Evolutionary Genomics of *Xenopus*

Evolutionary Genomics of *Xenopus*: Investigations Into Sex Chromosomes, Whole Genome Duplication, Speciation, and Hybridization

By Benjamin Louis Scott Furman, B.Sc. Specialization

A Thesis Submitted to the McMaster University School of Graduate Studies in the Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University
Department of Biology
Doctor of Philosophy (2018)
Hamilton, Ontario

TITLE: Evolutionary Genomics of *Xenopus*: Investigations Into Sex Chromosomes, Whole Genome Duplication, Speciation, and Hybridization
AUTHOR: Benjamin Louis Scott Furman , B.Sc. (Specialization) University of Alberta
SUPERVISOR: Dr. Ben J. Evans
COMMITTEE: Dr. G. Brian Golding, Dr. James Quinn, Dr. Ian Dworkin, Dr. External
NUMBER OF PAGES: xiv, 208

Acknowledgements

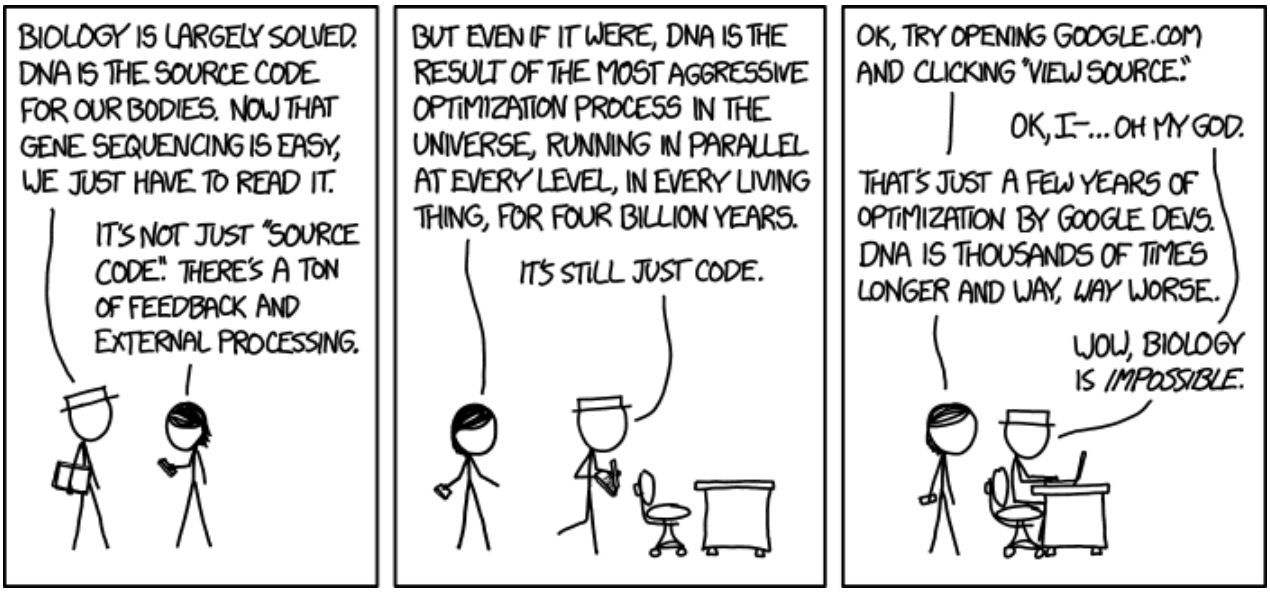
First and foremost, I want to thank my parents. Mom and dad, you have been nothing but supportive and loving, which has sustained me in the tough times and elevated me higher during the best times. What you have taught me offers more than anything I could learn in graduate school.

To my advisor, Ben Evans, it would be wrong to limit you to just that label as more importantly you have been my friend. It would not be possible to thank you enough for the guidance and opportunities you have given me during my time here. But thank you, Ben, for helping me become a scientist.

My dear Florence, nobody has held my hand more than you. You've been my closest confidant, partner in many adventures, and I know my time would have been more difficult if not for your love. You helped me define who I am, and encouraged me to grow. I could not have been more lucky to go on that bike ride with you. Monique and Graham, thank you for being my surrogate family, and providing me with that structure.

Grad school has given me some of my highest highs, and my friends are the reason for that. I have many great memories and many fuzzy ones spending time with all of you. From the conversations of science over coffee, to the jokes and nonsense with beer, I've enjoyed it all. So to my closest friends, Vogan, Chris, Adrian, Ramsha, Dr. Bewick, and a whole host of others that have played an integral part in my time here, thank you for the good times.

Brian Golding, your confusion during my lab presentations has taught me to keep the message simple and to the point, and the guidance you provided has made me a better scientist; I have thoroughly enjoyed our time together. Jim Quinn and Ian Dworkin, I can always count on you two for honest feedback and enlightening discussion. Thank you all for being on my committee.



Randal Munroe, XKCD comics

Abstract

African clawed frogs (*Xenopus*) have been scientific and medical model species for decades. These frogs present many curious features, and their genomic history is no exception. As such, a variety of evolutionary genomic questions can be addressed with these species in a comparative framework, owing to the great array of genetic tools available and a large number of abundant species. The sex chromosomes of this group are evolutionarily young, and this thesis establishes that there has been an additional change in what constitutes the sex chromosomes in one species of *Xenopus*. This allows us to compare the evolutionary trajectory of newly established sex chromosomes. By exploring the genetic content of these systems, profiling their recombinational activity, and assessing the extent of nucleotide divergence between the sex chromosomes, we find that sex chromosome evolution may be predictable in some aspects, and highly unpredictable in others. In addition, this genus is uncharacteristic for vertebrates in the frequency with which lineages undergo whole genome duplication. In this thesis, we explore the selective dynamics operating on duplicate genes over time, and the rate at which duplicate copies are purged from the genome from multiple *Xenopus* species. These investigations provide an animal perspective on the subject of biased subgenome evolution, characteristic of allopolyploids. The last two chapters of this thesis redefine the species boundaries for the most intensively studied *Xenopus* species (*X. laevis*), and explore the genetic extent of hybridization between the common *X. laevis* and the endangered *X. gilli*. Overall, this thesis provides a broad look at several aspects of *Xenopus* evolutionary genomics, providing novel contributions to the fields of sex chromosome research, whole genome duplication, and speciation and hybridization.

Contents

List of Figures

List of Tables

Declaration of Authorship

Each chapter in this thesis was a combination of individual and group efforts. Many colleagues assisted in field collection and raw data preparation (primarily sequencing of tissue samples). Most analyses were performed by me, with guidance from my advisor Dr. Evans. Co-authors and Dr. Evans handled some of the analyses. The major exception to this is the modelling of pseudogenization rates lead by Drs. Dang and Golding, with input by Dr. Evans and myself (chapter 4). Preparation of manuscripts was lead by myself, with input and advice from Dr. Evans, and editorial input by co-authors.

Chapter 1

Introduction

Welcome, to my thesis.

1.1 *Xenopus*

Xenopus, informally known as African clawed frogs, are a genus of Anurans (tailless amphibians) in the family Pipidae. These frogs are largely aquatic and exhibit a number of traits well adapted for this life style, including fully webbed feet, retaining their lateral line after metamorphosis, and a novel method of call production that does not involve the use of large, inflated resonating chambers that are characteristic of many other frogs (Tinsley and Kobel 1996; Tobias et al. 2011). The group has undergone fairly regular taxonomic revision (Cannatella and Trueb 1988; Cannatella and Sa 1993; Bewick et al. 2012; Lloyd et al. 2012; Evans et al. 2015, to name a few), and is currently divided into two subgenera, *Silurana* and *Xenopus* (Evans et al. 2015). These two genera are distinguished based on the number of chromosomes in a haploid cell, being either 10 for *Silurana*, or 9 for *Xenopus* (Evans et al. 2015). The majority of *Xenopus* species are polyploid, ranging from tetraploid to multiple independent events of dodecaploid species (Evans et al. 2015). In the subgenus *Xenopus*, all extant species are the descendants of a single whole genome duplication event that established a tetraploid ancestor between 18 and 34 million years (my) ago (Tymowska 1991; Evans et al. 2015; Session et al. 2016). This duplication event was the result of an allopolyploidization event, combining two fairly distinct diploid genomes (Tymowska 1991; Session et al. 2016); see below for further details. In the extant tetraploids, these two subgenomes are labeled the L (long) and S (short) subgenomes, representing that one set of chromosomes is physically longer than their homeolog (i.e., chromosome related by genome duplication) (Matsuda et al. 2015).

Xenopus have a wide array of interesting features that have made them useful for numerous scientific and medical investigations. First introduced as a pregnancy assay in the 1930s (Gurdon and Hopwood 2003), these species have subsequently been part

of many fields of study, including physiology (e.g., Hogan 2001), endocrinology (e.g., Olmstead et al. 2010), developmental biology (e.g., Watanabe et al. 2005), cellular biology (e.g., Wagner et al. 2000), investigations of neuronal control of sound production and behavior (Kelley and Tobias 1999; Tobias et al. 2004; Tobias et al. 2011), and of course, evolutionary biology of various kinds (e.g., Evans et al. 1998; Sémon and Wolfe 2008; Session et al. 2016). Due in part of the oddities of this group and its ability to be easily housed in lab, these species have been exported world wide (Cannatella and Sa 1993; Gurdon and Hopwood 2003) and have subsequently become invasive in many areas (Measey and Tinsley 1998; Measey et al. 2017), and implicated as a contributing factor to the spread of a deadly anuran pathogen (Weldon et al. 2004).

1.1.1 Objectives of this thesis

In this thesis, I present work advancing investigation of the *Xenopus* genomes, exploring various aspects of sex biased genome evolution, the consequences of polyploidization for selection and genome restructuring, and use genetic variation to assess species dynamics.

Part I

More is known about the sex chromosomes of *Xenopus* than any other amphibian. Previous research on the model species *Xenopus Xenopus laevis* has confirmed one of the few known master regulators of sex determination (Yoshimoto et al. 2008). Subsequent discoveries characterized the molecular origin of this master regulator, and hinted at its absence in some *Xenopus* species (Bewick et al. 2011). The goal of my sex chromosome research was to substantiate a potential change in sex chromosomes, and compare the evolutionary trajectories of different sex chromosome systems in this genus. Chapter 2 focuses on a phylogenetic analysis of the using representatives of various *Xenopus* lineages to establish the evolutionary relationships among species with each sex chromosome system, and contains a genome-wide scan of molecular variation to determine the sex chromosome of the new system. Chapter 3 covers a detailed analysis of recombination suppression and molecular divergence of these recently evolved sex chromosomes. Broadly, this work addresses the topic of predictability in sex chromosome evolution. By looking at two newly evolved systems, we can compare whether or not they evolved in similar ways, which would hint at general forces that govern sex chromosome evolution. As well, by studying the early stages of sex chromosome evolution, this work adds to our understanding of the fundamental steps that shape these chromosomes.

Part II

The evolutionary history of this group is characterized by reticulate evolution. There have been at least 10 whole genome duplication (WGD) events in this genus, which are thought to be from hybridization of different species (allopolyploidization), producing

tetraploid, octoploid, and dodecaploids (Evans et al. 2015). Previous work has compared selection acting on duplicate genes, but the recently sequenced genome of *X. laevis* (Session et al. 2016) and development of models for estimating the deletion rate of genes in the presence of missing data (Dang et al. 2016), allowed me to expand previous work to assess the rate of gene loss over time and evaluate selection in the context of genes contributed by the individual diploid ancestors of the *Xenopus* genus. To this end, chapter 4 leverages transcriptome data of several *Xenopus* species to explore questions related to the dynamics of subgenome evolution in an animal allopolyploid. WGD of some kind is thought to be a pivotal feature in the genomic history of many successful groups, including angiosperms (where WGD is rampant; Jiao et al. 2011) and even vertebrates (2R hypothesis; Dehal and Boore 2005). Overall, this work addresses the selective dynamics that have shaped genomes following duplication and adds an animal perspective to a field largely dominated by plant studies.

Part III

The species *X. laevis* has been a favorite of researchers for decades (Cannatella and Sa 1993; Gurdon and Hopwood 2003). Previous research on *X. laevis* established that genetically distinct populations show different physiological and developmental responses to agricultural contaminants in the environment (Du Preez et al. 2009). Through an Africa-wide comprehensive assessment of genetic variation, chapter 5 makes a reassessment of species status and relationships for the widely studied model clawed frog, *X. laevis*, redefining its distribution and elevating several close relatives to full species status. This detailed understanding of genetic structure for this species contextualizes the wide array of studies done on this frog, and may aid in understanding differential responses among individuals to various challenges. In addition, with its wide distribution, this species and its close relatives can offer insight into the biogeographic forces that influence species boundaries and distributions across Africa.

Finally, the reticulate history hints at a history of hybridization in this group. In extant species, there have been reports of hybridization between several pairs of species (Rau 1978; Picker 1985; Yager 1996; Fischer et al. 2000). One of these potential hybrid zones is between an endangered species (*X. gilli*) and the considerably common *X. laevis*, posing a genomic threat to wipe out the smaller, vulnerable *X. gilli* (Picker 1985; Measey et al. 2011; Villiers et al. 2016). However, genetic investigations into the extent of this introgression have been conflicting, with some reporting a great deal of introgression (Fogell et al. 2013), and others reporting almost none (Evans et al. 1998). In chapter 6, my colleagues and I perform the most comprehensive genetic investigation to date on this hybrid zone using single marker analyses and do not find any evidence of genetic introgression. This work underscores that although hybrids may be found between species, there is not necessarily introgression and a progressive genetic degeneration of one. The threat that the common *X. laevis* poses to the vulnerable and unique *X. gilli* is perhaps more related to ecological and direct competition for food and breeding resources.

Overall, these five chapters expand our knowledge of *Xenopus* biology, for both species dynamics and genome evolution. What follows are more detailed introductions to the three topics of Sex Chromosomes, Whole Genome Duplication, and *Xenopus* Speciation & Hybridization, covered in this dissertation. All chapters in this thesis are either published (chapters 2,3,5,6) or submitted (chapter 4).

1.2 An introduction to sex chromosomes, their evolution and diversity

> *Sex-chromosome evolution is rapid and quixotic, and plays by unique rules. We can begin to understand these rules by comparing the sex chromosomes in the different vertebrate lineages.*
>
> (Graves and Peichel 2010)

In many metazoans, not all chromosomes are equal. Typically, a pair of homologous chromosomes will contain genes that will dictate what sex an embryo will develop, and are called the sex chromosomes (Stevens 1905). These sex chromosomes evolve from a pair of autosomes and what makes them unique is their sex biased modes of inheritance (Ohno 1966). Sex chromosomes come in two forms, either ZZ:males/ZW:females, called female heterogamy and containing a female specific W chromosome, or XX:females/XY:males, called male heterogamy and containing a male specific Y chromosome. Their sex biased modes of inheritance leads to unique selective pressures and population genetic phenomena. For instance, there is a selective advantage to put genes that are beneficial to one sex on the chromosomes that spends most of its time in that sex (e.g., a female beneficial gene is best to have on a W in a ZW/ZZ system; Rice 1987). Another unique feature of sex chromosomes is that, from a population perspective, there are different copy numbers of each chromosome type. For every four copies of homologous autosomes (two in each sex), there are three shared sex chromosomes (two X's in females, one in males, the opposite for Z chromosomes), and one sex limited chromosome (one W in a female, or one Y in a male). This 4:3:1 ratio[1] reflects the relative numbers of chromosomal copies in a population of each type, referred to as the effective population size (N_e), and establishes an expectation for the strength of selection that can act on each chromosome type (and the expected level of population polymorphism for each chromosome type). Low N_e means that genetic drift[2] is able to overwhelm purifying selection, and mutations will accumulate. Conversely a high N_e means that selection is capable of removing more mutations from the population. Thus, the low N_e of the W and Y chromosome means

[1]This is assuming a stable and equal ratio of females and males in the population. If this ratio is changed, for example due to high variance in male reproductive success, or philopatry of one sex, then slightly more of one chromosomes at the expense of the other is expected (e.g., strong variance in male success will reduce the Y N_e further).

[2]Genetic drift is the process by which random alleles are not passed on to the next generation due to sampling error and affected by variable reproductive success among individuals and finite population sizes.

that these chromosomes are expected to accumulate mutations at a faster rate than the Z or X, and faster still than the autosomes.

This accumulation of mutations and sex specific genes surrounding the sex determining gene creates creates a pressure to suppress recombination between the Z and W, or the X and Y in the heteromagetic sex. Suppression of recombination further exacerbates the weakening of selection due to low N_e through Hill-Robertson interference, which is the competing of two beneficial mutations both vying for fixation (without recombination bringing them together on the same haplotype) or the interference of deleterious mutations hindering the fixation of beneficial ones (Hill and Robertson 1966; Gordo and Charlesworth 2001). As well, strong linkage disequilibrium from the suppressed recombination on the sex limited chromosome makes them vulnerable to losing diversity through selective sweeps (Smith and Haigh 1974) and background selection (Charlesworth et al. 1993). Deleterious mutations accumulate in these non-recombining regions in a process known as Müllers ratchet, where the least mutated allele is lost by genetic drift (Muller 1932, 1964; Felsenstein 1974; Gordo and Charlesworth 2001). Thus, over time, the W and Z (or X and Y) sex chromosomes diverge from one another in nucleotide sequence and gene content. The accumulation of mutations on the W and Y from the weakened purifying selection, owing to the small N_e and exacerbated by its inability to match with a recombining partner (two Ws or two Ys are not typically brought together), promotes the formation of heterochromatin and a selective pressure to move critical genes that cannot withstand the mutation accumulation off of the sex limited chromosome (Rice et al. 1994). The net effect is that the sex limited chromosome will often shrink over time, as is seen in most mammals and neognathe birds, with diminutive, gene poor Y and W chromosomes (Bachtrog et al. 2014).

Though the presence of females and males is fairly ubiquitous for vertebrates (present in many plant lineages, and sporadically in other lineages of organisms), and though the molecular cascade that leads to the development of one sex or the other is conserved across vertebrates, the sex chromosomes themselves are highly variable. They are variable in i) the amount of divergence and recombination between the two sex chromosomes, ii) the sex chromosomes that contain the genes and the state of them (ZW or XY) can be different even between closely related species, and iii) there is great diversity in what the master sex determining gene is (Bachtrog et al. 2014). While some sex chromosomes are highly diverged (termed 'heteromorphic'), like those of mammals with less than 100 genes on the Y compared to over 800 genes on the X and have recombination restricted to just the absolute tips of chromosomes (Hughes and Rozen 2012), others similar in gene content with non-diverged in nucleotide sequences, and may recombine throughout the chromosome. The most striking case of a lack of divergence is in the tiger pufferfish (*Takifugu rubripes*), which has a single base pair difference between females and males, and appears to lack recombination suppression entirely (Kamiya et al. 2012). Alternatively, some frogs appear to have no differentiation between the sex chromosomes, despite seemingly complete recombination suppression between the sex chromosomes (Stöck et al. 2011; Stöck et al. 2013). Changes between ZW and XY systems are also very common in some groups, transitioning at least 27 times in amphibians (Evans et al.

2012), 17–25 times in geckos alone (Gamble et al. 2015), and very frequently in fish (Devlin and Nagahama 2002). Finally, as for the master sex determining genes themselves, a wide variety are known, from growth factors (*GSDFY*, Myosho et al. 2012, to several different transcription factors (*SOX3/SR-Y*, Berta et al. 1990; Takehana et al. 2014; *DM-Y/DM-W/DMRT-1*, Matsuda et al. 2002; Yoshimoto et al. 2008; Smith et al. 2009), and hormone receptors (*AMHR2*, Kamiya et al. 2012; *AR*, Fujii et al. 2014). In some clades, like the medaka fish, several different types of genes can be found controlling sex among these closely related species (Myosho et al. 2015a).

Through all this variation, it's hard to see if there are any general rules governing sex chromosome evolution. However, there are some similarities among all these differences. For instance, many of the known master sex determining genes are variants of one another. The aforementioned *DM-Y* and *DM-W* are duplicates of the gene *DMRT-1*, all of which act as master sex determining genes. Similarly, the gene *SOX3* is the master regulator in mammals (as *SR-Y*), and is the most likely candidate in several fish species (Takehana et al. 2014; Myosho et al. 2015a). On a broader scale, large blocks encompassing many genes have independently become sex linked in many long diverged lineages (e.g., a gecko and birds, Kawai et al. 2009; another between several frogs, fish, and mammals, outlined in Chapter 2). Overall, these similarities point to the possibility that there are only a few options in the genome that can control the sex determining cascade, that some blocks of genes or only certain genes will be continuously co-opted to act as the sex chromosomes (Graves and Peichel 2010; O'Meally et al. 2012). There is certainly some ascertainment bias in these studies, in that we tend to look for things we know of, but the similarities across diverse taxa hint at an overall predictability for sex chromosome evolution. If there is a turnover event, and new sex chromosomes established, perhaps then there are only a limited number of possibilities as to what will constitute the sex chromosomes.

There is a solid theoretical basis for understanding what happens to sex chromosomes after formation, establishing the expectation that sexually antagonistic genes[3] will accumulate on the sex chromosomes, and this will enlarge the region of suppressed recombination (Rice 1987). If there is a build up of sexually antagonistic genes surrounding the master sex determining gene, then recombination breaking up the sex specific alleles would produce lower fitness offspring, promoting the suppression of recombination. Strongly sexually antagonistic genes are even thought to drive turnover of sex chromosome (Van Doorn and Kirkpatrick 2007). In a species of Malawi cichlids this may have been the case, as the sex determining gene has moved to a chromosome containing a gene for a gene controlling a sexually antagonistic color variant (Roberts et al. 2009). A recent compelling case for the role of sexual antagonism in modulating the boundaries of suppressed recombination was described in guppy fish. In areas of high predation pressure (the downstream populations), sexual selection for a male specific bright coloration is too costly and recombination suppression on the sex chromosomes is less expansive

[3]These are genes with a sex specific effect, beneficial to one sex and either neutral or even detrimental to the other sex. These genes may also simply have opposing selective strengths between the sexes (being strongly beneficial in one, and only weakly beneficial in the other).

than in low predation pressure populations where sexual selection is stronger (Wright et al. 2017). Thus, in this system there seems to be a strong link between sexual antagonism over coloration and the boundaries of recombination suppression. However, further support for the sound theoretical literature espousing the importance of sexual antagonsism is limited. This is due in part to the difficulties of determining whether the sexually antagonistic genes came before or after the suppression of recombination, as areas of low recombination on the sex chromosomes may promote the accumulation of sexually antagonistic genes (Rice 1987), rather than be caused by sexually antagonistic genes.

Sex chromosomes are often found to have inversions between them. These inversions are a possible mechanism to establish recombination suppression (Stevison et al. 2011), and the subsequent accumulation of sexually antagonistic genes in these regions can act as a selective pressure promoting fixation of the inversions (Charlesworth et al. 2005). Inversions disrupt chromosome alignment during paring, and interfere with the formation of crossovers. Often the effect to recombination extends beyond the breakpoints, promoting an even wider effect of recombination suppression (Stevison et al. 2011). But, like sexually antagonistic genes, inversions are more likely to occur in places where recombination is suppressed, making it difficult to establish what came first (Charlesworth and Charlesworth 1973; Navarro and Ruiz 1997). Recent work has demonstrated that recombination suppression preceded degeneration and divergence of the sex chromosomes. In *Neurospora*, a phylogenetic comparative analyses established that closely related lineages have different inversions on their sex chromosomes, but a consistent region of recombination suppression, suggesting that suppression was the ancestral condition and inversions came after (Sun et al. 2017).

