyte non normalized	2583
Lib.17018	Xenopus laevis oocyte cDNA subtracted library	84
Lib.16740	Xenopus laevis Lambda TriplEx cDNA Express Library	76
Lib.14127	Xenopus laevis Lambda TripleEx Express Library	6
egg	7 libraries	
Lib.4012	Blackshear/Soares normalized Xenopus egg library	19022
Lib.5324	Wellcome CRC pSK egg	9704
Lib.5661	Xenopus laevis unfertilized egg cDNA library	3211
Lib.8700	RIKEN Xenopus egg	3023
Lib.7258	Wellcome CRC pcDNAI egg	2753
Lib.4915	Soares NXEG	2322
Lib.9714	Wellcome CRC pRN3 St19 26 egg animal cap	517
cleavage	1 library	
Lib.12005	Xenopus Stage 6 cDNA Expression Library	1
morula	1 library	
Lib.17255	32-cell stage cDNA expression library	10

blastula	2 libraries	
Lib.15914	Blumberg Cho dorsal blastopore lip	3973
Lib.5323	Wellcome CRC pSK animal cap	3105
gastrula	15 libraries	
Lib.15679	Osada Taira anterior neuroectoderm (ANE) pCS105 cDNA library	69915
Lib.20680	Osada Taira anterior endomesoderm (AEM) pCS105 cDNA library	66334
Lib.10009	NIBB Mochii normalized Xenopus early gastrula library	40476
Lib.5575	NICHD XGC Emb1	15792
Lib.5659	Xenopus laevis gastrula non normalized	3659
Lib.7109	Wellcome CRC pRN3 dorsal lip	2604
Lib.2533	normalized Xenopus laevis gastrula	1609
Lib.5325	Wellcome CRC pSK St 10 5	1550
Lib.8599	Cho Li treated gastrula	1272
Lib.8754	Stage 10+ Gastrula Library	758
Lib.9669	Wellcome CRC pcDNAI St10 5	575
Lib.8804	Wellcome CRC pRN3 St10 5	522
Lib.1963	activin-induced ectoderm cDNA library	46
Lib.17257	LiCl-dorsalized gastrula cDNA expression library	46
Lib.17256	UV-ventralized gastrula cDNA expression library	31
gastrula/neurula cusp	4 libraries	
Lib.10252	Shibata Xenopus AEM lambda-ZAP II cDNA library	1043
Lib.1326	Xenopus neurula plasmid library	460
Lib.1962	Xenopus laevis ZAP Express endodermal cDNA library	17
Lib.10098	AEM cDNA library (lambda-ZAPII)	1
neurula	10 libraries	
Lib.10005	NIBB Mochii normalized Xenopus neurula library	28720
Lib.17620	NICHD XGC Emb10	10375
Lib.17619	NICHD XGC Emb9	5598
Lib.7111	Wellcome CRC pRN3 St13 17 egg animal cap	3976
Lib.4114	Harland stage 19-23	3172
Lib.2776	Xenla 13LiCl	840
Lib.8809	Wellcome CRC pRN3 St13 17	77
Lib.3800	Xenla 13	72
Lib.8711	Harland stage 19-23 Xenopus laevis cDNA	63
Lib.11059	RT-PCR product from stage 20 RNA	1
tailbud embryo	7 libraries	
Lib.10008	NIBB Mochii normalized Xenopus tailbud library	35548
Lib.8602	NICHD XGC Emb4	22270
Lib.5539	NICHD XGC Emb3	3244
Lib.8603	Wellcome CRC pRN3 head	2972
Lib.4113	Xenopus laevis tadpole stage 24	1852
Lib.8702	Wellcome CRC pcDNAI St24-26	103
Lib.12208	Xenopus embryonic liver diverticulum plasmid library	1
tadpole	7 libraries	
Lib.12612	NICHD XGC Tad1	10474
Lib.2550	Xenopus EST library	1917
Lib.10367	Cornea-lens transdifferentiation library	771
Lib.14491	Xenopus laevis AGM region stage 46-52	52
Lib.8565	cDNA from differential display on Platinum-treated embryos	19
Lib.10087	cDNA from differential display on mercury-treated embryos	2
Lib.10234	Xenopus laevis tadpole	2
metamorphosis	4 libraries	
Lib.12613	NICHD XGC Tad2	13898
Lib.19385	NICHD XGC int m	8947
Lib.19386	NICHD XGC limb m	5862
Lib.19387	NICHD XGC tail m	5605
adult	16 libraries	
Lib.7212	NICHD XGC Ov1	17255
Lib.8600	NICHD XGC Sp1	15807
Lib.8911	NICHD XGC Eye1	12318
Lib.15418	NICHD XGC Te2	12231
Lib.15412	NICHD XGC Te2N	11700

Lib.8910	NICHD XGC Brn1	11005
Lib.11985	NICHD XGC Kid1	9662
Lib.7211	NICHD XGC Lu1	6049
Lib.20092	NICHD XGC panc	5898
Lib.20093	NICHD XGC bone	5769
Lib.8704	NICHD XGC He1	4496
Lib.5540	NICHD XGC Li1	3956
Lib.12882	NICHD XGC Te1	2407
Lib.5329	Harland ovary	105
Lib.893	Xenopus laevis ovary (S.Hirohashi)	7
Lib.12209	Xenopus adult liver ZAP Express phage library	1
unknown embryonic stage	2 libraries	
Lib.16543	Xenla 13LiCl	146
Lib.16542	Xenla 13	87
unknown developmental stage	11 libraries	
Lib.17706	NICHD XGC FaBN	6245
Lib.19565	NICHD XGC thy	5862
Lib.17705	NICHD XGC FaB	5838
Lib.19645	NICHD XGC skin m	5768
Lib.19384	NICHD XGC sple PHA	5465
Lib.19644	NICHD XGC limb	5447
Lib.19388	NICHD XGC olfb	4011
Lib.1113	Xenopus laevis mitotic phosphoprotein cDNA	13
Lib.17189	Xenopus liver tumor cDNA library	2
Lib.1966	Xenopus laevis library (Cao Y)	1
Lib.19662	Xenopus total RNA	1
mixed	4 libraries	
Lib.20683	Yamamoto/Hyodo-Miura NIBB/NBRP Xenopus DMZ pCS2p+ cDNA library	69183
Lib.8601	Kirschner embryo St10 14	2906
Lib.7108	Wellcome CRC pRN3 St19 26	2687
Lib.8598	Wellcome CRC pCS2+ st19-26	662
not yet classified	1 library	
Lib.4650	Xenopus laevis intestine adult	5

Appendix Table 3: List of NCBI *X. laevis* EST libraries used in annotation


MEblack
MEpurple
MEbrown
METan
MElightcyan
MEmagenta
MEyellow
MEblue
MEgreen
MEred
MEcyan
MEsalmon
MEpink
MEturquoise
MEgreenyellow
log10(TPM)
−1.0
0.0
1.0
oo12
oo34
oo56
egg
st08
st09
st10
st12
st15
st20
st25
st30
st35
st40
Stage


MEred
MEcyan
MEsalmon
MEpink
MEturquoise
MEgreenyellow
MEmidnightblue
MEgrey
log10(TPM)
−1.0
0.0
1.0
oo12
oo34
oo56
egg
st08
st09
st10
st12
st15
st20
st25
st30
st35
st40
Stage