Recombination suppression can be achieved by means other than physical differences between sex chromosomes. Recombination modifiers are sequence variants that alter the local rates of recombination (Ji et al. 1999; Ortiz-Barrientos et al. 2016). These sites may be targeted by proteins like *PRDM9*, which coordinates the formation and repair of double stranded breaks (Baudat et al. 2010). Selective pressure to reduce the rates of recombination, perhaps by altering the sequences that proteins like *PRDM9* bind to, can then promote the formation of recombination cold spots, which could then lower the fitness cost of mutations like inversions that reinforce recombinaiton suppression (Butlin 2005). What exactly constitute recombination modifiers and how they work is still an active area of research (Ortiz-Barrientos et al. 2016), but these likely play a role in sex chromosome recombination suppression (Charlesworth et al. 2005; Chapter 3.3.2,3.4.1).

Relatedly, sex chromosomes are often found to have "strata" of genetic divergence (Lahn and Page 1999), meaning that chunks of the sex chromosomes have regions of divergence that are different from one another, indicating that they ceased recombining at different points in time. In mammals and in birds, nucleotide divergence between genes present on both of the sex chromosomes (X and Y, Z and W, respectively) falls into blocks of increasing levels of divergence, moving out from the sex determining genes (Lahn and Page 1999; Handley et al. 2004; Zhou et al. 2014). These strata may be related

to inversions that happen at different points in time (Lahn and Page 1999). But, for some sex chromosomes, divergence does not fall into discrete blocks. Instead, there appears to have been progressive expansion of the non-recombining region and with continuously decreasing divergence along sex chromosomes moving away from the sex determinign gene. For instance, divergence of the X and Y of *Silene latifolia* is highest near the sex determining gene, and has a rather smooth decrease in divergence moving away (indicative of a lack of strata with hard boundaries, Bergero et al. 2007). Stickleback sex chromosomes (resulting from the fusion of an autosome and an old Y chromosome) show a similar patter too, and in the Japan Sea lineage nucleotide divergence does not fall into discrete blocks (Natri et al. 2013). Thus, it seems that expansion of the non-recombining region may occur in stages, or may be a continuous process.

So, what we are left with is a conundrum over how sex chromosomes evolve. Sexually antagonistic genes may, or may not accumulate. Inversions may or may not happen. Degeneration may be stepwise or progressive, or not happen at all. And, an unexplained element is that the various features of sex chromosomes are not always related to the age of the sex chromosomes (Wright et al. 2016). There are old sex chromosomes, like those of Paleognathe birds, that have only stopped recombining on about $\frac{1}{3}$ of the chromosome, and show limited molecular differentiation despite being >80 my old (Vicoso et al. 2013; Yazdi and Ellegren 2014). And, there are young systems, only a few million years old, that already have considerable cytological and nucleotide differences (e.g., *Drosophila miranda*, Bachtrog et al. 2008). Thus, though there may be some degree of predictability in what can act as a sex determining gene or constitutes the sex chromosomes in terms of overall gene content, once established, there is little predictability in what will happen to the sex chromosomes.

This all leads to the question of how predictable the evolution of sex chromosomes may be. Once established, why do they degenerate in some cases and not others? What governs the dynamics and establishment of recombination suppression and its expansion or lack of? To answer questions regarding sex chromosome evolution, we need to study newly evolved, closely related systems in a comparative genomics framework, which allows us to understand shared and divergent forces acting on sex chromosomes soon after establishment. Studies of young systems can help elucidate what are the prominent features that set sex chromosomes off on one path or the other.

Here, we use *Xenopus* to explore the predictability of sex chromosome evolution, assessing both the regions that control sex and the trajectory of sex chromosomes after establishment. In this genus, the majority of species have a gene called *DM-W*, which resides on a W chromosome (Yoshimoto et al. 2008). Broad phylogenetic surveys of this genus have established that this gene is the result of a partial duplication of the S subgenome copy of *DMRT-1*, and is present in the majority of subgenus *Xenopus* (Bewick et al. 2011). *DM-W* is a female-dominant negative gene, whereby its presence will lead to the development of a female, and its absence means a male will develop, (and transgenic-*DM-W* ZZ individuals will develop as females, Yoshimoto et al. 2008). Transcriptional activity of *DMRT-1*, which normally turns on genes that lead to male

development, is antagonised in a dose-dependent manner by *DM-W*, likely due to their shared DNA-binding sequence (Yoshimoto et al. 2010). *DM-W* is thought to bind to the targets of *DMRT-1*, and lacking a functional transactivation domain, prevents the targets of *DMRT-1* from being expressed (Yoshimoto et al. 2010). *DM-W* resides at the tip of chromosome 2L in a W-specific segment, about 278 kb in size, and together with a small unique Z region (about 83 kb) are the only differentiated regions between the Z and W (encompassing <1% of the total chromosome length, Mawaribuchi et al. 2016). Overall, the sex chromosomes of *X. laevis* are highly homomorphic, as are other *Xenopus* that have been investigated (Tymowska 1991), but measurements of sex specific recombination rates and profiles were lacking (see chapter 3; Furman and Evans 2016).

However, *DM-W* is not present in all *Xenopus* species. Cloning and PCR based efforts were unable to detect *DM-W* in *X. borealis* and it's close relatives (though these efforts did detect a very diverged copy in *X. clivii*; Bewick et al. 2011). This monophyletic clade of *Xenopus* species (Evans et al. 2004; Evans et al. 2005; Evans et al. 2015) lacking *DM-W* was a candidate case of a sex chromosome turnover in the group. But, whether this clade could possibly contain an older system or a newer system was uncertain, as the phylogenetic relationship between this clade and other *DM-W* possessing *Xenopus* has had different supported resolutions, depending on the particular markers used (Evans et al. 2004; Evans et al. 2005; Evans et al. 2015).

In this thesis, I explore various outstanding questions related to *Xenopus* sex chromosomes, and through this aim to understand the evolutionary consequences of establishing novel sex chromosomes. These questions include: i) resolving the phylogenetic relationships among *Xenopus* to understand the history of *DM-W* (chapter 2), ii) testing for the sex specific presence of *DM-W* in the species most distantly related to *X. laevis* (chapter 2), iii) validating the presence of a new sex chromosome system in *X. borealis* (chapter 2), iv) exploring the extent of sex linkage and recombination suppression for these two sex chromosome systems (chapter 3), v) comparing genome-wide sex specific recombination rates in *X. laevis* and *X. borealis* (a first for ZZ/ZW amphibians; chapter 3), vi) profiling nucleotide differentiation of the newly derived sex chromosomes of *X. borealis* (chapter 3). The results indicate both predictability in sex chromosome evolution, supporting the notion of repeated co-option of some regions to act as sex chromosomes (chapter 2). But also underscore the unpredictability of sex chromosomes, as the two *Xenopus* sex chromosome systems have radically different evolutionary trajectories (chapter 3). Recombination suppression has not progressed in the older *X. laevis DM-W* based system, beyond the small region surrounding *DM-W*, but encompasses nearly 50% of the newer *X. borealis* sex chromosomes, underscoring that sex chromosome degeneration is not related to the age of the system. The new sex chromosomes of *X. borealis* have rapidly established widespread recombination suppression, exemplifying that suppression does not always occur in stages, nor must it be a slow, progressive process (chapter 3). And finally, despite recombination suppression being widespread in this new system, differentiation between the sex chromosomes is only modest, indicating that the extent of suppression is not necessarily related to the amount of sex chromosome differentiation (chapter 3).

1.3 An introduction to whole genome duplication

> *[N]atural selection [. . .] is an extremely efficient policeman which conserves the vital base sequence of each gene. [. . .] An escape from the ruthless pressure of natural selection is provided by the mechanism of gene duplication, [. . . emerging] as the major force of evolution.*
>
> (Ohno 1970)

Whole genome duplication (WGD) results in the doubling of every gene in the genome. This can happen either through the duplication of one species genome, termed autopolyploidization, or by the merging of two species where the entire genomic complement of each is maintained, called allopolyploidization. Duplication of the entire genome can bring about problems in the form of unstable meiosis, difficulty finding breeding partners that have also undergone WGD, and decreased level of purifying selection keeping deleterious mutations at bay (Comai 2005; Wang et al. 2010). However, duplications of the entire genome can also be beneficial, allowing for exploration of novel phenotypes (Ohno 1970), and provide greatly increased possible genetic combinations during gametogenesis (from exchange between subgenomes[4], Rieseberg 2001). As a testament to the benefits of WGD, many successful groups have signals of both ancient polyploidy and recent polyploidy (e.g., Teleosts and Salmonids: Allendorf and Thorgaard 1984; Inoue et al. 2015; angiosperms: Fawcett et al. 2009; Jiao et al. 2011). As well, WGD events have occurred at pivotal points, such as the origin of vertebrates (Dehal and Boore 2005), or during the domestication of various crop plants (e.g., corn, Woodhouse et al. 2014; cotton, Renny-Byfield et al. 2015; *Brassica*, Cheng et al. 2012; wheat, Feldman et al. 2012).

WGD through allopolyploidization brings together two diverged genomes of lower ploidy ancestors, and is a mode of reticulate evolution, potentially eroding species boundaries and sympatrically creating a new one. Because of the already present divergence of the genomes, allopolyploids probably immediately have disomic inheritance at meiosis (as opposed to the formation of multivalenets), giving greater genomic stability (avoiding pairing issues of diverged chromosomes and coordinating recombination events across more that two pairing partners, Comai et al. 2003). As well, the merging of diverged genetic material provides greater starting variation for selection to act upon (compared to an autopolyploid, which just doubles its own genome) and may confer heterosis (i.e., the advantage of being a hybrid) (Salmon et al. 2005). But, allopolyploidization is not without consequences, and could allow for release of transposable elements (TEs) potentially leading to "genomic shock", with widespread changes to gene expression (McClintock 1984; Comai et al. 2003).

[4] a 'subgenome' is the half of the genome inherited from the low ploidy ancestor. Thus a newly formed tetraploid from two diploid progenitors has two subgenomes.

There have been many plant allopolyploids that have been analyzed, including numerous crop and model species, such as cotton (Renny-Byfield et al. 2015), wheat (Feldman et al. 2012), coffee (Combes et al. 2013), *Brassica rapa* (Cheng et al. 2012), *Arabadopsis suecica* and possibly the ancient polyploidzation of *Arabadopsis thaliana* (Garsmeur et al. 2013). From these studies, it is evident that post-allopolyploidization bias fractionantion is observed, which refers to the preferential deletion of one parental genome over the other. As well, generally one parental genome is expressed more than the other and experiences stronger purifying selection (Adams 2007). These differences are attributed to differences between the genomes of the lower ploidy progenitors, and are perhaps the consequence of different TE populations between the subgenome. In corn, silencing of TEs has an off-target effects of silencing adjacent genes, making different TE populations between subgenomes a potential source of biased fractionation (Woodhouse et al. 2014; discussed further in the context of *Xenopus* polyploids in chapter 4).

Though polyploidization of one kind or another has played a role in the evolution of metazoans, and vertebrates are no exception (Dehal and Boore 2005; Canestro et al. 2013), there are few cases of confirmed vertebrate allopolyploidization. Though rare in the context of total number of species, many species of polyploid amphibians and fish are known (Mable 2004; Mable et al. 2011). For example, Salmonids, a diverse group of fish, have undergone further WGDs than the basal teolost duplication (Allendorf and Thorgaard 1984; Jaillon et al. 2004), but this probably not an allopolyploidization event, and may instead been autopolyploidization as biased fractionantion is not observed (Garsmeur et al. 2013; Berthelot et al. 2014). Over 64 species of amphibians are thought to be polyploids (Schmid et al. 2015), but few have had their route to polyploidy confidently determined. For some species, multivalent chromosome chains are formed at meiosis, indicating a likely autopolyploid origin (Schmid et al. 2015). Confirming allopolyploidy is difficult as lower ploidy progenitors may not be known[5]. Amphibians are uncharacteristic for vertebrates as they have many polyploid species with two sexes, whereas many other polyploid vertebrates tend to be parthenogenic (Schmid et al. 2015). In plants the two types of polyploids may be distinguished based on whether or not there is biased fractionation (Garsmeur et al. 2013), and that likely holds for animals. For amphibians, *Xenopus* were thought to represent a case of allopolyploidy based on kayrotypic differences of one homeologous chromosome being larger than the other (Tymowska 1991), and a recent genome sequence of *X. laevis* supports this as being the case based on different TE populations (details below; Session et al. 2016). Thus, studying *Xenopus* provides a rare animal perspective on the genomic and adaptive consequences of allopolyploidization.

Subgenus *Xenopus* species represent 44% of known polyploid amphibians (Schmid et al. 2015), and are minimally tetraploid (chromosome numbers of 2n = 4x = 36, gametes carry 18 chromosome that are two sets of nine homeologous chromosome). The diploid descendants of *Xenopus* ancestors that contributed to this allopolyploidization event

[5]In that case, allopolyploidy is indicated by close between species (interspecific) relationships for genes than within species. The half of the genome that came from the ancestor of the lower ploidy species will form a closer phylogenetic relationship than with it homeologous sequence within the polyploid species.

have not been found (Evans et al. 2015; Session et al. 2016). In the sister clade *Silurana* (2n = 20 chromosome pairs, gametes carry 10), a diploid species still exists (*X. tropicalis*), and is a likely descendant of the ancestral diploid that gave rise to the several tetraploid species (2n = 4x = 40, gametes carry 20 with two sets of 10 homeologous chromosomes) in this subgenus (Evans et al. 2015). There are also several octoploid (2n = 8x = 72) and dodecaploid (2n = 12x = 108) subgenus *Xenopus* species. WGD in *Xenopus* is thought to occur via an intermediate triploid stage, followed by multiple rounds of unreduced gametes and back crossing to parental species (Kobel and Du Pasquier 1986). First, two species breed, producing offspring of the same ploidy. This offspring then produces an unreduced gamete and backcross to one of the parental species, producing a triploid offspring[6]. This triploid offspring produces an unreduced gamete and backcrosses to the other parental species, producing a tetraploid offspring containing the genome of both parental species (Kobel and Du Pasquier 1986). Curiously, through this process the sex limited W-chromosome is not necessarily duplicated if the individual producing the unreduced gametes and the triploid stage is a female (Bewick et al. 2011). This process has been replicated in lab with various *Xenopus* tetraploid species to produce viable octoploids, and is assumed to be how WGD has been achieved in nature for this group (Kobel and Du Pasquier 1986; Tinsley and Kobel 1996). Why there are so many polyploids in this group remains an open question. Whether polyploidy provides benefits to animals generally is uncertain (Mable et al. 2011) (unlike what has been shown in plants, Comai 2005). In *Xenopus*, one interesting case points to a possible escape from parasite infections, wherein populations of co-occuring tetraploids and octoploids, the tetraploids carry parasites that do not infect the octoploids (Jackson and Tinsley 2001).

The timing of the ancestral WGD prior to speciation of extant subgenus *Xenopus* individuals is difficult to estimate. Since the diploid ancestors that contributed to this event are extinct, any comparisons of homeologous sequences aimed at estimating the time of their divergence only estimate the time at which the two diploid ancestors speciated from one another. Doing these comparisons sets an upper bound on the age of the WGD event, and a lower bound can be set by estimate the age of the most recent common ancestor (MRCA) of all extant *Xenopus* tetraploids. Until recently, the upper bound estimate ranged from 29–66 (Chain and Evans 2006; Hellsten et al. 2007), and lower bound estimates ranged from 17–41 (Evans et al. 2004; Chain and Evans 2006; Session et al. 2016)[7]. Session et al. (2016) took an interesting approach to try an estimate the actual time at which the two subgenomes were merged. They determined that the S-subgenome had a unique TE population not shared with the L-subgenome. Using relic TEs of this subgenome specific population, they computed JC-corrected nucleotide divergence between these relic sequences and an estimated ancestral sequence of them, and then divided it by their estimated mutation rate from synonymous sites of protein coding genes. This approach yielded an age of 17–18 my ago as the age of

[6]Triploid eggs and unreduced gametes can be produced by temperature shock, or by applying pressure to eggs to prevent extrusion of the polar body (Kobel 1981).

[7]These ranges are primarily due to choice of calibration points, being either the rifting of Africa and South America (Bewick et al. 2012), older fossil evidence (Henrici and Báez 2001), newer fossil evidence (Cannatella 2015), or calibration points from other non-*Xenopus* species (Hellsten et al. 2007).

the WGD event, but did leverage only the youngest estimated calibration point to do so (from Cannatella 2015). If this date is correct, then it means subsequent WGD events in *Xenopus* have happened quite regularly (~1 duplication per 2 my).

Duplicate gene retention rates in tetraploid *Xenopus* are high, as 60% of genes are still in duplicate copy in *X. laevis* (Session et al. 2016). How and why such a large fraction has been maintained could be a consequence of selection to maintain both duplicate copies. Neofunctionalization, where a duplicate copy acquires a new function (Ohno 1970), and subfunctionalization, with either partitioning of a diverse ancestral function or a sharing of a dosage requirement (Force et al. 1999) are two possible explanations invoking selection for the retention of duplicates. Both of these topics have been explored in *Xenopus*, but only a few genes fit with the expectations generated by these possibilities (Chain and Evans 2006; Sémon and Wolfe 2008), but see Hellsten et al. (2007) for a somewhat higher estimate. A large fraction of *Xenopus* duplicate gene pairs still have overlapping expression profiles (Chain et al. 2008; Sémon and Wolfe 2008), indicating that they still likely fulfill similar roles. Mostly, investigations of selective constraints acting on duplicate gene pairs in *Xenopus* consistently find that purifying selection is relaxed in the tetraploids, compared to the diploid *X. tropicalis*, indicating that the two hoemologs likely do not have unique roles (or else they should show purifiying selection levels similar to diploids, where the majority of genes are fulfilling unique roles) (Chain and Evans 2006; Hellsten et al. 2007; Chain et al. 2008; Sémon and Wolfe 2008). A limitation of these studies is that they often include only one or two *Xenopus* species, and typically just *X. laevis* (with *X. tropicalis* as a diploid comparison).

Dosage constraints can be a powerful force to retain duplicate copies. After WGD, all genes are in the same copy number relative to one another, thus any deletions of one would create a stoichiometric imbalance of epistatically interacting genes (Birchler and Newton 1981; Lynch and Conery 2000; Freeling 2009; Gout and Lynch 2015). There is an expectation that expression levels of the two homeologs may drift apart over time, and as one copy encompasses the majority of the needed expression level the fitness cost of silencing the other copy becomes lower (Freeling et al. 2012; Gout and Lynch 2015). In this manner, duplicate genes may be retained for long periods of time before becoming pseudogenized, after expression levels have sufficiently diverged. This model has support from the 350 my old *Paramecium aurelia*, where some genes have only recently become pseudogenized (Gout and Lynch 2015). In *X. laevis*, Session et al. (2016) reports finding 760 homeologs pairs where one homeolog shows little to no expression and has relaxed purifying selection. As well, Chain et al. (2008) found that some duplicate copies in *X. laevis* had different levels of overall expression (while being similar in tissue and timing), indicating that *Xenopus* have accumulated some level of expression divergence between hoemologs (also similar to the finding of Hellsten et al. 2007). With homeolog expression levels drifting apart over time in *Xenopus*, eventually genes will begin to pseudogenize (Zhang and Yang 2015), especially since these low expression genes have increased substitution rates (Session et al. 2016). As for the rate at which the other 40% were lost, little is known. It could be that these genes were rapidly pseudogenized after WGD, or that they have been slowly lost over time (supporting the

drift-pseudogenization hypothesis). Chapter 4 addresses this question.

Previous research on *Xenopus* duplicate gene expression has been limited to comparing duplicate genes to one another, but lacked the context of subgenome of origin (Channing and Howell 2006; Hellsten et al. 2007; Chain et al. 2008; Sémon and Wolfe 2008). Thus, across genes, researchers have not known which copies were from the same subgenome. The recent genome sequence of *X. laevis* alleviates this issue, meaning that a lot of these questions regarding neo-/subfunctionalization, expression divergence of homeologs, differential selective pressure, and genome restructuring post-duplication, can be revisited and addressed in the context of the subgenomes contributed from each of the diploid ancestors. The *X. laevis* genome analysis pointed to consistent differences in expression of and selection on genes between the subgenomes (Session et al. 2016). However, this analysis was limited to mostly just *X. laevis*, and did not thoroughly explore other *Xenopus* species. In this thesis, I present a *Xenopus* subgenus-wide analysis of the rates of gene loss and selection acting on duplicate copies, that explicitly compares the L- and S-subgenomes. Such an analysis can determine if the differences seen for *X. laevis* are universal in the genus, and assess general forces governing duplicate genome evolution in the group, as opposed to species specific influences. As well, using multiple species in a phylogenetic context allows for the exploration of how these forces change over time. Chapter 4 outlines that differences between the subgenomes were present at the time of duplications. Additionally, the previously described differences between the subgenomes (weaker purifying selection on the S and greater gene loss) do seem to be the case for all *Xenopus* tetraploids.

1.4 Speciation and Hybridization in *Xenopus*

> *I said, if they are going to name a Xenopus species after me, I want it to be named Xenopus benopus*
>
> Ben Evans, personal comm.

Xenopus are part of an early branching lineage of Anurans, called Pipoidea (Ford and Cannatella 1993), separating from other anurans 225 my ago (Roelants et al. 2007). These frogs include the fossorial Rhinophrynidae and the aquatic Pipidae families, the former including one extant genera *Rhinophrynus*, and the later including *Pipa*, *Hymenochirus*, *Pseudhymenochirus* and *Xenopus*. Currently there are 29 species of *Xenopus* divided into two subgenera, *Silurana* and *Xenopus* (Evans et al. 2015). *Xenopus* are spread across subsharan Africa, with high species diversity concentrated in the Albertine Rift, and central Africa (Tinsley and Kobel 1996; Evans et al. 2015). *Xenopus* species do not exhibit much morphological diversity, despite existing for tens of millions of years as a clade. However, *Xenopus* do have great vocal diversity, and produce more individual call types than most other amphibians, including females which are generally silent in amphibians (Tobias et al. 1998). These calls can be territorial in nature, or indicate various reproductive states such as advertisement, release from amplexus, or indicating

that they are unreceptive (Tobias et al. 2004; Tobias et al. 2011). *Xenopus* also have high diversity in the range of species distributions, with some species confined to a single lake (*X. longipes*), or spread out across thousands of square kilometers (for instance *X. laevis*, ranging from the south tip of South Africa, all the way to Namibia in the north-west and Malawi in the north-east, chapter 5). Some species represent interesting cases of allopatric divergence, such as *X. clivii* which has disjunct populations that were divided by the opening of the East Africa Rift Valley (Evans et al. 2011b). Polyploidization may have been a driver of sympatric speciation in this group, and there have been at least three independent events to generate octoploids, and four independent events generating dodecaploids (Evans et al. 2015). It is possible that the higher ploidy descendants of these events had higher fitness and replaced the low ploidy progenitors, as many of these progenitor species have not been found (Evans et al. 2015).

Of all these species, the most important to understand is the predominately South African *X. laevis*. This species had been exported for scientific and medical use as far back as 1920s (Cannatella and Sa 1993), and is a highly abundant and wide ranging species that is easily kept in lab and breed in large numbers (Gurdon and Hopwood 2003). This frog, along with several of its close relatives, has at various points in time been elevated to species status, and demoted to subspecies (a full explanation of this can be found in chapter 5 and Appendix D). *X. laevis* generally occur in stagnant water bodies, ranging from high sewage contaminated water bodies, agricultural ditches and water reserves, to pristine natural depressions (basically, they can live anywhere). This species has been introduced and has established populations world wide, including Portugal, France, Argentina, and California (Measey et al. 2017). Due to its world-wide introduction, it has been implicated in the spread of the devastating amphibian fungus *Batrachochytrium dendrobatidis*, which *X. laevis* can live asymptomatically with (Weldon et al. 2004). As outlined above, chapter 5 presents an investigation into the genetic structure of this species as understanding the genetic complexity of this species is necessary to properly stratify responses to the wide array of experimentation performed on it.

As mentioned, *Xenopus* phylogenetic history is characterized by reticulate evolution generated by species hybridization. At present, there have been reports of three hybrid zones between different *Xenopus* species (not producing higher ploidy offspring). There is one hybrid zone between the high elevation *X. borealis* and low elevation *X. victorianus* in Kenya, reported by Yager (1996) based on morphologically intermediate individuals and detection of aberrant breeding calls. Another hybrid zone is thought to exist between *X. laevis* and *X. mulleri* in North-East South Africa, also based on the capture of morphologically intermediate individuals and individuals with intermediate banding patterns of serum albumens and proteins (Fischer et al. 2000). Offspring beyond F_1 generation would be somewhat surprising for either of the hybrid zones as each species pair has a different set of sex chromosomes (chromosome 2L in *X. laevis* and *X. victorianus*, and chromosome 8L in *X. borealis* and probably *X. mulleri*; Yoshimoto et al. 2008; Bewick et al. 2011, and Chapter 2,3). The last hybrid zone between *X. laevis* and *X. gilli* has been studied for many years, with reports of F_1 offspring dating back to the 1970s

(Rau 1978; Picker 1985). The *X. borealis–X. victorianus* hybrid zone has not been investigated genetically (though that is a current objective in the Evan's lab), and the *X. laevis–X. mulleri* has only has some gel-protein work done, thus both lack a in-depth genetic investigation to determine if introgression is occurring (i.e., persisting beyond F_1 individuals). In laboratory breeding experiments, it has been demonstrate that it is possible to achieve fertile F_2 offspring between *X. borealis* and *X. laevis*, despite their different sex chromosome systems (Kobel et al. 1996, Ben Evans, unpublished data). As such, gene flow may be possible between these species. As for the *X. gilli* and *X. laevis* hybrid zone, these species are much more closely related and have the same sex chromosome system (Bewick et al. 2011), making it more likely that there is introgression. But, previous studies have been mixed on whether or not introgression is occurring, and a full description of this history is available in chapter 6. Chapter 6 uses markers spread throughout the genome, testing for signs of introgression.

Part I

Sex Chromosomes

Chapter 2

Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination

Benjamin L. S. Furman* & Ben J. Evans*

*Biology Department, Life Sciences Building room 328, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada

This paper was published in *Gene|Genome|Genetics* and is available here in its published form.

Abstract Sexual differentiation is fundamentally important for reproduction, yet the genetic triggers of this developmental process can vary, even between closely related species. Recent studies have uncovered, for example, variation in the genetic triggers for sexual differentiation within and between species of African clawed frogs (genus *Xenopus*). Here, we extend these discoveries by demonstrating that yet another sex determination system exists in *Xenopus*, specifically in the species *X. borealis*. This system evolved recently in an ancestor of *X. borealis* that had the same sex determination system as *X. laevis*, a system which itself is newly evolved. Strikingly, the genomic region carrying the sex determination factor in *X. borealis* is homologous to that of therian mammals, including humans. Our results offer insights into how the genetic underpinnings of conserved phenotypes evolve, and suggest an important role for cooption of genetic building blocks with conserved developmental roles.

2.1 Introduction

For nearly all vertebrates, two sexes are needed to secure the benefits of genetic recombination associated with sexual reproduction (Barton and Charlesworth 1998). It is, therefore, not surprising that the genetic control of sexual differentiation is tightly regulated, and has remained unchanged for millions of years in several lineages (Matsubara et al. 2006; Veyrunes et al. 2008; Graves and Peichel 2010; O'Meally et al. 2012). However, genetic control of sexual differentiation has diversified in some groups. For example, nonhomologous sex chromosomes have been detected in several closely related species or populations of stickleback (Ross et al. 2009), medaka (Myosho et al. 2015b), and cichlid (Roberts et al. 2009) fish, and rampant turnover of the sex chromosomes occurred over a broader phylogenetic scope in fish (Devlin and Nagahama 2002; Mank et al. 2006), gecko lizards (Gamble et al. 2015), and amphibians (Evans et al. 2012).

Among these turnover events, common elements have been independently coopted for sex determination in several instances. For example, one syntenic block of genes independently became sex-linked in a lizard (*Gekko hokouensis*) and birds (Kawai et al. 2009), and another separately became sex linked in a frog (*Rana rugosa*) and therian mammals (Wallis et al. 2007; Uno et al. 2008; Uno et al. 2013). In addition, individual genes with sex-related function have repeatedly evolved into the trigger for sexual differentiation. Examples include homologs of *doublesex and mab-3 related transcription factor 1* (*DMRT-1*), an important sex related gene in vertebrates (Zarkower 2001), which are triggers for sex determination in medaka fish, *Oryzias latipes* (Kondo et al. 2003, 2004), the African clawed frog *Xenopus Xenopus laevis* (Yoshimoto et al. 2008), probably the Chinese half-smooth tongue sole (Chen et al. 2014) and all birds (Smith et al. 2009, but see Zhao et al. 2010). Similarly, homologs of *SOX3*, which is another important sex related gene (Weiss et al. 2003), independently became triggers for sexual differentiation in the fish *O. dancena* (Takehana et al. 2014) and in the ancestor of therian mammals (Koopman et al. 1991). Turnover of sex chromosomes and the genes involved with sex determination provide opportunities to study how tightly regulated systems evolve, and in particular the extent to which this involves convergence, reversion to an ancestral state, or origin of genetic novelty.

2.1.1 Sex chromosomes of African clawed frogs

In addition to being model organisms for biology (Cannatella and Sá 1993; Hellsten et al. 2010; Harland and Grainger 2011), African clawed frogs (genus *Xenopus*) offer a promising system with which to study sex chromosomes. At least two species, *Xenopus Xenopus laevis* (Daudin 1802) and *Xenopus Silurana tropicalis* (Gray 1864), have a nonhomologous trigger for sex determination (Yoshimoto et al. 2008; Olmstead et al. 2010; Roco et al. 2015). These two species are members of different subgenera that are distinguished from each other by the number of chromosomes (x) carried by the gametes of their respective diploid ancestors, i.e., $x = 10$ for subgenus *Silurana* and $x = 9$ for

subgenus *Xenopus* (Evans et al. 2015). All extant species in subgenus *Xenopus* are polyploid, but with disomic chromosomal inheritance, and tetraploids in this subgenus have $4x = 36$ chromosomes. In *X. laevis*, a gene called *DM-W* is the master sex regulator of sex determination (Yoshimoto et al. 2008); this gene appeared in an ancestor of *X. laevis* after divergence from the ancestor of *X. tropicalis*, and is present in many close relatives of *X. laevis* (Bewick et al. 2011). In subgenus *Silurana*, *X. tropicalis* has a complex trigger for sex determination that resides on Y, W, and Z chromosomes (Roco et al. 2015). This system in *X. tropicalis* produces distorted sex ratios in some crosses (Roco et al. 2015). Thus African clawed frogs use at least two systems for sex determination, and at least one of them evolved during the diversification of this group.

Within subgenus *Xenopus*, species in a clade including *X. borealis* (Parker 1936a), *X. mulleri* (Peters 1844), and *X. fishcbergi* (Evans et al. 2015) appear to lack *DM-W* (Bewick et al. 2011), hinting at additional diversity of sex chromosomes in this group. The phylogenetic placement of this clade within *Xenopus* remains uncertain, making unclear the evolutionary histories of potentially diverse triggers for sex determination.

To further explore sex-related innovations in these frogs, we (i) used whole transcriptome information from several species to further resolve phylogenetic relationships within subgenus *Xenopus*. We (ii) tested whether *DM-W* is sex linked in the most distantly related species from *X. laevis* that is known to carry *DM-W*, i.e., *X. clivii* (Peracca 1898). Then, we (iii) used reduced representation genome sequencing and Sanger sequencing to identify the sex linked region in *X. borealis*, and (iv) established homology between the genes on the sex chromosomes of *X. borealis* and several other distantly related species. Our results identify a new sex determination system in *X. borealis* that evolved after the *DM-W* based system was already in place in an ancestor. Interestingly, the genomic regions involved in sex determination of *X. borealis* and therian mammals (including humans) are homologous. Rapid evolution of *Xenopus* sex chromosomes highlight a central role for genomic recycling in the evolution of important genetic pathways.

2.2 Materials and Methods

2.2.1 Exploring the origin of *DM-W*

Nuclear Data

In order to infer evolutionary relationships among representative *Xenopus* species that do and do not carry *DM-W*, we performed phylogenetic analyses on nuclear sequence data obtained from two sources. For the tetraploid species *X. laevis* and the diploid outgroup species *X. tropicalis*, we used Unigene databases (downloaded November 2015). These data sets had 31,306 and 36,839 unique sequences for *X. laevis* and *X. tropicalis*, respectively. For the tetraploid species *X. borealis*, *X. clivii*, *X. allofraseri*, and *X. largeni*, we extracted RNA from liver tissue using the RNAEasy extraction kit (Qiagen Inc.).

These four transcriptomes were multiplexed on 2/3rds of one lane of an Illumina HiSeq 2000 machine, with 100 base pairs (bp) paired end sequencing and using libraries that were prepared with the Illumina TruSeq RNA Sample Preparation Kit v2. This produced 18–20 million paired reads for each sample (data are deposited in the NCBI short read archive with accession numbers: *X. borealis* PRJNA318484, *X. clivii* PRJNA318394, *X. allofraseri* PRJNA318474, *X. largeni* PRJNA318404).

Low quality reads and bases were removed using TRIMMOMATIC version 0.30 (Bolger et al. 2014). We discarded the first and last 3 bp and then required the average Phred-scaled quality scores of retained sequences to be at least 15 in a sliding window of 4 bp. After imposing these requirements, we discarded all reads that were shorter than 36 bp. Across the samples, 88–95% of paired reads passed these filters. We then assembled the transcriptomes for each species with Trinity (version 2013_08_14), using default values for all settings including, for example, a kmer size of 25 and a minimum contig length of 200 (Grabherr et al. 2011; Haas et al. 2013). The resulting assemblies had 72,000–97,000 unique transcripts (*X. borealis* = 81,696, *X. clivii* = 72,019, *X. allofraseri* = 96,832, *X. largeni* = 82,695) and N50 values (the minimum length, in bp, for the longest 50% of reads) ranging from 885–1176 bp (*X. borealis* = 1,078, *X. clivii* = 885, *X. allofraseri* = 1,176, *X. largeni* = 1,000). Additional information on Illumina sequencing is presented in Table A2.1.

We used a reciprocal BLAST (Altschul et al. 1997b) approach between each tetraploid transcriptome (or Unigene database in the case of *X. laevis*) and the *X. tropicalis* Unigene database to collect sets of homologous sequences for phylogenetic analysis (Fig. A2.1). These sets of sequences included orthologous gene sequences (sequences in different species whose divergence was triggered by speciation), homeologous gene sequences (sequences in the same or different species whose divergence was triggered by genome duplication), and included splice variants, segmental duplicates, and assembly errors generated by Trinity (Grabherr et al. 2011). We performed a quality control step, retaining only those alignments whose ungapped length was above an arbitrary cutoff of 299 bps, and that contained sequences from at least three ingroup species with at least one species having at least two sequences. The need for the requirement that at least one species have two (possibly homeologous) sequences is discussed next.

Because our ingroup species are tetraploid, it was crucial for our phylogenetic analyses to distinguish orthologous from homeologous gene sequences. Since speciation occurred more recently than whole genome duplication in subgenus *Xenopus*, orthologous genes are expected to be more closely related to one another than they are to homeologous genes. In a gene tree with only one sequence from each species, it was therefore a concern that the relationships among the sequences could be orthologous or homeologous. Therefore, we developed a phylogeny-based bioinformatic filter that identified alignments whose estimated phylogeny allowed us to distinguish orthologous from homeologous gene sequences (Fig. A2.1). Importantly, we did not make any assumptions about how the orthologous sequences were related to one another. This filter involved three rounds of tree building, with each followed by assessment of sequence relationships using a

script and functions from the R packages Ape, Phytools, and Phangorn (Paradis et al. 2004; Schliep 2011; Revell 2012; R Core Team 2017; this script is available at Dryad repository; see Data Accessibility). The resulting alignments each included at least one species with two homeologous sequences, which diverged prior to speciation of extant tetraploids in subgenus *Xenopus*. Additionally, each alignment had at least three representative orthologous sequences. Similar BLAST and phylogenetic-based filtering approaches have been used in other studies to distinguish orthologous from homeologous gene sequences (Dehal and Boore 2005; Inoue et al. 2015). See Appendix A1.1 for full details.

Phylogenetic Analyses of Nuclear DNA

After filtering these alignments, we performed several phylogenetic analyses on these data including: (i) individual gene tree analyses for each alignment (Beast; Drummond and Rambaut 2007), (ii) concatenated Bayesian analyses (Beast), (iii) concatenated maximum likelihood analyses (RAxML; Stamatakis 2014b), (iv) a gene tree to species tree analysis using MPest (Liu et al. 2010), and (v) a multi-species coalescent analysis using *Beast (Heled and Drummond 2010). For Analysis (i), a model of evolution was selected for each gene alignment using the Akaike Information Criterion MrModelTest2 (Nylander 2004). We set the root height to be 65 million years (my), with a standard deviation of 4.62 my (Bewick et al. 2012) and assumed a strict clock, and ran 2 chains, for at least 75 million generations. 197 files failed to converge with substitution model selected by MrModelTest, so we instead used the HKY+Γ model. For all analyses we assessed convergence of the posterior distribution using loganalyser (part of the Beast package), and removed a 25% burnin from each chain. For Analysis (i), we summarized relationships across the combined post-burnin posterior distribution of all individual gene analyses using an approach described in Appendix A2.2. We analyzed two datasets for Analyses (ii) and for (iii). The first dataset was a concatenation of all gene alignments. The second dataset had all sites with gaps or missing data removed from the concatenated alignment. For Analysis (ii), for both datasets, we set a GTR+I+Γ substitution model (as selected by MrModelTest using AIC) and a strict clock with an exponential distribution for the rate with a mean rate of 1.0 and a SD of 0.33 (default settings in Beauti). The root height was set to 65 my ($\pm$ 4.62) as detailed above (Bewick et al. 2012). For each dataset, we ran four independent chains, for 50 million generations, and tested for convergence by inspecting the plots of parameter estimates and calculating ESS values using Tracer. Based on this inspection, we removed a 25% burn-in from each chain and constructed a consensus tree using TreeAnnotator. For Analysis (iii), we used the GTR+Γ model and performed 500 bootstrap replicates to assess support.

For Analysis (iv), we used the individually constructed Beast consensus chronograms that were generated from Analysis (i). We selected a random sample of 250 trees from the post-burnin posterior sample of tree topologies from each gene tree analysis to act

as the "bootstrap" replicates, which MPEST uses to assess support (Seo 2008). These trees were uploaded to the STRAW server (Shaw et al. 2013) to run the MPEST analysis.

To perform Analysis (v), we used only those gene alignments that had orthologous sequence data for all species (i.e., five aligned orthologs within one homeologous lineage), and retained only the longest sequence in the other homeologous lineage (or a randomly selected sequence if there were multiple equally long sequences). Because the homeologous sequences are equivalently diverged from a set of orthologs, it did not matter from which species this latter homeologous sequence was derived. The result was a dataset that had gene sequence for all taxa, and minimizing missing data to only incomplete sequencing of a gene and insertion deletion mutations. We ran *BEAST with a strict clock that was linked across all partitions. The GTR+Γ model of evolution was used and was linked across partitions. The tree topology, however, was free to vary among genes (i.e., it was unlinked). We ran two independent chains for 500 million generations each. Convergence was assessed using effective sample size values calculated with TRACER. Based on this, we removed a 25% burn-in from each chain. This analysis did not include calibration points because all attempts to set one failed to converge on the posterior distribution. Instead, in order to assign dates to the nodes, trees in the resulting posterior distribution were rescaled using an R script that used functions from the phytools library (Revell 2012). As above, the root node age was drawn from a normal distribution with a mean of 65 and a SD of 4.62 (Bewick et al. 2012), and the rest of the nodes were assigned based on branch length from the root.

Phylogenetic Analysis of Mitochondrial DNA

We downloaded the previously sequenced mitochondrial genomes for *X. tropicalis* (direct GenBank submission: NC_006839.1), *X. borealis* (GenBank accession no. X155859; Lloyd et al. 2012) and *X. laevis* (GenBank accession no. HM991335; Irisarri et al. 2011). We used the *X. borealis* mitochondrial genome as a BLAST query to recover matches from the transcriptomes of *X. clivii*, *X. allofraseri*, and *X. largeni*, retaining hits with less than an $< e^{-10}$ match. Then, using these assembled mitochondrial DNA (mtDNA) sequences, and the previously sequenced mtDNA genomes, a multispecies alignment was performed using MAFFT (Katoh and Standley 2013) followed by manual adjustment. In order to remove sections that were poorly aligned or had ambiguous homology, GBLOCKS (Castresana 2000) was used with default parameters. We then performed a BEAST analysis of these data, a root node age set to 65 my and a standard deviation of 4.62 (Bewick et al. 2012), a GTR+I+Γ substitution model (as determined by AIC with MRMODELTEST2), and ran 13 chains. For comparative purposes, we ran this analysis with a relaxed clock and with a strict clock, and the suitability of each clock model was assessed by comparing the harmonic means of the postburn-in likelihood values. We also performed a RAxML analysis with a GTR+Γ model and 1,000 bootstrap replicates to assess support.

2.2.2 Assessing sex specificity of *DM-W* in *X. clivii*

The phylogenetic results (discussed below) suggests that *X. clivii* is the most distantly related species to *X. laevis* that carries *DM-W*. Therefore, we tested whether *DM-W* is found only in *X. clivii* females by attempting to amplify *DM-W* in several wild-caught individuals for which sex was inferred based on external morphology (Evans et al. 2011b). We designed primers from a sequenced clone of *DM-W* from this species (Bewick et al. 2011; Table A2.2) and attempted to amplify this gene in 12 females and 13 males.

2.2.3 The sex determining region of *X. borealis*

X. borealis and *X. laevis* families

We generated *X. borealis* and *X. laevis* families from adults obtained from Xenopus Express (Brooksville, Florida, USA). To promote mating, parents each received 50U of human chrionic gonadotropin followed by 200U and 50U for the female and male, respectively, six hours later. The *X. borealis* offspring were reared to sexual maturity, killed with an overdose of MS222, and dissected to determine sex based on presence of testis or ovary. For *X. laevis*, tadpoles were reared for 4 weeks and then euthanized with MS222. Sex of the *X. laevis* tadpoles was determined based on amplification or lack of amplification of a portion *DM-W*; amplification of *DMRT-1* was used as a positive control (Yoshimoto et al. 2008; Bewick et al. 2011). For both families, DNA was extracted using DNEasy kits (Qiagen, Inc) from either fresh liver tissue (*X. borealis*) or tadpole tail tissue (*X. laevis*).

Genotype by Sequencing

To identify the sex determining region of *X. borealis*, we performed Genotype by Sequencing (Elshire et al. 2011) on parents and offspring of the *X. borealis* cross. DNA was extracted for 23 male and 24 female siblings, and both parents using DNEasy extraction kits (Qiagen, Inc). For the mother and father, we sequenced multiple technical replicates to increase coverage 10-fold for each parent compared to each offspring. Library preparation using the EcoT22I restriction enzyme and sequencing was performed at Cornell University Institute of Biotechnology Genome Diversity Facility. Sequencing (100 bp, single end) was performed using an Illumina Hi-Sequation 2500 machine; 96 samples, of which 67 were *X. borealis* samples for this study, were repeated on two Illumina lanes at 96-plex each; the resulting sequence files were merged prior to processing.

We then used Tassel v.3.0 (Glaubitz et al. 2014), employing the UNEAK pipeline (Lu et al. 2013), to perform SNP calling of GBS data without the use of a reference genome sequence. Tassel also does demultiplexing, quality checking, and barcode trimming of sequences. During the process, reads were truncated to a maximum of 64 bp, and high quality reads with < 64 bp were padded with "A" nucleotides to bring them to

the 64 bp length. We set the minimum number of times a read must be present (-c option) to five, and set the error tolerance rate (i.e., the number of mismatched base pairs between reads) to 0.03 when forming groups of homologous sequences. The minimum and maximum allele frequencies of SNPs were set to 0.05 and 0.5, respectively, and the minimum and maximum call rate (i.e., the proportion of all individuals that must have a sequence to call a SNP for a stack of reads) was set to 0.0 and 1.0, respectively. We then trimmed the dataset to only sequence tags that had SNP calls for at least 90% of individuals.

One concern we encountered was "under calling" of heterozygous sites, wherein sites that are actually heterozygous were called as homozygous. For instance, if the parental genotype calls were A/T and A/A, and an offspring was T/T, then it is likely the offspring was actually T/A because the coverage of the parents was ~10X higher. To cope with this, we used a Perl script (deposited in Dryad, see above) to compare offspring genotype calls to those of parent genotype calls for each locus in order to identify biologically implausible genotypes. If $< 10\%$ of offspring had a biologically implausible genotype call, then the implausible genotype calls were changed to missing genotypes. If $> 10\%$ of the offspring had implausible calls, then the site was discarded. With this Perl script, we then identified completely sex biased inheritance of parental SNPs, and used this information to determine whether such sites had inheritance consistent with a female heterogametic (ZZ/ZW) or male heterogametic (XX/XY) sex determining system. We limited our search to loci that were completely sex biased (i.e., only daughters or only sons were heterozygous).

Comparative analysis of the *X. borealis* sex determining region

We used BLAST with the consensus sequences (64 bp long) surrounding the sex-linked SNPs (hereafter "tags") from *X. borealis*, generated by TASSEL, as a query to find matches in the *X. laevis* genome assembly v.7.1 (Bowes et al. 2008). Matching *X. laevis* scaffolds were then aligned to the reconstructed *X. tropicalis* chromosomes in the v.9.0 genome, using the program NUCMER (part of the MUMMER package; Delcher et al. 2002). Settings for NUCMER included a minimum length of a maximal exact match of 50 (`-l 50`), gaps between cluster of matching sequence was set to 500 (`-g 500`), match separation was set at 0.08 (`-d 0.08`) and the minimum cluster length was set to 150 (`-c 150`).

As discussed below, this analysis indicated that a genomic region containing three sex-related genes – *sex determining region Y-box 3* (*SOX3*), *androgen receptor* (*AR*), and *fragile X mental retardation 1* (*FMR1*) – might be sex linked in *X. borealis*. To test this, we amplified and sequenced portions of these three genes in our *X. borealis* family using Sanger sequencing. Primers for both homeologs of *SOX3* and *FMR1* were designed from the *X. laevis* v.7.1 genome or from unpublished *X. borealis* genome sequence data. For *AR*, we used the primers detailed in (Evans et al. 1998), which target the hyper-variable region of one homeolog of the *AR* gene. Primer sequences are reported in Table A2.2.

Using BLAST, we identified chromosomes or scaffolds in the *X. laevis* genome v.9 that are orthologous to these homeologous sequences in *X. borealis*. We also sequenced these genes in wild caught *X. borealis*, including individuals of both sexes and from multiple localities.

As an additional independent test of whether the sex-determining regions of *X. laevis* and *X. borealis* reside in non-homologous genomic regions, we evaluated sex linkage of a *RAB6A* homeolog that is located near *DM-W* (Uno et al. 2013), in both the *X. laevis* and *X. borealis* families. We designed primers for both homeologs using *X. laevis* genome v.7.1 (Table A2.2) and amplified in parents and offspring of both crosses, followed by Sanger sequencing.

Data Availability

Representative individuals from the sex linked alignments and wild samples were deposited in Genbank (accession *SOX3*:KX765742–KX765751; *FMR1*:KX765752–KX765762; *AR*:KX765731–KX765741) and transcriptome and GBS sequences in the NCBI short read archive (accessions PRJNA318484, PRJNA318394, PRJNA318474, PRJNA318404, and PRJNA319044). The phylogenetic trees, gene sequence alignments, BEAST XML files for final gene trees, important scripts used in this study, and full alignments of sex linked genes are deposited in dryad (doi:10.5061/dryad.00db7).

2.3 Results & Discussion

2.3.1 *DM-W* originated before speciation of *X. laevis*, *X. clivii*, *X. borealis*, and other 4x=36 tetraploids.

The gene *DM-W* triggers female sexual differentiation in the African clawed frog *X. laevis* and is located on the female-specific portion of the W sex chromosome (Yoshimoto et al. 2008). This gene is carried by several other *Xenopus* species, but has not been detected in *X. borealis* (Bewick et al. 2011). The most distantly related species from *X. laevis* known to carry *DM-W* is *X. clivii*; however phylogenetic relationships among these three species remain unresolved. If *X. borealis* does indeed lack *DM-W*, two possibilities exist: either (i) *DM-W* arose after divergence of *X. borealis* from the most recent common ancestor (MRCA) of species that carry this gene, including *X. laevis* and *X. clivii*, or (ii) *DM-W* evolved prior to this in the MRCA of species that do and do not carry *DM-W*, and was subsequently lost in a more recent ancestor of *X. borealis*. Analyses of partial mtDNA sequences support the former hypothesis (Evans et al. 2004; Evans et al. 2011a; Evans et al. 2015) and analysis of two linked nuclear DNA (nDNA) genes supports the latter (Evans et al. 2005; Evans 2007; Evans et al. 2015). We therefore estimated phylogenetic relationships among tetraploid species that represent the major *Xenopus* clades, in which *DM-W* has and has not been detected, using new and publicly available sequence data

from nuclear and mitochondrial DNA from *X. largeni*, *X. allofraseri*, *X. borealis*, *X. clivii*, and *X. laevis*, and the diploid outgroup species *X. tropicalis*.

From these data, we recovered 1,585 sets of homologous nuclear gene sequences (Appendix A1.1). Each set consisted of at least one species with two homeologous sequences (i.e., generated from tetraploidization), at least 300 bp for all species, and a minimum of three ingroup taxa for at least one set of orthologs. When combined, these data included 2,696,030 bp. Data from a given ingroup species were missing from the gene alignments as rarely as 14% of the gene alignments (for *X. laevis*) to as much as 64% of the gene alignments (for *X. clivii*, Appendix A1.1). These data formed the basis of Analyses (i–iv). Analyses with gapped sites removed (alternate Analysis ii & iii), included a total of 788,627 aligned bps. The *Beast analysis (Analysis (v)) included 151 gene alignments (238,606 bp, 70,233 sites sequenced for all taxa, with some gaps due to insertion-deletion mutations or incomplete gene sequences).

All of the multigene analyses (ii–v) strongly supported, with a posterior probabilities of 1.0 (or bootstrap support of 100%), two reciprocally monophyletic clades, with the first including *X. borealis* and *X. clivii* and the second including *X. laevis*, *X. largeni*, and *X. allofraseri* (Fig. 2.1 and Fig. A2.2; Appendix A2.1). Similar to previous studies (Evans et al. 2004; Evans et al. 2005; Evans 2007; Evans et al. 2011a; Evans et al. 2015), these analyses failed to resolve relationships among *X. laevis*, *X. largeni*, and *X. allofraseri* with strong support (Fig. 2.1 and Fig. A2.2; Appendix A2.1). Analyses of individual genes (i) identified substantial gene tree discordance among chronograms estimated from each gene (Table A2.3; Appendix A2.2). Despite this discordance, in the pooled post-burnin posterior distribution of these chronograms, a sister relationship between *X. borealis* and *X. clivii* was at least twice as common as any other relationship with either of these species (Table A2.3; Appendix A2.1).

Because previous phylogenetic inferences from mtDNA and nDNA differed with respect to the placement of *X. clivii*, we reexamined mtDNA relationships with additional data from the liver transcriptome sequences of *X. clivii*, *X. largeni*, and *X. allofraseri*, and complete mtDNA genome sequences from *X. tropicalis*, *X. laevis*, and *X. borealis*. After gaps and ambiguously aligned portions were removed, the alignment length was 8318 bp, which spans about 50% of the complete mtDNA genomes of *X. laevis*, *X. borealis*, and *X. tropicalis*. When analyzed with a relaxed molecular clock Bayesian analysis, or with a no clock maximum likelihood analysis, a phylogeny that was topologically consistent with the nDNA analyses was recovered. This topology included a clade containing *X. borealis* and *X. clivii*, although support for this clade was lower than the multigene analyses of nDNA described above (posterior probability was 0.75 and bootstrap support was 66%; Fig. A2.3). Analysis with manual removal of ambiguously aligned sequences instead of Gblocks (16,260 bp aligned) recovered the same topology for both analyses and with similar levels of support (results not shown).

Analysis with a strict molecular clock supported an alternative mtDNA topology, with *DM-W* containing species forming a monophyletic group, as was found by previous studies (Evans et al. 2004, 2015). However, Bayes factors calculated following Nylander

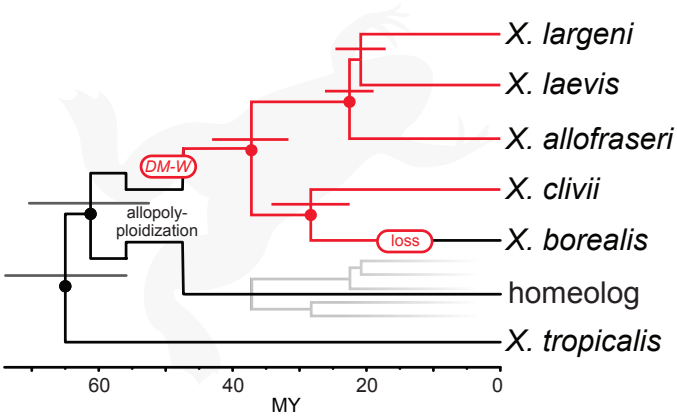


FIGURE 2.1: Phylogenetic relationships inferred from representative species in subgenus *Xenopus* suggests *DM-W* was gained before diversification of ($4x$=36) tetraploids and then lost in an ancestor of *X. borealis*. This phylogeny was recovered from *BEAST analysis of transcriptome data and is topologically consistent with those recovered from other analyses of nuclear DNA and of mitochondrial DNA. Dots over nodes indicate 1.0 posterior probability; bars above nodes indicate the 95% credible intervals for divergence time in millions of years (MY). All species depicted are tetraploids except the outgroup species, *X. tropicalis*, which is diploid. For this analysis, one homeolog from any one of the tetraploid species was included for each gene, and is indicated by the gray subtree (Appendix A1.1). The timing of the origin of *DM-W* with respect to the allopolyploidization event (whether before or after) is unclear. *Xenopus* silhouette from Phylopic by Sarah Werning, CC04 license.

(2004), indicate that a relaxed clock model is strongly preferred over the strict clock (BF = 9.3; Kass and Raftery 1995). An important difference between this and previous mtDNA analyses is that this study is based on a sixfold larger dataset. Similar to the nDNA analyses, mtDNA analyses failed to confidently resolve the relationships of *X. laevis*, *X. allofraseri*, and *X. largeni* (Fig. A2.3).

Although the support for a sister relationship of *X. borealis* and *X. clivii* is lower in the mtDNA analysis than in the nDNA analyses, this relationship has more support than any alternative. Thus, using the most favored models of evolution we considered, the most strongly supported phylogenetic relationships among nDNA and among mtDNA are both consistent with an origin of *DM-W* prior to the diversification of the most recent common ancestor of all of our ingroup taxa (*X. laevis*, *X. largeni*, *X. allofraseri*, *X. clivii*, and *X. borealis*). Results from mtDNA and nDNA, thus, both suggest that *DM-W* originated before the diversification of extant ($4x = 36$) tetraploids in subgenus *Xenopus*.

2.3.2 *DM-W* is sex-linked in *X. clivii*

Our phylogenetic results indicate that *X. clivii*, a species that carries *DM-W*, is closely related to several species in which *DM-W* has not been detected, including *X. borealis* (Bewick et al. 2011). *DM-W* was previously amplified in one female *X. clivii* individual, but it is not clear whether this gene is also sex-linked in this species. Put another way, although *DM-W* arose before *X. laevis* and *X. clivii* diverged from one another, it is possible that *DM-W* acquired its role as a trigger for sexual differentiation (and thus its female-specific mode of inheritance) in an ancestor of *X. laevis* after divergence from an ancestor of *X. clivii*. Therefore, we tested whether *DM-W* is found only in *X. clivii* females, including in our assay males and females from the populations on each side of the Ethiopian Rift Valley (Evans et al. 2011b). We were able to amplify *DM-W* in a subset of females (8 of 12 females) from both sides of the Rift Valley, but no males (0 out of 12 males; a 13^{th} male also failed to amplify in a positive control; Fig. A2.5). The failure of *DM-W* to amplify in four female samples, which were also from both sides of the Ethiopian Rift Valley, could be due to divergence at our primer sites or misidentification of the sex of these individuals when sampled in the field (specimens of these individuals were not available for examination). It is also possible that additional sex determining systems may also be present in *X. clivii*, as is the case in *X. tropicalis* (Roco et al. 2015). Either way, female-specific amplification is consistent with the hypothesis that *DM-W* is found only in female *X. clivii*, that this gene triggers female sexual differentiation in at least some *X. clivii* individuals, and (more broadly) that *DM-W* was the ancestral trigger for female differentiation in subgenus *Xenopus*.

2.3.3 The sex determining region of *X. borealis* is different from that of *X. laevis* and that of *X. tropicalis*

Our inability to detect *DM-W* in *X. borealis* could be because this gene is not present, or because divergence at primer sites prevented amplification with the polymerase chain reaction. To find the sex-linked region of *X. borealis*, we examined patterns of inheritance of SNPs identified in our GBS data from the *X. borealis* family. Of the 89,000 SNPs identified by TASSEL (Table A2.1), 21,000 were successfully genotyped in at least 90% of the offspring, and 15,632 of these passed our filter because they had "undercalled" genotypes in , 10% of the offspring (Materials and Methods). Of these, variation in 25 SNPs had a completely sex linked pattern of inheritance (in offspring one sex is completely homozygous and the other completely heterozygous). By inspecting the genotypes of the parents, we could then distinguish ZZ/ZW from XX/XY systems (Fig. A2.4). All 25 tags were consistent with female heterogamy. In 24 of them, the mother and daughters were heterozygous and the father and sons were homozygous; a pattern best explained by a SNP on the W chromosome. In one of the 25 tags, the mother and sons were heterozygous and the father and daughters were homozygous; a pattern consistent with a SNP on the Z chromosome of the mother that was not present in either Z chromosome of the father. Overall, these results support genetic sex determination and female

heterogamy in *X. borealis*, at least in the strain we examined, which is also the case in *X. laevis* (Mikamo and Witschi 1966) and possibly all other *DM-W*-containing *Xenopus* species.

To evaluate homology of the sex determining regions of *X. borealis*, *X. laevis* and *X. tropicalis*, we aligned the *X. borealis* tags to the *X. laevis* genome assembly. This resulted in tags matching either (i) one region in *X. laevis*, (ii) two regions, (iii) multiple regions, or (iv) no regions. Scenario (ii) is likely the result of the short tags matching both homeologs in the *X. laevis* genome with similar strength. Scenarios (iii) and (iv) are not surprising given the short length of the tags and the divergence between *X. laevis* and *X. borealis* (Fig. 2.1), and we discarded these tags. Ten of the 25 tags had only one or two *X. laevis* scaffold matches below our BLAST threshold ($< e^{-5}$). Six of these 10 scaffolds (either the single match or a randomly retained scaffold if there were two matches) aligned to *X. tropicalis* chromosome XTR8, two scaffolds had a split alignment with portions of each matching two different *X. tropicalis* chromosomes (XTR1 and XTR5 or XTR3 and XTR6, respectively), one matched *X. tropicalis* chromosome XTR4, and one matched *X. tropicalis* chromosome XTR7.

Most of the tags mapped to the XTR8 chromosome, suggesting that the sex chromosomes in *X. borealis* might be homologous to this *X. tropicalis* chromosome. To test this, we designed homeolog-specific primers based on *X. laevis* sequences, to amplify and sequence three genes (*SOX3*, *AR*, and *FMR1*) in our *X. borealis* family that are known to reside on chromosome XTR8 in *X. tropicalis* (Uno et al. 2013). This effort identified sex-linked polymorphisms in *X. borealis* in one homeolog of each gene, and each was consistent with a female heterogametic (ZZ/ZW) sex chromosome system. For *SOX3*, *AR*, and *FMR1*, we successfully amplified and genotyped 93, 41, and 54 offspring, respectively, including 47, 24, and 30 daughters, respectively. For all three of these genes, we identified at least one heterozygous site in themother of the cross that allowed us to confirm sex linkage and female heterogamy (Fig. A2.4; alignments of all sequences are deposited in Dryad and representative sequences are deposited in GenBank; see Data availability). For the *AR* amplification, the father appeared to have a null allele, but importantly, this did not compromise our ability to assess sex linkage and female heterogamy, which was based on patterns of inheritance of a heterozygous SNP from the mother (Fig. A2.4), resulting in completely sex associated genotypes in the offspring. The top BLAST hit of the sex-linked *X. borealis SOX3* and *FMR1* homeologs to the *X. laevis* genome indicated that these sequences were orthologous to *X. laevis* chromosome XLA8L (and thus homeologous to XLA8S); *AR* was orthologous to an unplaced scaffold (scaffold 37), but fluorescent in situ hybridization studies place this gene on XLA8L (Uno et al. 2013).

In wild-caught *X. borealis*, we successfully sequenced amplifications from three females and three males for *SOX3*, and amplifications from the same individuals plus a fourth male for *AR* and *FMR1*. Two of three females tested had the same heterozygous genotypes in *SOX3* and *FMR1* as the females in our lab family; for *AR*, neither of these samples had the same sex-linked polymorphism as the lab family. The wild-caught

females also had other polymorphic sites, some of which were shared with male wild samples. These results indicate either that these genes reside in the pseudoautosomal region in *X. borealis*, that there is variation in the sex determining system within *X. borealis*, or some combination of these possibilities. It is also possible that the sex of some of the wild-caught individuals was misidentified based on external morphology; unfortunately, specimens of these individuals were not available for examination. Examination of other wild-caught individuals whose sex is determined surgically is an important next step for further characterizing the sex-specific region of the sex chromosomes of *X. borealis*.

Analysis of polymorphisms in Sanger sequences of homeologs of the *RAB6A* gene, indicated that one homeolog is linked to *DM-W* in *X. laevis*, as indicated by sex linked inheritance (Appendix A2.3), in agreement with a findings from fluorescence *in situ* hybridization (Uno et al. 2013). This analysis also revealed that the ortholog of *RAB6A* that is sex linked in *X. laevis* is not sex-linked in *X. borealis* (Appendix A2.3). Overall, these results demonstrate that the genomic region containing the trigger for sex determination differs between the *X. borealis* strain we examined and *X. laevis*.

Analysis of polymorphisms in Sanger sequences of homeologs of the *RAB6A* gene indicated that one homeolog is linked to *DM-W* in *X. laevis*, as indicated by sex-linked inheritance (Appendix A2.3), in agreement with a finding from fluorescence in situ hybridization (Uno et al. 2013). This analysis also revealed that the ortholog of *RAB6A* that is sex-linked in *X. laevis* is not sex-linked in *X. borealis* (Appendix A2.3). Overall, these results demonstrate that the genomic region containing the trigger for sex determination differs between the *X. borealis* strain we examined and *X. laevis*.

2.3.4 Some genomic regions are good at sex determination

Our results indicate that the sex chromosomes of *X. borealis* are homologous to *X. tropicalis* chromosome XTR8, orthologous to *X. laevis* chromosome XLA8L, and homeologous to *X. laevis* chromosome XLA8S (Fig. 2.2). In the diploid species *X. tropicalis*, the gene that triggers sex determination is unknown, but resides on the distal end of the petite arm of chromosome XTR7 (Olmstead et al. 2010; Wells et al. 2011; Roco et al. 2015). XTR7 is homologous to *X. laevis* autosomes XLA7L and XLA7S (Uno et al. 2013; Matsuda et al. 2015). The sex chromosome of *X. laevis* is XLA2L; this chromosome and its homeologous chromosome XLA2S are homologous to XTR2 of *X. tropicalis* (Uno et al. 2013; Matsuda et al. 2015). Thus, at least three sets of nonhomologous sex chromosomes are present within the African clawed frogs (Fig. 2.2). The sex chromosomes of *X. laevis* and *X. borealis* occur in orthologous subgenomes (i.e., portions of their respective allotetraploid genomes that are derived from the same diploid ancestor). There is still the possibility that *DM-W* has been translocated to the newer sex chromosomes in *X. borealis*, but all efforts to detected it have failed (Bewick et al. 2011, and here).

Another frog species, *R. rugosa*, has sex chromosomes that are at least partially homologous to those in *X. borealis* (and this may be true for two other *Rana* species; Miura

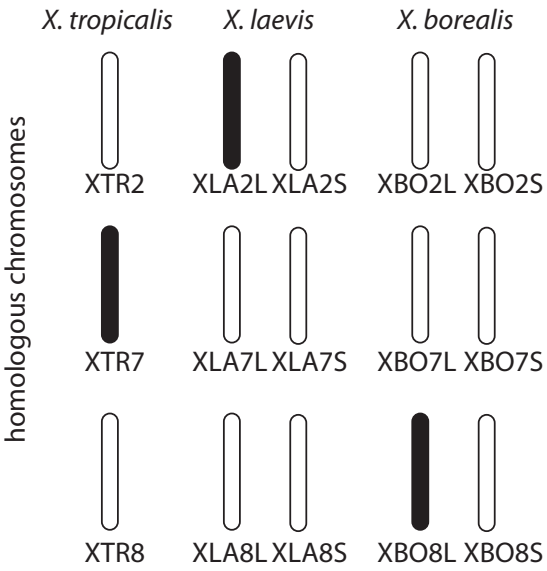


FIGURE 2.2: Sex chromosomes, indicated in black, in three species of African clawed frog are not homologous. For the tetraploid species, *X. laevis* and *X. borealis*, both homeologous (L and S) chromosomes are shown. Chromosome nomenclature for *X. tropicalis* and *X. laevis* follows (Matsuda et al. 2015).

2008). In both species, *SOX3* and *AR* are located on the sex chromosomes (Fujii et al. 2014; Uno et al. 2015). Interestingly, this inference extends even farther: orthologs of *AR*, *SOX3*, and *FMR1* are also present on the X chromosome of therian mammals, including humans (Uno et al. 2013), and *SOX3* is a new trigger for sex determination in a fish (*O. dancena*; Takehana et al. 2014). Similarly, the Z chromosome of lacertid lizards is partially homologous to the X chromosomes of therian mammals (Rovatsos et al. 2016). The phylogenetic placement of these lineages with respect to other species that have different sex determining systems (specifically *X. laevis* and monotremes; Veyrunes et al. 2008) strongly suggests several independent origins of sex linkage of these homologous regions, or minimally of regions containing *SOX3*. Similarly, another region containing *DMRT-1*, an ortholog of which is related by partial gene duplication (paralogous) to *DM-W*, independently became sex-linked in birds and a gecko lizard (Kawai et al. 2009). Taken together, these observations are consistent with the proposal that certain genomic regions contain blocks of genes that are particularly suited to perform the task of triggering sex determination (Graves and Peichel 2010; Brelsford et al. 2013).

2.3.5 Conclusions

Sex chromosomes carry the genetic trigger that initiates sexual differentiation, a crucial developmental phenomenon that is generally required for reproduction (Matzuk and

Lamb 2008). Sex chromosome turnover could occur by translocation among chromosomes of a conserved genetic trigger, or via a novel mutation creating a new trigger on an autosome. That sex chromosomes in African clawed frogs and several other lineages have frequently turned over contrasts sharply with other lineages with ancient sex chromosomes, such as therian mammals. Indeed, transitions in sex chromosomes appear to be more frequent when sex chromosomes are cytologically homomorphic and/or nondifferentiated (Bachtrog et al. 2014), which is the case in *Xenopus*, including *X. borealis* (Tymowska 1991), but not therian mammals. However, the evolutionary dynamics of these systems are highlighted by loss of the Y chromosome in various therians (Just et al. 1995; Sutou et al. 2001) and duplication of SRY, an ancient trigger for sex determination in this group (Geraldes et al. 2010).

A lack of recombination in the genomic region carrying the trigger for sex determination causes sex chromosomes to diverge from one another (Rice 1987). If the region of suppressed recombination expands, as it did in therian mammals, genomic elaborations such as loss and dosage compensation of sex-linked genes may arise and act as "evolutionary traps" that impede evolutionary change or, more specifically, future sex chromosome turnover (Bull 1983; Pokorna and Kratochvíl 2009; Gamble et al. 2015). In theory, before such evolutionary traps evolve, genes with sexually antagonistic function could catalyze sex chromosome turnover by increasing the fixation probability of new sexdetermining genes that arise on a linked autosomal region (Van Doorn and Kirkpatrick 2007). Related to this, dosage compensation has not been detected in species with female heterogamy (Mank 2009b; Vicoso and Bachtrog 2009) or in anurans (frogs) in general (e.g. Schmid et al. 1986), and is unlikely to exist in *Xenopus* species whose female heterogametic sex chromosomes are homomorphic at the cytological (Tymowska 1991) and molecular level (Bewick et al. 2013). An absence of dosage compensation may prevent sex chromosome divergence (Adolfsson and Ellegren 2013) leaving a permissive environment for sex chromosome turnover in the presence of maintained homomorphic sex chromosomes, thereby avoiding these evolutionary traps. However, this is not always the case, as some snakes have differentiated sex chromosomes despite a lack of global dosage compensation (Vicoso et al. 2013; Rovatsos et al. 2015). More information about the nature of the master trigger for sex determination in *X. borealis*, on sex-linked genes, and on sex-biased expression of genes elsewhere in the genome may cast additional light on the drivers of sex chromosome turnover in these frogs. The drivers could include the role of alternative mechanisms that could resolve sexual conflict, such as gene duplication (Gallach et al. 2011; Wyman et al. 2012), which is a potentially important factor in these tetraploid species.

The sex determination system we detected in *X. borealis* is set apart from most other rapidly evolving systems, in that it is derived from an ancestral trigger that itself was newly evolved (i.e., *DM-W*), as opposed to groups with diverse mechanisms that are each potentially once evolved (autapomorphic). Our results support the hypothesis that *X. borealis* and *X. clivii* are sister taxa, and that *DM-W* is restricted to female *X. clivii*. This suggests that female sexual differentiation was triggered by *DM-W* in the ancestor of all extant species of subgenus *Xenopus*. This also suggests that the new system we

report in *X. borealis* is derived with respect to the *DM-W*-based system. Thus, the sex chromosomes of *Xenopus* are an example of multiple important biological novelties arising in rapid succession.

Perhaps most interesting, however, are the aspects of sex determination that convergently evolve in distantly related organisms in the context of frequent sex chromosome turnover. These aspects include the participation of key sex-related genes (e.g., *DMRT-1* in *X. laevis* and in *O. latipes*) and the role of homologous genomic regions (e.g., carrying *SOX3*, in *X. borealis* and in *O. dancena*). This study contributes to a growing body of evidence that, in lineages with rapidly changing sex chromosomes, the turnover is catalyzed by cooption of genetic building blocks that are already involved in the development and maintenance of sexual differentiation.

Acknowledgments. We thank Brian Golding for access to computing resources and Adam Bewick for assistance with rearing tadpoles. Funding for this study was provided by the Natural Science and Engineering Research Council of Canada (RGPIN/283102-2012) and the Museum of Comparative Zoology, Harvard University.

Supplemental Information. Supplemental methods, results, and figures available in Appendix A, or online with the original publication.

Chapter 3

Divergent evolutionary trajectories of two young, homomorphic, and closely related sex chromosome systems

Benjamin L. S. Furman* & Ben J. Evans*

*Biology Department, Life Sciences Building room 328, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada

This paper was published in *Genome Biology and Evolution* and is available here in its published form.

Abstract There exists extraordinary variation among species in the degree and nature of sex chromosome divergence. However, much of our knowledge about sex chromosomes is based on comparisons between deeply diverged species with different ancestral sex chromosomes, making it difficult to establish how fast and why sex chromosomes acquire variable levels of divergence. To address this problem, we studied sex chromosome evolution in two species of African clawed frog (*Xenopus*), both of whom acquired novel systems for sex determination from a recent common ancestor, and both of whom have female (ZW/ZZ) heterogamy. Derived sex chromosomes of one species, *X. laevis*, have a small region of suppressed recombination that surrounds the sex determining locus, and have remained this way for millions of years. In the other species, *X. borealis*, a younger sex chromosome system exists on a different pair of chromosomes, but the region of suppressed recombination surrounding an unidentified sex determining gene is vast, spanning almost half of the sex chromosomes. Differences between these sex chromosome systems are also apparent in the extent of nucleotide divergence between the sex chromosomes carried by females. Our analyses also indicate that in autosomes of both of these species, recombination during oogenesis occurs more frequently and in different genomic locations than during spermatogenesis. These results demonstrate that

new sex chromosomes can assume radically different evolutionary trajectories, with far-reaching genomic consequences. They also suggest that in some instances the origin of new triggers for sex determination may be coupled with rapid evolution sex chromosomes, including recombination suppression of large genomic regions.

3.1 Introduction

Sex chromosomes originate when an autosome acquires a mutation that triggers development of one sex or the other. Recombination between sex chromosomes (the X and Y or Z and W) can be suppressed in regions that include and flank the sex determining mutation, which causes sex-specific inheritance of a sex determining trigger (Charlesworth 1991). Portions of sex chromosomes that lack recombination (e.g., the sex specific portions of the Y or W) and portions that have a reduced level of recombination compared with the autosomes (e.g., the nonpseudoautosomal regions of the X or Z) are subject to distinct population genetic phenomena from autosomes. These genomic regions generally have a lower effective population size than autosomes and thus experience weaker purifying selection (Rice et al. 1994). Portions of each sex chromosome that have a sex-biased mode of inheritance may also have distinct mutation rates (Makova and Li 2002) and generation times (Amster and Sella 2016). Differences in the variance of reproductive success between each sex can further contribute to the disparity in the extent of genetic drift (the effective population size) of these regions (Charlesworth 2009).

A lack of recombination causes portions of the two sex chromosomes to diverge from one another in nucleotide sequence, gene content, and the abundance and distribution of transposable and other repetitive elements (Charlesworth and Charlesworth 2000; Bachtrog 2013). Additionally, the non-recombining region may expand due to accumulation of sexually antagonistic genes, because sex-biased inheritance can mitigate sexual antagonism (Rice 1987; Wright et al. 2017). Over time, these factors can lead to cytological distinctions between the sex chromosomes, a condition known as sex chromosome heteromorphy. In various taxa (e.g., some mammals, birds, and plants), divergence of sex chromosomes occurred incrementally along the length of the sex chromosomes due to sequential inversions or natural selection on recombination modifiers, expanding the non-recombining regions in a stepwise fashion (Coop and Przeworski 2007; Bergero and Charlesworth 2009; Vicoso et al. 2013).

Interestingly and perhaps counterintuitively, the age of the sex chromosomes does not seem to be tightly correlated with whether or not sex chromosomes are cytologically distinct (heteromorphic) or indistinct (homomorphic) (reviewed in Wright et al. 2016). In some old sex chromosomes, for example those of neoaves (> 100 million years (my); Zhou et al. 2014) and therian mammals (~150 my; Graves 2006), and also some young sex chromosomes, such as those of *Drosphila miranda* (~1 Mry; Bachtrog and Charlesworth 2002) and *Silene latifolia* (10–20 my; Bergero et al. 2007), divergence between the sex chromosomes is pronounced. In contrast, in the old sex chromosomes of ratite birds (> 100 my; Zhou et al. 2014), recombination is suppressed over large regions of the sex chromosomes, but accompanied at the nucleotide level by relatively modest differentiation between the sex chromosomes and minimal cytological differentiation (Vicoso et al. 2013; Yazdi and Ellegren 2014). An extreme case of homomorphy exists in the young sex chromosomes of tiger pufferfish, where a single mutation appears to control sexual differentiation and there is no evidence of suppressed recombination (Kamiya et al. 2012). In the young sex chromosomes of hylid tree frogs (~5 my old) and Palearctic green toads

(~3.3 Mry old), recombination appears to be low or absent in heterogametic males, but there is not substantial nucleotide divergence (Stöck et al. 2011; Stöck et al. 2013). Why sex chromosomes of some species are homomorphic whereas those of others are heteromorphic, and why some heteromorphic sex chromosomes are more cytologically diverged than others remains enigmatic (Wright et al. 2016).

3.1.1 Sex Chromosomes Evolved Multiple Times in *Xenopus*

Insights into the origin of variation among species in sex chromosome divergence may be gained by examining whether, to what extent, why, and for how long recombination is suppressed in genomic regions flanking the sex determining locus in multiple species. For this reason, we quantified and compared recombination on the sex chromosomes of the African clawed frog, *Xenopus Xenopus laevis*, and the Marsabit clawed frog, *Xenopus Xenopus borealis*. The most recent common ancestor of these two species experienced allotetraploidization ~18-34 Ma (Evans et al. 2015; Session et al. 2016). These and other allotetraploid species in subgenus *Xenopus* have $2n{=}4s{=}36$ chromosomes, where n refers to the number of chromosomes in a haploid gamete and s refers to the number of chromosomes in an ancestral gamete prior to genome duplication. Chromosomes in tetraploids in subgenus *Xenopus* are numbered 1–18 followed by an L or an S, indicating from which of two diploid ancestors each chromosome was derived (Matsuda et al. 2015).

Species in genus *Xenopus* have homomorphic sex chromosomes (Tymowska and Fischberg 1973; Tymowska 1991), and three nonhomologous sex determining systems have been identified in this group. One is on chromosome 2L of the allotetraploid species *X. laevis* (Yoshimoto et al. 2008) and also several other allopolyploid *Xenopus* species (Bewick et al. 2011). In these species, the W chromosome carries a gene called *DM-W* that triggers female sexual differentiation (Yoshimoto et al. 2008). *DM-W* originated after the whole genome duplication event ancestral to subgenus *Xenopus* species (Bewick et al. 2011). A second sex determination system in *Xenopus* is located on chromosome 8L in the allotetraploid species *X. borealis* (Furman and Evans 2016). This sex determination system evolved in *X. borealis* from an ancestor that carried *DM-W* (Furman and Evans 2016). A third sex determination system in *Xenopus* is located on chromosome 7 in the diploid species *Xenopus Silurana tropicalis* (Olmstead et al. 2010; Evans et al. 2015). In *X. tropicalis*, Z, W, and Y chromosomes segregate (Roco et al. 2015). Overall then, of the three sets of sex chromosomes in *Xenopus*, at least two – those of *X. laevis* and *X. borealis* – are newly evolved, and the system of *X. borealis* is proposed to be derived with respect to (i.e., younger than) the system of *X. laevis* (Fig. 3.1; Furman and Evans 2016).

This variation in sex chromosomes among *Xenopus* species presents an opportunity to compare the evolutionary trajectories of two newly established sex chromosome systems (i.e., the sex chromosomes of *X. borealis* and *X. laevis*). Some differences between the W and Z chromosomes of *X. laevis* have been detected, including differences in gene content, insertion-deletion mutations, and nucleotide divergence, but this is limited to

only a few hundred Kb (< 1% of the chromosome length; Mawaribuchi et al. 2016). However, in general, in *X. laevis* and most other *Xenopus* species little is known about fundamental evolutionary genomic characteristics of sex and recombination, such as sex chromosome-wide levels of divergence, the extent of sex-linkage of genes on sex chromosomes, genome-wide variation in rates of recombination, or sex differences in rates of recombination. We therefore used reduced genome sequencing of parents and offspring of each species to assess sex-linkage of SNPs and to construct sex specific linkage maps for both species. We found that these two systems differ greatly in the extent of sex chromosome recombination suppression during oogenesis, with the younger system in *X. borealis* exhibiting a substantially larger region than the older system of *X. laevis*. Whole genome sequence data indicate that the nonrecombining portions of the *X. borealis* sex chromosomes have a modest, but detectable, level of nucleotide divergence. Finally, linkage mapping in both species demonstrates that females have higher rates of recombination than males of both species, and that the location of crossovers is distinctive between females and males in both species, but similar in same sex comparisons across species. These findings demonstrate that newly evolved sex chromosomes in different species may rapidly assume radically different evolutionary trajectories.

3.2 Materials and Methods

3.2.1 Reduced Representation Genome Sequences from *X. laevis* and *X. borealis* Families

To assess genome wide sex-linkage we used reduced representation genome sequencing (Genotype by Sequencing (GBS): Elshire et al. 2011 and restriction site associated DNA sequencing (RADSeq): Baird et al. 2008) on parents and offspring of an *X. borealis* family and an *X. laevis* family, respectively. For the *X. borealis* family, we used GBS data that we previously reported (Furman and Evans 2016), with a female and male obtained from *XenopusExpress* (Brooksville, FL, USA). These GBS data included mother, father, 24 daughters, and 23 sons (22 and 17 individuals, respectively, after filtering, see Appendix B1.1), with offspring sex determined by dissection after euthanasia. The GBS data were 100 base pairs (bp) single-end sequences; library preparation and sequencing was performed at Cornell University Institute of Biotechnology Genome Diversity Facility on an Illumina HiSeq 2500; other details about these data available in Furman and Evans (2016). For the *X. laevis* family, we obtained female and male individuals from Boreal Science (St. Catharines, ON, Canada). We induced breeding with injection of human chorionic gonadotropin and determined the sex of tadpoles using primers for *DM-W*, which amplifies only in females, and sex-shared primers for *DMRT-1*, which is present in both sexes, as a positive control (Yoshimoto et al. 2008). The RADSeq library was generated by Floragenex (Portland, Oregon, USA) on both *X. laevis* parents, 17 daughters, and 20 sons and 150-bp single-end sequencing was performed at the University of Oregon using an Illumina HiSeq 2500 machine. Though slightly different procedures

were used to generate reduced representation genome sequences from each species, the nature of the data is essentially the same – both methods produced sequence data from many homologous regions in most or all individuals from each family.

GBS or RADSeq data from each *X. borealis* or *X. laevis* individual were demultiplexed, trimmed, and aligned to the *X. laevis* genome version 9.1 (www.xenbase.org) followed by genotyping and filtering steps that are described in the Appendix B1.1. This yielded a panel of SNPs for each family that were used to study recombination as described next. We discuss the potential impacts that the differences in the datasets of *X. borealis* and *X. laevis* may have on our study in Appendix B1.1, Fig. B1.4.

3.2.2 Sex-Linked Genomic Regions

In *X. laevis* and *X. borealis*, females are the heterogametic sex (Yoshimoto et al. 2008; Furman and Evans 2016). Using the filtered data for both families, we thus calculated maternal genotype association with the phenotypic sex (male or female) of each individual SNP following Goudet et al. (1996). Significance was assessed using a false discovery rate correction on the P value of association with sex ($\alpha = 0.05$, using R; R Core Team 2017) and we discarded from this analysis maternal SNPs that were also heterozygous in the father. In order to make inferences discussed below about the region of suppressed recombination that flanks the trigger for sex determination, for each maternal SNP we also determined the frequency of the most common genotype in daughters and then the frequency of this same genotype in sons. We refer to this frequency as the "major daughter genotype frequency." At a completely sex-linked site that was heterozygous in the mother and homozygous in the father, we expected offspring genotypes to be homozygous in one sex and heterozygous in the other (which sex is heterozygous depends on whether the SNP was on the maternal Z or W). Thus, the major daughter genotype frequency at a completely sex-linked site would be 1.0 for daughters, and 0.0 for sons. Conversely, at an autosomal site the major daughter genotype frequency in daughters should be about 50% (but always $\geq$50% because we excluded from this analysis positions with more than two variants). In sons, the major daughter genotype frequency should also be about 50% at autosomal sites, but could be lower or higher than this value.

3.2.3 Linkage Maps

We set out to evaluate rates and locations of recombination events in the mother and the father of our laboratory crosses. To accomplish this, we used the R package OneMap (Margarido et al. 2007) to construct linkage groups based on variable sites from the *X. borealis* and *X. laevis* families that mapped to each of the 18 *X. laevis* chromosomes in the reference genome. For each *X. laevis* chromosome and separately for each species, linkage groups were constructed with a maximum recombination fraction of 0.4 and a LOD threshold of five. With perfect synteny between the *X. laevis* and *X. borealis* and an even genomic distribution of genotyped SNPs, there should be one linkage group

per *X. laevis* chromosome. However, we frequently identified several linkage groups per *X. laevis* chromosome in each species cross and we suspect that this was a consequence of genotyping and mapping errors (see below) and regions with sparse SNPs due to poor mapping of *X. borealis* reads to the *X. laevis* reference genome. For the *X. borealis* family, rearrangements between *X. borealis* and *X. laevis* could also break up a chromosome-specific linkage group. For either species, genome assembly errors could also prevent assembly of one linkage group for a chromosome. We note that our linkage maps did not include a particularly large number of offspring (39 in *X. borealis* and 37 in *X. laevis*), and this contributed to a lack of statistical power to form whole-chromosome linkage groups. However, this was not a concern for (or an objective of) our analyses, which focus on genomic regions for which assembly of linkage groups was possible.

In order to evaluate rates of recombination in the mother and father of each species, we selected the largest linkage group from each chromosome and divided the markers in each linkage group into those that were heterozygous in the mother, in the father, or in both parents. Then, using each of the maternal and paternal sets of markers from each of the largest linkage groups per chromosome, we recomputed recombination fractions between the sets of sex-specific markers and constrained marker order to match the mapping position in the v.9.1 *X. laevis* genome. For the *X. borealis* family, some chromosomes had very few or no double heterozygous sites (sites that were heterozygous in both parents), which is a consequence of the lower overall amount of data for this cross compared with the *X. laevis* cross (due to mapping of *X. borealis* but not *X. laevis* data to a diverged reference genome, and the lower overall coverage we obtained from the GBS data compared with the RADSeq data). This meant that the recombination fractions between male and female markers were unable to be estimated for some chromosomes, and thus the first step of creating a joint linkage group could not be performed. For these chromosomes, we instead selected the largest female-specific and largest male-specific linkage group for each chromosome independently to estimate sex-specific linkage maps. Thus for these chromosomes, the male and female linkage groups do not span identical genomic regions.

3.2.4 Error Correction and Haplotype Estimation

Genotyping errors create genotypes resembling recombined haplotypes that distort linkage maps and lead to inflated map lengths (Hackett and Broadfoot 2003). Although we filtered incompatible parent-offspring genotypes (Appendix B1.1), undercalling of heterozygous sites can also produce incorrect homozygous genotypes in offspring that are nonetheless compatible with parental genotypes. To deal with this problem, we identified putative genotype errors based on phased offspring haplotypes. Each parent has two haplotypes per chromosome, and sites inherited by offspring can be assigned to one or the other haplotype for each parent. Recombination during gametogenesis creates new combinations of the two parental haplotypes within an offspring, with the "phase" referring to which parental haplotype an offspring site comes from (see Fig. B1.1 for a visual explanation). Genotyping errors appears as a change in phase for a single SNPs

(or a few SNPs in a row) when compared with surrounding SNPs. This pattern at one or few sites can also arise biologically from a double recombination (a crossover on either side of a variable position). However, double recombination events in small genomic windows are considered to be rare because of recombination interference (reviewed in Zickler and Kleckner 2016).

To identify putative genotype errors, we used the parental phase estimated during linkage map construction (using OneMap; see Wu et al. 2002 for details on phase estimation of outcross maps) to estimate the parental haplotypes inherited by each offspring individual, for each chromosome-specific linkage map (Fig. B1.1a,b). Under the assumption that double recombination events are rare in small genomic windows, we set to missing data any single genotype supporting a phase change in an individual at just that site (i.e., sites whose flanking genotypes were consistent double recombination event around a single genotyped site). As well, any genotypes in an individual that indicated a double recombination event that only encompassed a small genomic window of < 5 Mb were set to missing data (i.e., a series of sites within 5 Mb who were in an alternate phase compared with adjacent sites). For the *X. laevis* cross, which involved substantially more markers than the *X. borealis* cross, there were more of these potential genotyping errors (4% of all genotyped sites in individuals indicated a double recombination at either a single site or phase changes encompassing < 5 Mb in the *X. laevis* maps, compared with <0.5% for either in the *X. borealis* map; Table B1.1). Over 90% of the putative genotyping errors that were identified based on double recombination like phase changes in the *X. laevis* maps were homozygous, which is consistent with the bulk of these putative errors having been generated by under-called heterozygous positions (Table B1.1). After setting these genotypes to missing data in the affected individuals, we reestimated linkage maps for each chromosome, for each parent for each species. Map distances were then calculated using the Kosambi function (Kosambi 1943).

To quantify recombination events across all maps, we counted all phase changes in each linkage map for each individual based on haplotypes that were constructed from phased SNPs in each offspring. The location of recombination events was approximated as half the distance between the two markers bordering a recombination event in the *X. laevis* reference genome. We assessed the relationship between linkage map length and the amount of bp covered (on the *X. laevis* genome) by each map using a linear model, fitting an interaction between sex and species, along with a three-way interaction between sex, species, and the Mb covered by a linkage map (after scaling and centering Mb) using R. This strategy allowed us to assess for each sex and species slopes for the relationship between cM and Mb. We then used the `confint` function to compute confidence intervals on the estimates.

3.2.5 Divergence between the W and Z chromosomes of *X. borealis*

As discussed below, our analysis identified a large region of the *X. borealis* sex chromosomes that had sex-linked inheritance. If recombination has been suppressed in this

region for a protracted period of evolutionary time, we expected molecular polymorphism in the mother to be higher than the homologous region of the father due to the accumulation of diverged sites between the W and Z. For this reason, we also predicted that polymorphism in this region of the maternal sex chromosomes would be higher than other recombining portions of the maternal genome.

To explore the effects of this lack of recombination at the nucleotide level, we performed whole genome sequencing on the parents of our *X. borealis* family using the Illumina HiSeqX platform at The Center for Applied Genomics (Toronto, Canada), with both individuals multiplexed across two lanes. We trimmed the data, mapped it to the *X. laevis* reference genome, and genotyped and filtered the data as described in the Appendix B1.4. Mapping to a diverged reference genome could lead to a bias of more conserved sequences mapping, than sequences that have evolved quickly. With sex chromosomes, faster-Z (i.e., rapid evolution of Z-linked genes) or degeneration of the W sequences could lead to an underrepresentation of rapidly evolved sequences, leading to an underestimation of divergence. Contrary to this expectation, however, the number of reads mapped to the sex linked region of chromosome 8L in the female (10.2 million) was similar to other identically sized regions of other chromosomes (range 7.8–11 million).

One concern in the quantification of divergence in the nonrecombining portion of the sex chromosomes is that intergenic regions may have many mapping errors due to repetitive sequences. For this reason, we focused our calculation of nucleotide diversity on genomic regions that are within and flank genes, because these areas contain less repetitive DNA (at least in *X. tropicalis*; Shen et al. 2013). We used the *X. laevis* genome annotation (version 9.1 primary gene models gff file; www.xenbase.org) to separately calculate nucleotide diversity (π) in each parent for coding sequence of genes (hereafter CDS), introns, 5′ and 3′ untranslated regions (hereafter UTR), 5,000 bp upstream of the 5′-UTR, and 5,000-bp downstream of the 3′-UTR for genes on all chromosomes. We considered only estimates that were generated from at least 200 bp of contiguous data from both *X. borealis* individuals. Overall, we measured π in 30,876 CDS regions, 3,092 5′-UTRs, 14,954 3′-UTRs, 119,420 introns, 30,326 upstream regions, and 30,270 downstream regions (for a total of 230,016 genomic regions) in the female and the male *X. borealis* individuals.

To test whether the W and Z chromosomes were more diverged in the mother than the homologous Z region in the father, we used a linear mixed model implemented by the LME4 package in R (Bates et al. 2015). We set as fixed effects sex (female or male) and sex-linkage (defined as sex-linked if between bp 4,605,306 and 51,708,524 (corresponding to 100% sex linked tags; Fig. 3.1) of chromosome 8L as defined by the analysis of sex-linked GBS tags discussed below). The six categories of gene regions (CDS, 5′- and 3′-UTRs, introns, up/down-stream) were set as a random effects. The model also included an interaction between the two fixed effects (sex and sex-linkage). We then used likelihood profiles (using the `profile` command in LME4) to calculate confidence intervals on the estimated coefficients.

To visualize and test for differences in divergence within the sex-linked region, we calculated median π for the mother and father in 1 Mb windows of chromosome 8L, using the π estimates from each of the genomic regions (intragenic, 5′-UTR, 3′-UTR, introns, 5,000 bp upstream of genes, and 5,000 bp downstream of genes). Because the mother and father had different levels of polymorphism, we needed to control for this difference in our comparisons between genomic regions of each individual. We therefore first calculated the median π value of all 1 Mb windows across chromosome 8L for each individual. We then standardized the maternal and paternal estimates of π by dividing by their corresponding chromosome-wide median. In order to compare these standardized values of diversity, we then divided the standardized estimates of π measured in each 1 Mb window of the mother by the standardized estimates of π measured in the homologous window of the father. With no difference in level of divergence between alleles we expected this ratio to be equal to one; if the W and Z chromosome were more diverged from each other in the mother than the two Z chromosomes were from each other in the father, this ratio should be greater than one. We tested for a difference between the sex-linked and non-sex-linked portions using a Wilcoxon rank sum test and the measured disparity between parents of each 1 Mb estimates of standardized π. We also explored whether there was a higher rate of synonymouns and nonsynonymous substitutions in genes on the nonrecombining portion of the sex chromosomes to the rest of the genome using the WGS sequence data as described in detail in the Appendix B1.5. Finally, we explored the possibility of an accumulation of deletions and/or insertions on the sex chromosomes. Further details of these analyses are presented in the Appendix B1.6.

3.2.6 Validation of *X. borealis* Sex Chromosomes and Recombination Suppression

To explore whether the expansive region of suppressed recombination in *X. borealis* was limited to our lab raised family, we raised a second family of *X. borealis* using different parents. We then sequenced two genes (*SOX3* and *NR5A-1* [alternatively, *SF-1*]) located 25 Mb apart within the sex linked region (according to placement in the *X. laevis* genome v9.1) to look at coinheritance of alleles from parents to offspring. We also surveyed a panel of adults that were not used in either cross from both sexes to assess linkage of alleles at these two genes. Further details of these assessments are in the Appendix B1.3.

3.3 Results

3.3.1 Diverse Evolutionary Fates of Newly Evolved Sex Chromosomes

Our analysis of the sex chromosomes of *X. borealis* and *X. laevis* identified a far larger region of sex-linked SNPs in *X. borealis* (Fig. 3.1). In *X. borealis*, 40 maternal SNPs spanning ~52 Mb (43%) of the sex chromosome (8L) had a significant association with

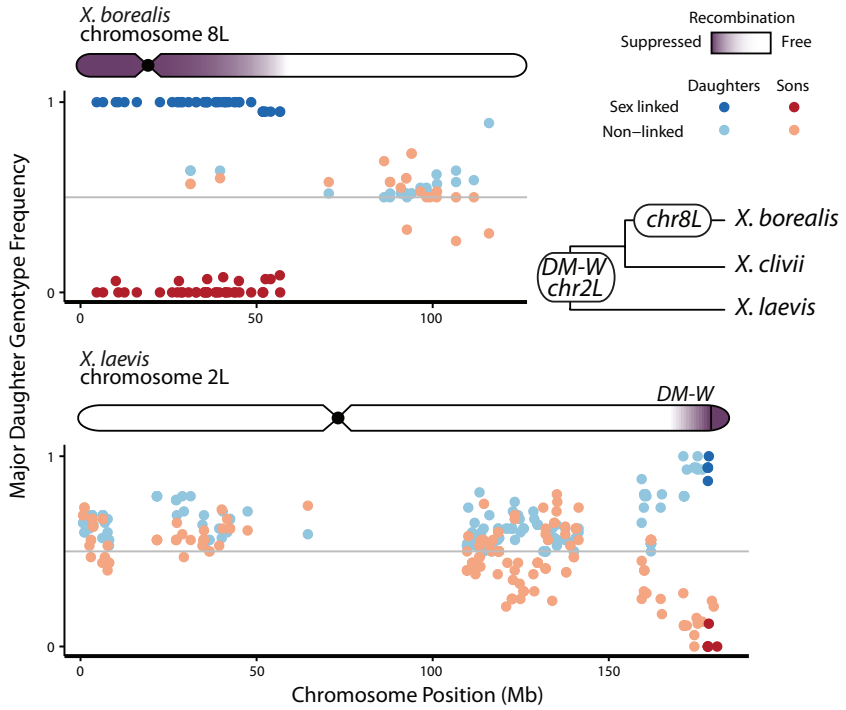


FIGURE 3.1: Sex-linkage of SNPs on sex chromosomes of *X. borealis* and *X. laevis*. In each graph, the x-axis is the position on the sex chromosome using the coordinates of the *X. laevis* reference genome and the y-axis is the major daughter genotype frequency in sons and daughters (see Methods for details) with colors as defined in the key indicating whether or not a SNP is significantly associated with sex (FDR corrected $P < 0.05$). For each species a diagram of a chromosome is shaded darker in the region of suppressed recombination. The inset phylogeny is from Furman and Evans (2016); *DM-W* is carried by female *X. clivii*, but its presence on chr2L has not been confirmed.

the phenotypic sex of offspring (positions 4,605,306 – 56,690,925 of a total chromosome length of ~120 Mb in the *X. laevis* genome assembly; $P < 0.05$ after FDR correction; Fig. 3.1,B1.2). Within this region, daughters had identical genotypes at 34 of the 40 SNPs, with only one daughter differing for the last seven in the region (see below). Similarly in most sons, maternally inherited molecular variation in this genomic region was also almost entirely sex-linked, with exceptions discussed below. Across the entire genome after filtering, the SNP dataset consisted of 1,813 variable positions and there were more heterozygous SNPs in the mother than the father (1,103 and 644 SNPs in the mother and father, respectively, and 66 positions were heterozygous in both parents, with 15–133 SNPs per chromosome, and a mean of 61.8 maternal SNPs per chromosome). For maternal heterozygous positions used for assessing sex linkage in *X. borealis*, daughters had a median depth of 68 and genotype quality of 99 (maximum possible value), sons had a depth of 31 and a genotype quality of 99 (Fig. B1.4). Aligning to the diverged

X. laevis genome substantially reduced the number of SNPs recovered to about 10% of the *de novo* SNP discovery method that did not involve mapping to the *X. laevis* genome (Furman and Evans 2016).

In sharp contrast, on the *X. laevis* sex chromosomes (2L) significant sex-linkage was only detected at only six maternal SNPs spanning 2 Mb (1%; positions 178,144,865 to 180,779,644, and possibly to the end of the chromosome at ~181,296,000; $P < 0.05$ after FDR correction; Fig. 3.1,B1.5). In *X. laevis*, SNPs immediately adjacent to the statically associated SNPs also had a strongly sex-biased pattern of inheritance, which is consistent with recombination suppression of this region (Fig. 3.1). A lack of a statistically significant sex-linkage of some SNPS in this small genomic region may be a consequence of under-called heterozygous positions (Table B1.1 and see Methods). Across the entire genome, there were 7,779 SNPs, and in this family. The father was more polymorphic (1,618 and 4,547 in mother and father, respectively, and 1,614 positions were heterozygous in both parents). For maternal heterozygous positions used in the sex linkage analysis of *X. laevis*, daughters had a median depth of 67, and a genotype quality of 99, sons had a depth of 61 and a genotype quality of 99 (Fig. B1.4).

Within the sex-linked region of *X. borealis*, there was a section with no recombination, and an adjacent section with reduced recombination between positions 51,708,524–56,690,925 of chromosome 8L (Fig. 3.1). Seven consecutive SNPs on the end of this region indicated recombination between the W and Z in one daughter, who had the same genotype as the sons at these positions (Fig. 3.1). Additionally, by inspecting changes in parental phase in the offspring (see below), another maternal recombination event was observed immediately adjacent to the region of completely suppressed recombination in one of the sons (Fig. B1.3). We note that additional information from more offspring or other families could potentially identify more recombination events within the genomic region where we did not observe recombination.

The genomic locations of several SNPs in the *X. borealis* family suggested genotyping or mapping error (Fig. B1.2). For a few sites within the otherwise completely sex-linked region of chromosome 8L, different individual sons had the same genotype as their sisters (Fig. 3.1). If this were due to a real recombination event, we would expect these sons to have the same genotype as their sisters at adjacent SNPs as well. Although this pattern could arise from independent double recombination events around these single sites in different sons, a more plausible explanation is that these are genotyping errors.

We observed three SNPs that mapped to the middle of the sex-linked region of chromosome 8L that were not associated with sex ($P > 0.05$, following FDR, two sites are overlapping on the plot; Fig. 3.1), and we also found five SNPs that were completely sex-linked that mapped to chromosome 8S. These genotypes are best explained by mapping error between *X. borealis* sequence reads and the *X. laevis* genome, or perhaps assembly error in the *X. laevis* genome wherein homeologous portions of the 8L and 8S chromosomes are intermingled in the assembly. It is also possible that sections of homeologous sequences of *X. laevis* and *X. borealis* were lost in an asymmetric fashion after whole genome duplication, such that chromosome 8L in *X. laevis* is missing portions that were

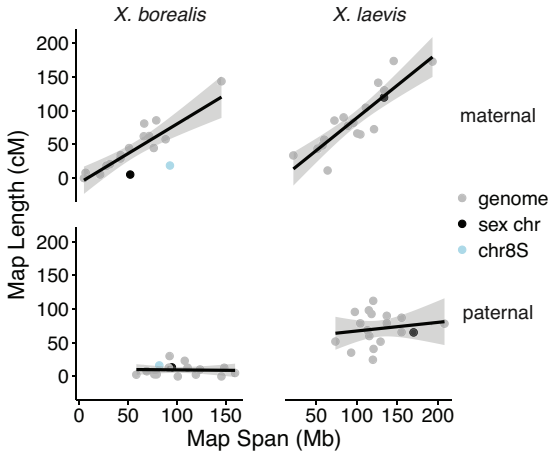


FIGURE 3.2: Linkage map length (in cM) is positively correlated with the number of bp spanned by the map (based on the *X. laevis* genome) for maternal but not paternal linkage maps. Black "sex chr" dots indicate the linkage map of the sex chromosome of each species (chromosome 8L in *X. borealis*, chromosome 2L in *X. laevis*). Lines reflect linear model relationships; gray shading indicates the 95% confidence interval of this relationship. Additionally, chromosome 8S is highlighted for *X. borealis*, because it is the homeolog of the sex chromosome 8L (see Results for details).

not lost in *X. borealis*. This could cause reads from *X. borealis* to map to homeologous sequence in the *X. laevis* genome, instead of to the missing orthologous sequence in *X. laevis*.

We also identified a sex-linked site in *X. borealis* that mapped to *X. laevis* chromosome 5S (Fig. B1.2). We blasted sequence from the GBS tag that contained this SNP to a *de novo* assembly of the maternal *X. borealis* HiSeqX data that were assembled using SOAPDENOVO v.2.04, with a kmer = 23, and default parameters. We then blasted the top hit scaffold back to the *X. laevis* genome and found that its best matches were chromosomes 8S and 8L with similar affinities. This suggests that that this site could be a translocation between *X. borealis* and *X. laevis*, an assembly error in the *X. laevis* genome, or a mapping error due to the short sequence length (< 100 bp) of each GBS tag.

3.3.2 Recombination is Higher in Females of Both Species

Sex differences in the linkage maps revealed higher recombination rates in females of both species. The female linkage maps of both species were longer (*X. laevis* = 1,572 cM; *X. borealis* = 719 cM) than the same-species male linkage maps (*X. laevis* = 1,275 cM; *X. borealis* = 165 cM; Fig. 3.2). Longer female maps were recovered despite female markers spanning fewer base pairs of the *X. laevis* genome in both species (*X. laevis*

female = 1.76 giga base pairs (Gb), male = 2.28 Gb; *X. borealis* female = 0.96 Gb, male = 1.72 Gb; Fig. 3.2). Consistent with this, the number of crossovers is higher in oogenesis than spermatogenesis in both species (*X. laevis*: oogenesis = 558 total; 15.1/offspring, spermatogenesis = 467 total; 12.6/offspring; *X. borealis*: oogenesis = 270 total, 7.3/offspring; spermatogenesis = 62 total; 1.6/offspring).

Also of note is that the locations of crossovers were distinctive in females and males of both species. Female crossovers we more concentrated in the middle of the chromosomes, whereas male crossovers occurred more often at the ends of chromosomes (Fig. 3.3). Possibly related to this (see Discussion), the length in cM of female linkage maps of both species was positively correlated with the number of bp covered by a map, but this relationship was not found in the male linkage maps from either species (linear model slope estimates, 95% confidence intervals: *X. borealis* female = 36.96, 24.78–49.13, male = -0.50 -14.96–13.95, *X. laevis* female = 40.80, 30.66–50.94, male = 5.40, -8.04–18.83; Fig. 3.2). Similar results were recovered when total length of chromosome was used instead of the number of bp covered by the linkage map, or when the number of crossover events was used instead of total cM (results not shown).

For the *X. borealis* family, the largest female linkage group on chromosome 8L (the sex chromosome, which includes both the Z and the W chromosomes) was formed from markers that mapped to the sex-linked portion (Fig. 3.1), and did not include markers from the nonsex-linked portion (see Materials and Methods for possible explanations). This region spanned 52 Mb (43% of the total *X. laevis* chromosome 8L) and was only 5 cM in length. That this recombination probability is not 0 cM is attributable to two recombination events at the end of the region, each of which is illustrated in plots of offspring haplotype assignment (Fig. B1.3). The female linkage map of chromosome 8L was much shorter in recombination probability (cM) than other female and male linkage maps that spanned similar numbers of bp on other chromosomes (Fig. 3.2). The male map of chromosome 8L in the *X. borealis* family, which corresponds to a pair of Z chromosomes, spanned almost the entire chromosome, and had a length of 13 cM, which is similar to other chromosomes (Fig. 3.2). In the father, we detected five recombination events within the portion of chromosome 8L (i.e., between two Z chromosomes) that had suppressed recombination in the mother (i.e., the region where there was almost no recombination between the W and Z chromosomes; Fig. B1.3).

Interestingly, even though it is not a sex chromosome, the maternal linkage map of the *X. borealis* chromosome that is homeologous to the sex chromosome – chromosome 8S – was also substantially shorter in cM than other linkage maps spanning a similar amount of megabases (it was below the best fit line; Fig. 3.2). This suggests that recombination is less frequent on this homeologous chromosome than other autosomes, even though it is not sex-linked.

The *X. laevis* female linkage map of chromosome 2L did not include the last 20 Mb, which is where *DM-W* resides (Session et al. 2016), and where we detected sex-linked SNPs (Fig. 3.1). Therefore, we did not detect any restricted recombination in this map,

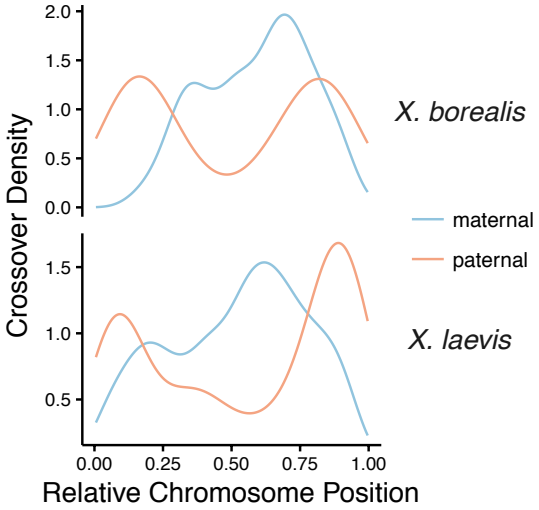


FIGURE 3.3: Density plots of recombination events with respect to the relative position along chromosomes (chromosome length scaled to be between 0 and 1) in the maternal and paternal linkage maps of *X. borealis* and *X. laevis*.

and the size (in cM) of the linkage map of this chromosome was similar to the size of the linkage maps for other chromosomes spanning similar amounts of Mb (Fig. 3.2).

3.3.3 Divergence Between the Sex-Linked Portions of the W and Z Chromosomes of *X. borealis*

We analyzed genotypes inferred from whole genome sequencing data from the mother and the father to test whether we could detect evidence of sex chromosome divergence between sex-linked portions of the W and Z sex chromosomes. Compared with the pseudoautosomal portion of chromosome 8L and also to the autosomes, the sex-linked portion of chromosome 8L had the highest median nucleotide diversity in the female (pairwise nucleotide diversity (π) = 0.012; Fig. 3.4a). In this female genome, diversity within the nonsex-linked (pseudoautosomal) portion of chromosome 8L was similar to that of other chromosomes ($\pi = 0.009$; Fig. 3.4a). In the male genome, diversity of each portion of chromosome 8L fell within the range of estimates from other chromosomes from this genome (sex linked: $\pi = 0.0072$; non-sex linked: $\pi = 0.009$; Fig. 3.4a). The nucleotide diversity measured for these chromosomes is far less than the 7% divergence of homeologous sequences (Evans and Kwon 2015); the considerably lower π estimates reported here suggest that cross mapping of reads across subgenomes was relatively rare.

Analyses of nucleotide diversity in and around genes (divided into six categories; see Materials and Methods), which used a linear mixed model, recovered a significant interaction between sex and sex-linkage, indicating that the mother had a higher π than the father in the sex-linked portion of chromosome 8L compared with the rest of the

genome, and after controlling for differences in polymorphism between these individuals (estimate of the increase in female diversity in the sex linked region = 0.0018, 0.0009 – 0.0027 95% CI, t-stat = 4.09; Fig. 3.4b). For this analysis, we discarded the first four million base pairs of chromosome 8L because we lacked information on whether this region is also sex-linked (Fig. 3.1).

We note that nucleotide diversity in the sex-linked portion of the female sex chromosomes includes fixed differences between the W and Z chromosomes and also positions that are segregating on the Z chromosome. Thus, this measurement is influenced by demographic differences between the female and male (the female genome is more polymorphic; Fig. 3.4). However, we found that standardizing the estimates of nucleotide diversity by the genome-wide average for each individual (by dividing diversity estimates from the male or female genome by the corresponding genome-wide mean for each genome) did not affect the results of the linear mixed model (see Results and Appendix B1.4). In the analysis of nucleotide diversity, the sex linked portion of chromosome 8L stood out as the most polymorphic region in the female genome, supporting the existence of fixed divergent sites between the W and Z chromosomes.

The disparity between the female and male in nucleotide diversity along chromosome 8L was greater in the sex-linked portion than the pseudoautosomal portion of chromosome 8L (Wilcoxon rank sum test: $P < 0.001$; Fig. 3.4c). This result is consistent with the results of the linear mixed model (above). There was also a peak of divergence near end of the chromosome in the non-sex-linked region (Fig. 3.4c), that overlapped with a region where *X. borealis* daughters were mostly inheriting the same allele, suggesting partial sex-linkage (Fig. 3.1). This could be due to an inversion, although we did not explore this possibility in our data.

Within coding regions, dN and dS were very slightly, but significantly (statistically) elevated in the sex-linked region of *X. borealis* compared with the rest of the genome for both the female and male, but dN/dS was not (based on a permutation test; see Appendix B1.5). But, unlike the analysis of all SNPs (above), which included more data, the sex linked region was not the highest for any value (dN, dS, or dN/dS) compared individually to the other chromosomes. This emphasizes the subtlety of the divergence in the sex linked region and indicates that the time since recombination suppression is recent. We did not recover evidence of substantial differences in coverage between the female and male on the sex chromosomes (see Appendix B1.6).

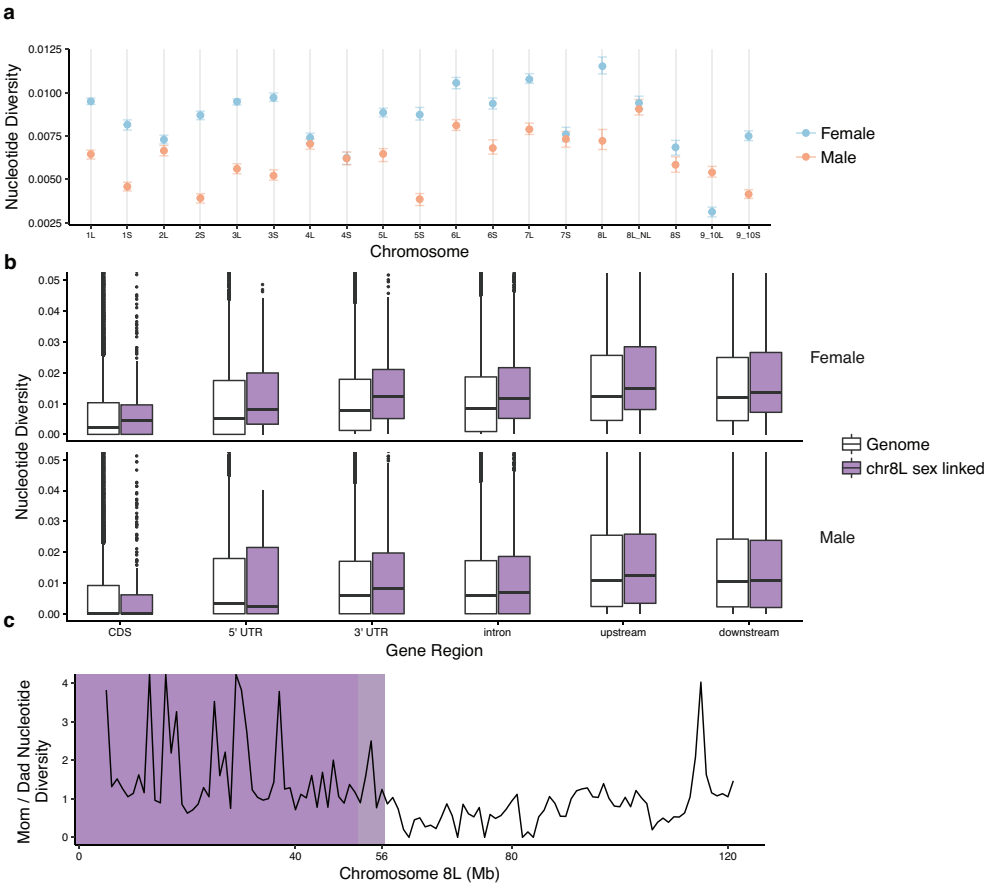


FIGURE 3.4: Nucleotide diversity (π) in *X. borealis* based on WGS data mapped to the *X. laevis* reference genome. (a) Median π by chromosome as measured in the six genomic categories; error bars indicate 95% CI bootstrap estimates (for further information on differences see Appendix B1.4). The 8L_NL category refers to the diversity measured on chromosome 8L in the non-sex-linked region (57 Mb–120 Mb). (b) Box and whisker plot of π across six genomic categories (described in Methods); the y-axis is truncated at 0.05 for clarity. (c) Standardized nucleotide diversity of the female divided by the standardized nucleotide diversity of male in 1 Mb windows across chr8L; the completely sex-linked region is highlighted in dark purple, and the significantly sex linked region with suppressed recombination in light purple (see Fig. 3.1).

3.4 Discussion

3.4.1 More Expansive Recombination Suppression on Younger Sex Chromosomes

The homomorphic sex chromosomes of *X. borealis* and *X. laevis* experienced distinctive evolutionary histories since they originated. In *X. laevis*, the sex-linked region is restricted to a small portion on the end of a chromosome (2L). In *X. borealis*, however, the sex-linked region encompasses almost half of a chromosome (8L; Fig. 3.1), even though this sex chromosome system is thought to be derived with respect to the sex determination system of *X. laevis* (Furman and Evans 2016). Within the region of suppressed recombination of both of these species, there is evidence of sex chromosome divergence at the molecular level (*X. borealis*: Fig. 3.4a-c, Appendix B1.5; *X. laevis*: Mawaribuchi et al. 2016). Although the magnitude of sex chromosome divergence in the large sex-linked region of *X. borealis* is modest, it appears that recombination has been suppressed over sufficient evolutionary time for these differences to be detectable, presumably for many thousands of generations or more. Supporting this, our second family of lab-reared *X. borealis* and the surveyed panel of adults also had completely suppressed recombination in this large region (there were some sex linked female heterozygous sites that appeared in both families and others that were unique to one family or the other, see Appendix B1.3). Together, these findings are consistent with observations made in other, more diverged species that the extent of recombination suppression need not be more expansive in older than younger sex chromosomes (reviewed in Wright et al. 2016). They further demonstrate that newly established sex chromosomes may assume radically different evolutionary trajectories.

We infer here that the younger sex chromosomes of *X. borealis* have a larger region of suppressed recombination than the older sex chromosomes of *X. laevis*. One possibility is that this is due to a large scale genomic change, such as an inversion or deletion leading to widespread recombination suppression (Charlesworth et al. 2005). We were unable to characterize rearrangements in the sex chromosomes of *X. borealis* here due to the nature of our WGS data (short reads and relatively low coverage). However, there were two crossover events detected in the sex linked region (Fig. 3.1,B1.3). As well, the level of divergence between the W and Z was lower in the last 1/3 of the sex linked region, consistent with a more recent cessation of recombination (and possibly indicating the presence of genomic regions – strata – with different levels of divergence). These results suggest that a single large scale inversion encompassing the entire sex-linked region is not a likely reason for suppressed recombination. We cannot rule out the possibility that there are smaller inversions within the sex linked region that causes recombination suppression in flanking regions. In some sex chromosome systems, inversions are not thought to be the driver of recombination suppression. For example, in the plant *S. latifolia*, inversions in the nonrecombining portion of the sex chromosomes may have occurred after recombination suppression evolved (Bergero et al. 2008). We did not recover any evidence of major coverage differences between the sequenced female and

male *X. borealis* (Appendix B1.6), suggesting a lack of deletions or insertion differences between the Z and W. However, our inference is limited by a lack of a con-specific reference genome, because unique or rapidly evolving sequences on the sex chromosomes of *X. borealis* may not map to the homologous portion of or be present in the *X. laevis* reference genome.

Alternatively, modifiers of recombination can be favored by natural selection to suppress recombination (Charlesworth et al. 2005; Coop and Przeworski 2007). These genetic factors control chiasmata formation during meiosis, possibly by modifying chromosome structure, or via the action of genes or repetitive elements (Ji et al. 1999; Otto and Lenormand 2002). Curiously, chromosome 8S in *X. borealis* also had a lower recombination rate that other chromosome linkage maps of similar size (Fig. 3.2). This chromosome is homeologous (i.e., related by genome duplication) to the sex chromosomes 8L (Session et al. 2016). This result offers the intriguing possibility that whatever is acting to suppress recombination on the sex chromosome may also influence recombination of homeologous sequence on chromosome 8S (genome-wide, the L and S nucleotide divergence is about 6%; Session et al. (2016)). This is unlikely to be an artifact of mapping errors because linkage groups would not form from markers that were a mix of chromosome 8L and 8S, because SNPs on different chromosomes should have a recombination fraction of about 0.5 (above our threshold; Materials and Methods).

Sex-linkage with minimal divergence (similar to our observations in *X. borealis*) has also been found in other species. For instance, the Japan sea population of stickleback fish have a recently evolved set of sex chromosomes, which were generated by a fusion of the ancestral sex chromosome and an autosome (Kitano et al. 2009). In this system, recombination suppression spread from the point of sex chromosome fusion to an ancestral autosome along a large fraction of the neosex chromosome (Natri et al. 2013). Sex-linked genomic regions with variable levels of divergence suggest that the boundaries of recombination suppression evolve over time, and may encompass areas that are not yet diverged. As such, recombination may occasionally happen in these regions until a hard recombination boundary is established (Bergero and Charlesworth 2009). In some other amphibians, periodic recombination may prevent divergence of the sex chromosomes (Perrin 2009; Stöck et al. 2011; Dufresnes et al. 2014). Though recombination was not detected in this region for either family of *X. borealis*, it is possible that over long timescales the sex chromosomes of *X. borealis* may occasionally recombine. However, the divergence detected here between the Z and W, though modest, indicates that recombination is not happening frequently enough to completely prevent divergence (Fig. 3.4).

3.4.2 The Relative Ages of the Sex Chromosomes of *X. laevis* and *X. borealis*

Our inference that recombination suppression expanded more quickly in *X. borealis* than *X. laevis* is based on (i) the inferred origin of *DM-W* in subgenus *Xenopus* after the

whole genome duplication event shared by all extant subgenus *Xenopus* species (Bewick et al. 2011) and (ii) inferred phylogenetic relationships within subgenus *Xenopus* (Furman and Evans 2016), which indicates that the *DM-W* based sex determination system is ancestral to the system of *X. borealis* (Fig. 3.1). If this phylogenetic inference were erroneous and instead the sex determining system of *X. borealis* were ancestral to the *DM-W* based system of *X. laevis*, the rate that recombination suppression expanded over the sex chromosomes of *X. borealis* could be slower than it seems here.

However, there are several lines of evidence that argue against *X. borealis* having the older sex chromosomes than *X. laevis*. First, the strongest phylogenetic signal found using 1,585 genes supports a paraphyletic clade of *DM-W*-possessing species (Fig. 3.1; Furman and Evans 2016). More specifically, the alternate hypothesis of monophyly of *DM-W*-possessing species is supported by substantially fewer genes than the hypothesis of paraphyly of *DM-W*-possessing species with a sister relationship between *DM-W*-possessing *X. clivii* and *X. borealis* (as presented in Fig. 3.1; Furman and Evans 2016). In fact, the hypothesis of monophyly of *DM-W*-possessing species has an equal support to another paraphyletic relationship among *DM-W*-possessing species where *X. borealis* is more closely related to *X. laevis* than *X. clivii* is to *X. laevis* (Furman and Evans 2016).

Additional evidence against the possibility of older sex chromosomes in *X. borealis* is provided by divergence of orthologous autosomal genes of *X. borealis* and *X. laevis* (e.g., divergence of synonymous site of ~14%; Chain et al. 2008) that is substantially greater than that observed between the nonrecombining regions of the *X. borealis* sex chromosomes (Fig. 3.4). Likewise, homeologous coding sequences (including nonsynonymous and synonymous sites) also have higher divergence (~7%; Evans and Kwon 2015) than the non-recombining region of the *X. borealis* sex chromosomes. These genomic patterns are consistent with the proposal that suppressed recombination in the sex chromosomes of *X. borealis* occurred after allotetraploidization. Thus, even if previous phylogenetic inferences (Furman and Evans 2016) are incorrect, the level of divergence between these sex chromosomes still argues that the expansion of the nonrecombining region occurred after the origin of *DM-W* (i.e., post-whole genome duplication in subgenus *Xenopus*) after or at least within a similar time frame.

3.4.3 More Recombination in Females than Males, and in Different Genomic Regions

Heterochiasmy refers to differences in sex-specific rates of recombination. Here, in two independently derived sex chromosome systems with female heterogamy, we observed heterochiasmy with females having a higher rate of recombination than males. In some species of bird and crab with female heterogamy, recombination rates appear to be similar between the sexes (Groenen et al. 2009; Backström et al. 2010; Cui et al. 2015; Nietlisbach et al. 2015). But in some fish and other bird species the rate of recombination is higher in heterogametic females (Hansson et al. 2010; Ruan et al. 2010), or higher in homogametic males (Kawakami et al. 2014). In vertebrates with male heterogamy, the

rate of recombination is often higher in females, particularly in XY mammals (Wong et al. 2010; Ottolini et al. 2015), though exceptions are known where rates are similar between the sexes, or higher in males (Mank 2009a; Johnston et al. 2016, respectively).

In several other frog species with male heterogamy, heterochiasmy has been observed with a higher recombination rate in females (Berset-Brändli et al. 2008; Brelsford et al. 2016). This was interpreted to be consistent with the Haldane-Huxely Rule (Haldane 1922; Huxley 1928) which postulates that when one sex does not recombine (i.e., when one sex is achiasmatic), that sex is the heterogametic sex (Berset-Brändli et al. 2008; Brelsford et al. 2016). Our results suggest instead that in species with heterochiasmy, the sex with lower recombination is not strongly linked to which sex is heterogametic (Lenormand and Dutheil 2005). Heterochiasmy may be more prominently influenced by haploid selection (Lenormand and Dutheil 2005), sexual antagonism (Mank 2009a), or other explanations.

The locations of recombination events were sex-biased in both species of *Xenopus* investigated, with recombination most frequent in the center of chromosomes in females, versus the ends of chromosomes in males (Fig. 3.3). Sex specific differences in crossover location have been observed in other taxa, including, for example, frogs, dogs, and primates (Wong et al. 2010; Venn et al. 2014; Ottolini et al. 2015; Brelsford et al. 2016). Female linkage map length (in cM) and the number of crossover events was positively correlated with the amount of bp covered by the map and the total length of a chromosome, whereas in males this relationship was not observed (Fig. 3.2). A similar disparity between the sexes in the relationship of cM and Mb spanned by linkage maps has been observed in the frog *Hyla arborea* (Brelsford et al. 2016) and in humans (Ottolini et al. 2015). This sex specific difference could be due to the differences in recombination location. In females, because recombination is spread out across the middle of chromosomes, longer chromosomes may permit more recombination events to occur without crossover interference. In males, where recombination occurs mostly on the tips of chromosomes, crossover interference is less likely to vary among chromosomes with different lengths. Similar findings have been recovered in soay sheep, where male recombination is mostly biased to the last 18 Mb of each of the chromosome tips, with chromosomes ranging in size from ~50–200 Mb (Johnston et al. 2016), encompassing the chromosome length variation of *Xenopus* (Session et al. 2016). Why females and males have differences in recombination locations is potentially due to differences in meiosis. During speramtogenesis there appears to be more control over formation and number of crossover events compared with oogenesis, with crossovers stopping in the presence of errors and more often restricted to one per arm (Hunt and Hassold 2002; Hassold et al. 2004; Coop and Przeworski 2007). As well, maintenance of favorable allelic combination by haploid selection, which is generally stronger in males, may limit the breadth of possible crossover locations to genomic regions, such as chromosome tips, that have low gene density (Lenormand and Dutheil 2005).

One possible caveat to our conclusions on sex specific differences in recombination rate is that in some cases maternal and paternal linkage groups spanned nonoverlapping

genomic regions, which themselves may vary in the local rate of recombination (Groenen et al. 2009; Kawakami et al. 2014; Ottolini et al. 2015). Since male recombination rate is biased towards tips of chromosomes (Fig. 3.3), it is possible that crossover events were not accounted for in these linkage maps if tags do not span to the ends of chromosomes. Kawakami et al. (2014) also noted that RAD based studies in birds may also underestimate linkage map lengths, because they underrepresent microchromosomes and ends of chromosomes. In this study, the disparity between female and male linkage map lengths in *X. laevis* (1.2:1 ratio of map length) is much less than *X. borealis* (4.4:1). The total map lengths in *X. laevis* (females: 1,572 cM and males: 1,275 cM) was not far from a total map length of 1,800 cM, which is the expected length if there were an obligate rate of one crossover per chromosome arm. This suggests our estimate of recombination in *X. laevis* is not unreasonably low. As well, the female to male map length ratio in *X. laevis* of 1.2:1 is within the range of a wide variety of other species (1.4:1 for a fish, Ruan et al. 2010; 1.2:1 for a mammal, Wong et al. 2010; 1.1:1 for a bird, Kawakami et al. 2014). Thus, the sex specific differences detected in *X. laevis* are likely genuine. We note that the magnitude of the sex difference in recombination rate for *X. borealis* (females: 719 cM and males: 165 cM) may be exaggerated due to lower genomic coverage in the *X. borealis* family (though large differences in recombination between closely related species is known Kawakami et al. 2014). Furthermore, our linkage maps are not capturing all recombination events in either species because the per gamete rates of recombination are much less than the expectation of one event per chromosome of 18 (Results). As such, caution should be used when interpreting linkage maps from reduced genome sequencing technologies (e.g., RADseq, GBS), especially when a closely related reference genome is lacking to assess marker distribution across chromosomes.

3.4.4 Drivers of Sex Chromosome Evolution and Stasis

Information from a diversity of organisms suggest that the age of sex chromosomes is not a strong predictor of the amount divergence between sex chromosomes within a species (Wright et al. 2016). Our findings from the sex chromosomes of *X. borealis* and *X. laevis* support this inference. One possible explanation for these observations is that the genomic context in which a new sex chromosome system is established plays a large role in determining the extent of divergence a newly established will experience. For example, the ability to cope with dosage imbalances or the potential for dosage compensation mechanisms to evolve could strongly influence whether sex chromosomes become heteromorphic or not (Batada and Hurst 2007, but see Mank 2009b). If, for instance, the sex chromosomes of *X. laevis* (chromosome 2L), contains more dosage sensitive genes than the sex chromosomes of *X. borealis* (chromosome 8L), this could hinder the expansion of recombination suppression in *X. laevis* but not *X. borealis*. In ratites, for example, an inability to accommodate dosage imbalances may prevent sex chromosome divergence beyond the limited regions thought to no longer recombine (Adolfsson and Ellegren 2013; Vicoso et al. 2013; Yazdi and Ellegren 2014). As well, the life history or ecological context of a population can influence the fate of sex chromosomes. Guppies, which similar to

X. borealis have a large sex linked region without extensive degeneration, show variability in the extent of sex linkage on the chromosomes depending on an interplay between the strength of sexual antagonism and predation pressures in the population (Wright et al. 2017). A compelling direction for further inquiry is to explore factors that govern sex chromosome divergence and stasis in African clawed frogs, including the role of natural selection (e.g., favoring balanced gene dosage between the sexes, sexually antagonistic selection, haploid selection (Rice et al. 1994; Lenormand 2003; Adolfsson and Ellegren 2013)), and nonselective events (e.g., recombination in sex reversed individuals; Perrin (2009), or large scale inversions).

Supplementary Material

Supplementary figures and tables are available in Appendix B.

3.5 Acknowledgements

We thank Brian Golding for providing computational resources. We also thank Natural Sciences and Engineering Research Council (NSERC) for funding support (CGSD3-475567-2015 to BLSF, RGPIN/283102-2012 and RGPIN-2017-05770 to BJE).

Part II

Whole Genome Duplication

Chapter 4

Divergent subgenome evolution after allopolyploidization in African clawed frogs (*Xenopus*)

Benjamin L. S. Furman*, Utkarsh J. Dang†, Ben J. Evans*, G. Brian Golding

*Biology Department, Life Sciences Building room 328, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada
†Department of Health Outcomes and Administrative Sciences, School of Pharmacy and Pharmaceutical Sciences, Binghamton University, State University of New York, Binghamton, USA

This chapter has been submitted to the *Journal of Evolutionary Biology* and is currently under review.

Abstract Whole genome duplication (WGD), the doubling of the nuclear DNA of a species, contributes to biological innovation by creating genetic redundancy. One mode of WGD is allopolyploidization, wherein each genome from two ancestral species becomes a 'subgenome' of a polyploid descendant species. The evolutionary trajectory of a duplicated gene that arises from WGD is influenced both by natural selection and by gene silencing (pseudogenization). Here, we explored how these two phenomena varied over time and within allopolyploid genomes in several allotetraploid clawed frog species (*Xenopus*). Our analysis demonstrates that, across these polyploid genomes, purifying selection was greatly relaxed compared to a diploid outgroup, was asymmetric between each subgenome, and that coding regions are shorter in the subgenome with more relaxed purifying selection. As well, we found that the rate of gene loss was higher in the subgenome under weaker purifying selection and has remained relatively consistent over

time after WGD. Our findings provide perspective from vertebrates on the evolutionary forces that likely shape allopolyploid genomes on other branches of the tree of life.

4.1 Introduction

Whole genome duplication (WGD) creates redundancy in genetic pathways and can lead to biological innovation (Ohno 1970). For instance, WGD is thought to have contributed to phenotypic diversity in jawed vertebrates, which experienced at least two rounds of WGD (2R hypothesis) (Dehal and Boore 2005), and the success of angiosperms, which experienced numerous WGD events (Fawcett et al. 2009; Jiao et al. 2011). However, despite the potential evolutionary advantages of WGD, the most common evolutionary outcome of a gene pair generated by WGD (homeologs) is that one becomes non-functional ("pseudogenization") (Lynch and Conery 2000).

One possible route for both gene copies to be retained is "neofunctionalization", where one homeolog acquires a novel function (Ohno 1970). Another route involves a partitioning of ancestral function among duplicated genes, thus making both copies necessary ("subfunctionalization") (Force et al. 1999; Stoltzfus 1999). Additionally, because WGD doubles entire genetic pathways, natural selection may favor the functional persistence of gene duplicates in order to maintain the stoichiometric balance of epistatic interactions among the protein products of duplicated genes (Papp et al. 2003; Gout et al. 2010; Qian et al. 2010).

Post-WGD, genes involved in dosage sensitive functions, such as protein complexes and transcription factors, were preferentially retained in several species (Blanc and Wolfe 2004; Makino and McLysaght 2010; McGrath et al. 2014). However, analysis of ancient WGD events indicate that many of these genes will also eventually be lost, with only a small proportion of duplicates surviving over the long haul: 8% for yeast (Scannell et al. 2006); 18% for teleost WGD (Inoue et al. 2015); likely $< 10\%$ for 2R vertebrate WGD (Dehal and Boore 2005); 20-30% for *Brassicaceae* WGD (Liu et al. 2014). One exception appears in *Paramecium*, in which 40-50% of its homeologous pairs have remained functional after 350 my (Aury et al. 2006; McGrath et al. 2014).

There is evidence that some genes are rapidly lost after WGD (Scannell et al. 2006; Inoue et al. 2015). This makes sense if gene copies are initially functionally redundant, and if natural selection to retain both homeologs is correspondingly weak. However, if gene dosage is important after WGD, it may take a long time for gene loss to be selectively neutral (Gout and Lynch 2015). Thus, the rate of pseudogenization could be constant or potentially increase over time (Gout et al. 2010; Gout and Lynch 2015). As well, if WGD is the result of allopolyploidization, duplicates may not be fully redundant due to divergence in lower-ploidy progenitors (Adams 2007). This divergence could also introduce biases in rates of pseudogenization between each half of an allopolyploid genome, i.e., each subgenome (Comai 2000; Evans 2007).

Rates of pseudogenization and the extent of purifying selection on homeologs are interrelated and potentially dynamic through time. Thus, to best understand the interactions of each one, a comparative approach – that uses data from multiple species in a time-calibrated phylogenetic context – is paramount (Inoue et al. 2015). To better understand these phenomena, we examined several tetraploid African clawed frog (*Xenopus*) species that are derived from a shared allotetraploid ancestor (4x=2n=36 chromosomes; haploid number (n) of 18 chromosomes) (Tymowska 1991; Evans et al. 2015). Each of these allotetraploid species have two subgenomes of 9 homologous chromosome pairs each (the 'L' and the 'S' subgenome; Matsuda et al. 2015), that were inherited from different diploid ancestral species that generated the shared allotetraploid ancestor. Estimates for the time for the initial allopolyploid WGD range from as recently as 17 my ago (Session et al. 2016) to between 25–65 my ago (Chain and Evans 2006; Hellsten et al. 2007; Bewick et al. 2011; Furman and Evans 2016), depending on the calibration point and analytical methods used. In one of these species, *X. laevis*, about 60% of the homeologous pairs are functional, and there exists substantial asymmetry in subgenome evolution (Session et al. 2016). For instance, the S-subgenome of *X. laevis* experienced more genomic rearrangements compared to the L and has fewer intact and functional genes (Session et al. 2016).

Using a phylogenetic framework and sets of expressed gene sequences from Furman and Evans (2016), we explore duplicate gene evolution and pseuodgenization post-WGD in several allotetraploid *Xenopus* species. Our findings indicate that selection is substantially relaxed post-WGD, and has not returned to pre-duplicate levels. The extent of this relaxation differs between the two subgenomes and the S-subgenome with a more relaxed level of purifying selection has shorter coding sequences. Using a probabilistic model in a maximum likelihood framework (an extension of Dang et al. 2016), we also found that the rate of pseudogenization is higher in the S-subgenome across the *Xenopus* clade. Furthermore, we find that these rates have remained relatively constant over time. Our results are consistent with those of the *X. laevis* genome sequencing project (Session et al. 2016), but extend them by demonstrating across multiple polyploid species that subgenome differences are a phenomena prevalent across the whole *Xenopus* subgenus. We conclude that genome restructuring post-WGD is an ongoing feature of the *Xenopus* subgenus, drawn out over millions of years.

4.2 Methods

4.2.1 Homeolog Identification

Sequence data analyzed in this study was obtained from a recent phylogenetic analysis of *Xenopus* (Furman and Evans 2016). The dataset was generated from a total of six species (1 diploid, 5 allotetraploids), including previously published RNAseq data from four *Xenopus* (*X. borealis*, *X. clivii*, *X. largeni*, and *X. allofraseri*) and downloaded

Unigene libraries from *X. laevis* and *Xenopus Silurana tropicalis* (Unigene database, last modified March 2013).

As described in Furman and Evans (2016), we identified homeologous and orthologous sequences using a reciprocal BLAST approach (Altschul et al. 1997a) with each species transcriptome assembly and the Unigene database of *X. laevis* to the *X. tropicalis* database. Multiple rounds of tree building and parsing allowed us to distinguish orthologous from homeologous sequences. We used a bioinformatic filter that required closer interspecific than intraspecific relationships among orthologous sequences to match an expectation associated with WGD preceding speciation of the allopolyploids (see Furman and Evans (2016) for full details). We analyzed only those gene alignments that had data for both homeologs for at least one species. Orthologous sequences for each species included only the longest coding region for each gene and allelic and splice variants were not analyzed. We considered only those alignments with at least 300 bp of ungapped sequence. These data included a total of 1,585 genes. From these data, we realigned each using the codon aligner MACSE (Ranwez et al. 2011). We then removed alignments with no sequence data for *X. laevis* (a constraint of our pseudogenization model, see below), leaving 1,503 genes.

In the original dataset of Furman and Evans (2016), homeologous lineages in each alignment were randomly assigned the label of "alpha" or "beta", which did not consistently refer to the same subgenomes of *Xenopus* because the genome sequence from Session et al. (2016) was unavailable for genomic analyses at that time. For this study, we have now been able to assign the "alpha" and "beta" lineages to the L and S-subgenomes of the *X. laevis* genome v9.1 for each alignment. To accomplish this, we selected one *X. laevis* sequence from each alignment (randomly, if two were present), and used BLAST to find the best subgenome match in the *X. laevis* genome. If the sequence did not have an L or S chromosome as the best hit, the alignment was discarded, otherwise all sequences in that lineage ("alpha" or "beta") were assigned to the corresponding subgenome match and the other lineage to the other subgenome. After these steps, the dataset encompassed 1,417 genes spanning 2,235,636 base pairs (bp) (711,340 bp ungapped characters total across alignments), with a wide range of L- or S-subgenome copies missing across species (10%:*X. laevis* – 64%:*X. clivii*).

4.2.2 Quantifying Selective Constraint Over Time

To assess selective constraints on DNA sequences over time, we obtained estimates of ω (dN/dS) that were specific to a lineage or a group of lineages (described below), using CODEML (part of the PAML package; Yang 1997, 2007). We first used a Perl script to remove any stop codons in each alignment. We then concatenated sequences across all genes for each subgenome for each species such that each allotetraploid species was represented by two concatenated sequences – one from the L- and one from the S-subgenome. To generate a starting tree topology, we used RAxML v.8.2.4 (Stamatakis 2014a) and set a `GTRGAMMA` model, followed by 500 bootstrap replicates to assess support. Strongly

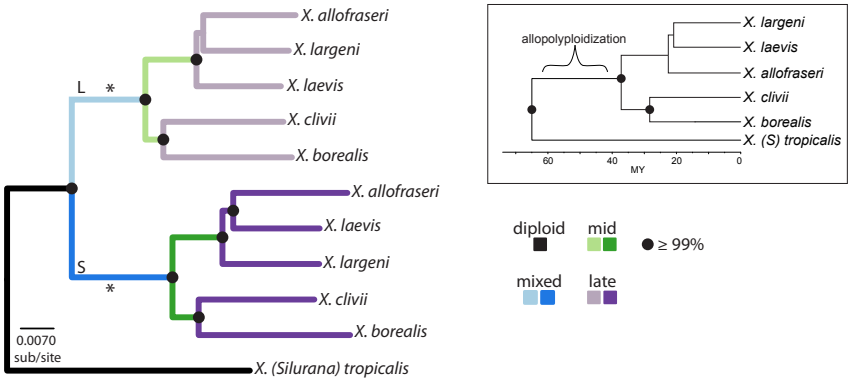


FIGURE 4.1: Phylogram recovered from a RAxML analysis of concatenated sequence data. L and S refer to the two subgenomes of *Xenopus* species (Matsuda et al. 2015), and the colors and corresponding labels (diploid, mixed, mid, late) refer to how branches were grouped for CODEML analysis and modeling of pseudogenization rates. Somewhere along the "mixed" branch an allopolyploidization event took place (indicated by an asterisk), generating a tetraploid ancestor of all extant *Xenopus*. The different resolution of sister relationships among the *X. laevis*, *X. allofraseri*, and *X. largeni* clade between the L and S-subgenomes reflect the poor resolution obtained by any analysis (Furman and Evans 2016).

supported nodes are consistent with the phylogenetic analyses of Furman and Evans (2016), but include different relationships among *X. laevis*, *X. allofraseri*, and *X. largeni* between the L-lineage compared to the S-lineage. This is due to poorly supported, short internal branch lengths, and a lower mutation rate of the L-subgenome (Fig. 4.1).

We used six evolutionary models to evaluate the impact of different subgenomes and time on purifying selection that are distinguished by the number of ω values and by which branches in the phylogeny were pooled for each ω value estimate (see Fig. 4.2 for a visualization of all models). The simplest "ploidy" model has three separate ω (dN/dS) values that are defined by the ploidy of the branches. One ω value was estimated for the diploid branch, one was estimated for the pair of branches where the whole genome duplication took place across both subgenomes (the "mixed" lineages that has part diploid and part tetraploid histories), and one ω value was estimated for the other branches that are entirely allotetraploid.

From this simplest model accounting for just differences in ploidy, models took one of two forms. Either, they tested for the effect of different subgenomes on purifying selection, or they tested the effect of time since duplication on purifying selection. A final model evaluated both of these factors together. The first 'subgenome' model extends the 'ploidy' model by two parameters, independently estimating ω values for each of the mixed ploidy branches (one for L and one for S) and for each subgenome of the tetraploid branches (again, one for L and another for S) (a total of five ω parameters). Extending

this 'subgenome' model by two more parameters, the 'subgenome-species' model estimates ω separately for each of the two major *Xenopus* clades (the laevis/largeni/fraseri clade, and the borealis/clivii clade), within each subgenome (a total of seven ω parameters). This was done because the tempo of speciation in each clade is quite different, with the most recent common ancestor (MRCA) of the borealis/clivii clade being much older than the other clade.

In the other set of models the allopolyploid branches were divided into groups that represented distinct time intervals following WGD, ignoring the different subgenomes. Extending the most simple model of just differences in 'ploidy' by one parameter, the first 'time' model estimated ω values separately for the internal allopolyploid branches (referred to in Fig. 4.1 as the 'mid' tetraploid branches), and the tip branches (referred to as the 'late' tetraploid branches) (a total of four ω values). This set of 'late' tetraploid branches does include a small internal branch in the laevis/largeni/fraser clade that stems from an unresolved node (Fig. 4.1). Then, extending this 'time' model by two parameters, the 'time-species' model separately estimating ω in each of the major allotetraploid clades, as described above, for both the 'mid' and 'late' tetraploid branches (a total of six ω values).

The last, and most complex 'time-subgenome-species' model, tested the effect of both time and subgenome in tandem. This model extended the ploidy model by estimating ω values within each subgenome (L and S), within each major clade (borealis/clivii and laevis/largeni/fraseri), and separately for each time interval (mid and late) since duplication (a total of 11 ω values).

By analyzing these epochs with unique ω parameters, we assessed how selective constraints on duplicate copies changed over time. As well, by assigning subgenome of origin, we evaluated whether each genome that contributes to an allopolyploidization event experiences different selective constraints after allopolyploidization. For the best fit model, we estimated 95% confidence intervals for the ω parameters by analyzing 100 bootstrap replicates where codons were re-sampled with replacement. For model selection, we used the Bayesian information criterion (BIC) (Schwarz 1978):

$$\mathrm{BIC} = 2l(\hat{\boldsymbol{\Theta}}) - \log n \times m,$$

where n is the number of codons in the alignment (745,212), and $\hat{\boldsymbol{\Theta}}$ is the log-likelihood estimate of the model. m is the number of parameters estimated for each model and included nine parameters for the estimated codon frequencies (`CodonFreq = 2`), 19 parameters estimated for the branch lengths (`clock = 0`, $2 * n - 3$, where n is the number of tips, as outlined in the PAML manual), one for the estimated value of κ (the transition/transversion ratio, `fix_kappa = 0`), and then the estimated number of ω values for each model (either three, four, five, six, seven, or 11; see outlined models above and Fig. 4.2). Thus, m was 32, 33, 34, 35, 36, 40 for the "ploidy", "time",

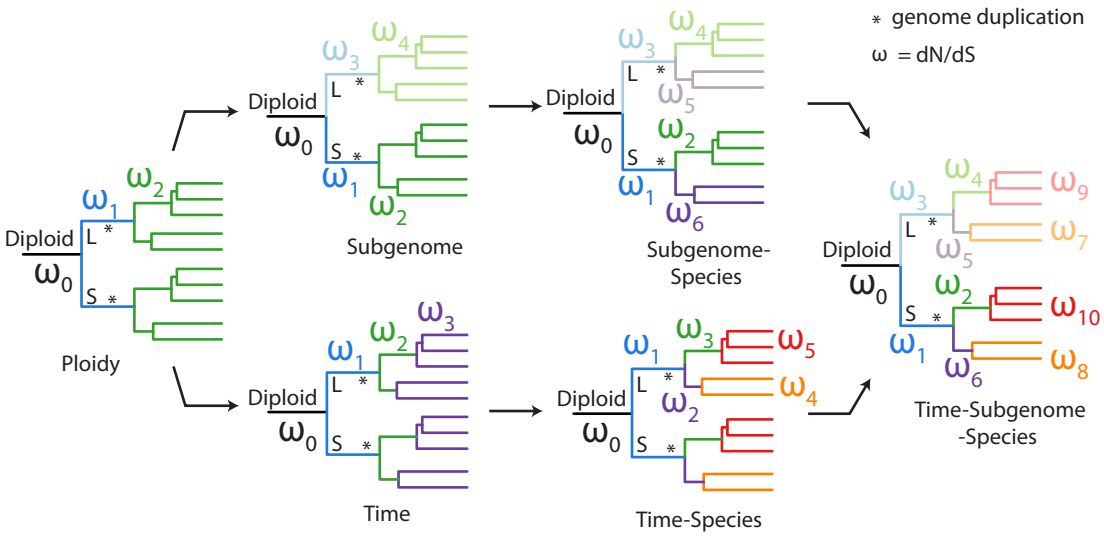


FIGURE 4.2: Model schemes for the CODEML analysis of purifying selection. The * marks the branches where the whole genome duplication took place, thus these branches were diploid for some length of time and tetraploid for the rest. The diploid branch is the lineage that extends to *X. tropicalis*. Analyses were performed on unrooted trees. Colors reflect branches that have a distinct dN/dS value estimated for them.

"subgenome", "time-species" ,"subgenome-species", and "time-subgenome-species" models, respectively.

4.2.3 Coding Sequence Length

To evaluate changes in selective constraints as evidenced by the evolution of premature stop codons, we tested if homeolog coding sequences were of different lengths between the subgenomes. To accomplish this, we first measured the ungapped sequence length of each of the ingroup sequences for each alignment (using a Perl script), retaining only those that had copies from both subgenomes. We then analyzed these data using a Markov chain Monte Carlo generalized linear mixed model (MCMCglmm) using the R package MCMCGLMM (Hadfield 2010). We set as a fixed effect the subgenome of origin (L or S), and used random effects of gene and phylogeny. The phylogeny used was the species tree estimated by *BEAST, described in more detail in the pseudogenization model analysis below. We ran the Markov chain for 1000000 generations, with a 10000 generation burn-in and a thinning interval of 500. We set an inverse-Gamma prior for both random effects (phylogeny: $V = 1, \nu = 2$; gene: $V = 1, \nu = 0.002$) and the residual effect ($V = 1, \nu = 0.002$, and a Gaussian family link; following similar example analyses in the package documentation and in Garamszegi (2014), and we modified this method to ensure adequate exploration of the posterior distribution of parameter values).

4.2.4 Modeling Variation in the Rate of Duplicate Gene Loss Over Time

We used a model-based approach to infer the rate of pseudogenization in our duplicate gene copy dataset over different time intervals demarcated by speciation events, and allowing for different pseudogenization rates in each subgenome. Because our data are from transcriptomes, pseudogenization of duplicate gene copies here represents genes that are either no longer expressed, or expressed in a low enough amount to not be detected. The latter category of genes (low expression) are quite possibly on their way to becoming pseudogenized (Gout and Lynch 2015), and we explore this further in the discussion section. Additionally, there may be genes not expressed in the tissue we analyzed (liver) at the developmental stage we surveyed (adult), and these would be picked up as "pseudogenized" genes in our analysis. If it is common for genes to be missing for these reasons, our estimate of the rate of pseudogenization will be upwardly biased. But, because our interest is in comparing subgenomes to one another and time periods to one another (and subgenomes across time periods), we focus on comparisons of estimated rates rather than the magnitudes of these rates. Our model does attempt to estimate the proportion of gene copies that were not detected due to missed missing data (e.g., low expression level, low sequencing coverage). Whole genome sequencing of these species and analysis using the model developed below could help further account for some of these issues.

In the context of accounting for duplication and loss events using the principle of parsimony, Eulenstein et al. (2010) note that accounting for loss events can be problematic, because "it is impossible to differentiate between gene loss and missing data". Here, the probabilistic methodology utilized accounts for sampling bias and missing data, based on the framework used in Dang et al. (2016), while evaluating variation in the evolutionary rates of duplicate gene loss following WGD. While such models have also been used by Han et al. (2013) to correct for missing data, like the models in Dang et al. (2016), our modeling needs were different for this data. The models by Han et al. (2013) were implemented for gene family size-type data, the models used a birth death process, and there was no way to constrain the transitions between the different states (which we needed to do, see below). We also simultaneously estimate what they call the errormodel matrix (and provide standard errors) along with the rates. In contrast, Han et al. (2013) first estimate the errormodel matrix (without knowledge of this already from an external source) and then the gene family size expansion or reduction rates.

Here, we investigate whether there is a difference in rates soon after the WGD event (corresponding to the "mid tetraploid" group) compared to later in time ("late tetraploid" group; see Fig. 4.1). As well, we investigate the rates for each of these time periods within each of the subgenomes (L and S). Similar to the CODEML analyses, the rates are estimated individually for the borealis/clivii clade and the laevis/largeni/allofraseri clade because the "mid tetraploid" branches in each clade have different lengths.

Model Overview

Markov models have been successfully used to estimate evolutionary rates (e.g., insertion and deletion rates) of gene families in closely related sequences (Hao and Golding 2006; Marri et al. 2006; Cohen and Pupko 2010). These likelihood-based analyses typically require that the sequences being investigated have complete genome sequences available to ensure that phenomena such as genome rearrangement does not affect detection of homeologs (Hao and Golding 2006). Because we are working with transcriptome data, gene absence may not be due to pseudogenization only, and may additionally reflect incomplete gene copy detection (see details below). Here, using a continuous time Markov chain model, we simultaneously estimate the rate of pseudogenization and the fraction of missing (or mis-recorded) data in the fashion of Dang et al. (2016). The model constructed here also accounts for substantial sampling bias in the data due to the aforementioned constraints imposed by Furman and Evans (2016) during dataset construction (see details below).

We first re-coded the data from each allotetraploid species in the form of the number and type of gene copies present in each species in each gene alignment. An observation of (1,1) denotes the presence of two gene homeologs (L and S), (0,1) and (1,0) indicate presence of one homeolog and not the other (only L, or only S, respectively), and finally (0,0) would indicate that neither homeolog was recorded as present for a species. Thus, for a given gene, our data consists of membership in the form of (1,1), (1,0), (0,1), and (0,0) categories for each species. These form a phyletic pattern of presence/absence of homeologous sequence for the species in the phylogenetic tree (Table 4.1). The tree used in this analysis is based of the topology obtained by the *Beast (Heled and Drummond 2010) analysis performed by Furman and Evans (2016), and is the same as was used for the coding length analysis above. To obtain a chronogram (the original analysis was done without calibration points), we used mcmctree (Yang 2007) and a 65 million year (±4.62) divergence estimate of the *Silurana* and *Xenopus* subgenera (Bewick et al. 2011). The original phylogenetic analysis included 73 genes (representing a set of genes that had orthologous sequence for each of the five species and at least one homeologous sequence; i.e., five L-subgenome sequences plus one or more S-subgenome sequences, or five S-subgenome sequences with one or more L-subgenome sequences; see Furman and Evans (2016) for full details), which we partitioned into the three codon positions and concatenated all genes for each position. We then set an HKY+Γ model and ran two chains for 50 000 steps (following a 10 000 generation burnin), ensuring that acceptance parameters were within boundaries specified in the manual and that each chain reached results. We trimmed the resulting tree to only include the five ingroup taxa of interest (by removing the diploid outgroup and the homeolog lineage; Fig. C0.1).

Naively, failure to detect a gene copy could be interpreted as a sign of gene loss (pseudogenization). However, because these data are derived from transcriptome sequencing and Unigene databases, and not from perfectly assembled genomes, an undetected gene copy may be missing for technical reasons such as low sensitivity of sequencing to detect low abundance RNA, because the gene was not expressed in the tissue when the RNA

TABLE 4.1: Frequencies of the different observed states in the data. (0,1) indicates the presence of an L-subgenome homeolog and not an S homeolog; (1,0) denotes that the S homeolog is present, but not L. Less S homeologs were recovered for each species.

Species	*Dataset A* ($N = 1417$)			
	(0,0)	(0,1)	(1,0)	(1,1)
X. laevis		151	121	1145
X. largeni	370	525	372	150
X. allofraseri	314	475	397	231
X. clivii	480	470	371	96
X. borealis	346	530	404	137

was harvested, or because sequence divergence prevented detection with the reciprocal BLAST hit approach that we used. Previously, in the context of accounting for sequence errors in nucleotide models, Yang (2014) recommended that at least one genome be free of sequence errors. Erring on the side of caution, Dang et al. (2016) assumed a minimum of three taxa did not possess any missing data. Here, to distinguish missing data from pseudogenization, our model assumes that there is one species that has near perfect sampling of gene copies (i.e., any non-detected gene copies are truly due to pseudogenization), and that all species are closely related. We assume that *X. laevis* represents the near perfect sampling of gene copies – a reasonable assumption given that the data for *X. laevis* come from well curated Unigene databases generated by numerous RNA and DNA sequencing studies. Note that in our dataset, *X. laevis* has both homeologs present for 80% of the alignments, which is a greater presence of homeologous copies than the ~60% estimated from the recent genome sequencing (Session et al. 2016).

Model Details

Using a three state continuous time Markov chain, we model rates of duplicate gene pseudogenization. The three states are (0,1), (1,0), and (1,1). For this model, it is assumed that pseudogenization occurs independently for each gene in each species and at constant rates. In addition, we assume that the probability of both duplicate copies being truly pseudogenized (i.e., complete gene loss) is extremely low and so it is assumed that only one copy, either the L- or S-subgenome homeolog, is truly pseudogenized. Then, the substitution rate matrix for the Markov chain is

$$\boldsymbol{Q} = \begin{array}{c} \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{ccc} (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} - & 0 & 0 \\ 0 & - & 0 \\ \theta_S & \theta_L & - \end{pmatrix} \end{array}, \tag{4.1}$$

where the rows (columns) represent the current (future) state, respectively. The substitution rate matrix $\boldsymbol{Q}$ only allows moves from the (1,1) state to a (0,1) or (1,0) state, i.e., from a two gene copy state to having either the L or S-subgenome homeolog. Because going from two functional copies, for example (1,1), to only the L gene, namely (0,1), means pseudogenization of the S-subgenome gene (and no change to the presence of the L-subgenome gene), this is denoted as θ_S. Similarly, θ_L denotes the pseudogenization rate for the L-subgenome. Alternatively, we can model the instantaneous rate of change from a two-gene copy state to any single gene copy state, as just a single θ (not distinguishing if the L or S copy was lost). We employ both variants in the model variants outlined below. Furthermore, in assessing the impact of time on pseudogenization rates, we fit an individual pseudogenization rates matrix for individual portions of the tree, with specifics outlined below.

The model assumes that the MRCA of the extant tetraploid *Xenopus* contained both copies of all genes (i.e., the root of our five taxon tree has a (1,1) state for all genes). Thus, the model excludes the point of time from the WGD event until the speciation of extant *Xenopus* species. However, our duplicate gene dataset also does not include genes where one copy was lost during this interval, because at least one extant species must contain both homeologs (see above). Therefore, our estimated rates of pseudogenization apply only to genes that escaped immediate pseudogenization, and thus that were co-expressed duplicates prior to the speciation of the *Xenopus* allopolyploids we studied. Overall and with these assumptions and limitations, our model simultaneously corrects for incomplete gene copy membership data and provides probabilistic rate estimates of pseudogenization of duplicate gene copies, starting from a two-copy state.

The pruning algorithm (Felsenstein 1973, 1981) is employed to calculate the likelihood for the Markov chain model on the phylogenetic tree (Fig. C0.1). In the pruning algorithm, at any node i, conditional probabilities of observing data at the tips that are descendants of node i are calculated. Conditional probability vectors are calculated for all nodes in a post-order tree traversal fashion. Finally, the likelihood is calculated as a weighted sum using the conditional probability vector at the root of the tree and the prior root probabilities (π_{x_0}) of the states. Here, the prior root probability vector for states (0,1), (1,0), and (1,1) is $(0, 0, 1)'$ reflecting that at the root of the tree in Fig. C0.1, all genes of interest are known to exist in duplicate (as discussed above). Traditionally, at the tips of the tree, the conditional probability equals 1 if state g_i is observed at node i and 0 otherwise. However, Felsenstein (2004) notes that the vector of conditional probabilities $\boldsymbol{L}^{(i)}$ is not restricted to sum to one because they are probabilities of the same observation conditional on different events. Here, the missing data (0,0) and the uncertainty around (0,1) and (1,0) due to the transcriptomic nature of the data is accommodated using this definition, as in Dang et al. (2016).

We assume that complete loss of both homeologs of a gene in a species is not possible, hence an observation of (0,0) is considered to actually correspond to one of the other three states. Our model allows for the possibility that due to errors in the data collection

mechanism, an observation of a (0,0) was truly either (0,1), (1,0), or (1,1). The probability that one gene copy is not correctly recorded for a given taxa (as compared to closely related taxa on the tree) is denoted by δ, regardless of whether the copy corresponds to the L- or S-subgenomes. Then, assuming independence of the copies being successfully recorded, $(1-\delta)^2$ is the probability that both copies are correctly recorded for that taxon. Moreover, this model also allows for the possibility that a (1,1) is mis-recorded as a (0,1) or a (1,0), a (1,0) is mis-recorded as a (0,0), or (0,1) is mis-recorded as a (0,0). Here, it is assumed that each copy of each gene has an equal probability of not being recorded as present. The matrix of probabilities used in the model and in the pruning algorithm are summarized in Table 4.2. Extending this formulation to multiple taxa of interest, a vector of missing proportions $\boldsymbol{\delta}$ can be modeled such that $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_w)'$ for w number of taxa. Here, δ is assumed to be unique for all species for which a missing data proportion is modeled, i.e., for all species except *X. laevis*.

TABLE 4.2: Here, δ denotes the probability of one gene copy not being correctly recorded as present.

True \ Observed	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	δ	$1-\delta$	0	0
(1,0)	δ	0	$1-\delta$	0
(1,1)	δ^2	$\delta(1-\delta)$	$(1-\delta)\delta$	$(1-\delta)^2$

Separate from missing data, there are certain gene presence/absence patterns at the tips of the tree that are not observed in the dataset, given the constraints necessary to identify homeologs. For instance, it is required that at least one species has sequence data present for both homeologs to distinguish homeologs from orthologs (Furman and Evans 2016), thus there are no loci without at least one species having a (1,1). Additionally, to maintain data quality and retain topologically informative datasets, Furman and Evans (2016) restricted the data to contain at least 3 ingroup species, thus there will not be any genes where there are more than two (0,0) patterns for the species.

Dealing with this sampling bias is straightforward. Recall that *X. laevis* does not have any (0,0) states to fulfill the requirement that one taxon has perfect data sampling (see above). Now, let the set of all unobservable patterns be denoted by U. The number of patterns in U can be calculated in the following fashion. There are $3^4 \times 2 = 162$ patterns where a (1,1) observation is not seen in any of the taxa (and a (0,0) is not seen for *X. laevis*). The number of patterns where exactly k (0,0)s are observed in data with g taxa and a observed states is $\binom{g}{k} \times (a-1)^{g-k}$. Then, the number of patterns where exactly three (0,0)s are observed for all taxa but *X. laevis* (i.e., for four taxa) and four observed states (but *X. laevis* is observed to have only three states) is

$$\binom{4}{3} \times (4-1)^{4-3} \times 3 = 36. \tag{4.2}$$

Similarly, the number of patterns with exactly four (0,0)s can be calculated as 3. Thus for the dataset, a union of the 162 and 39 (36 + 3) patterns results in 183 unobservable patterns.

Because only certain patterns are present in the data, the excluded patterns can be corrected for in the likelihood by conditioning on observing an observable pattern (Felsenstein 1992; Lewis 2001; Hao and Golding 2006; Cohen and Pupko 2010; Dang et al. 2016). The probability of the model producing each unique unobservable pattern is calculated and then summed to estimate the probability of the model producing an unobservable pattern. Note that this quantity is the same for all duplicated genes. The complement of this quantity is the probability of seeing an observable data pattern. The likelihood of the data is then conditioned on this latter quantity. Then, the conditional probability of the h^{th} phyletic pattern is

$$L_+^h = \frac{L^h}{1 - L_-},$$

where $L_- = \sum_{s \in U} L_-^s$ and L_-^s is the probability of the s^{th} unobservable phyletic pattern. The log-likelihood for the n observed phyletic patterns can then be calculated as $l(\boldsymbol{\Theta}) = \sum_{h=1}^{n} \log(L_+^h)$.

The model is implemented in R (R Core Team 2017) and C++, with parameter estimates obtained from numerical optimization using PORT routines (Gay 1990) as implemented in the `nlminb` function in R. Code was adapted and extended from the INDELMISS (Dang et al. 2016) and MARKOPHYLO (Dang and Golding 2016) packages for R. Time expensive computations were written in C++ using RCPP (Eddelbuettel et al. 2011). We performed simulations to ensure reliable parameter recovery for the rates, and to show that the model can differentiate between missing data and pseudogenization; see Appendix C2 for details.

Model Fitting

To these data, we fit four versions of this Markov model, correcting for sampling bias in each, estimating a proportion of missing data for each of the non-*X. laevis* taxa, and generated confidence intervals for all parameters with 1000 bootstrap replicates. As mentioned, these models were all fit using the species tree, and as with the CODEML analyses, we estimate rates separately for the 'borealis/clivii clade' and the 'laevis/largeni/allofraseri clade'. The first version of this model estimated a single rate of pseudogenization for the borealis/clivii clade, and another for the laevis/largeni/allofraseri clade (i.e., homogeneous rates with no partitioning of subgenome or time). For the other three, we followed a similar partitioning scheme as the CODEML analyses above. The second model assessed the effect of time on the rate of pseudogenization, allowing for unique rates for the mid and late-tetraploid linages, within each clade, but estimated a single pseudogenization rate (i.e., not distinguishing whether an

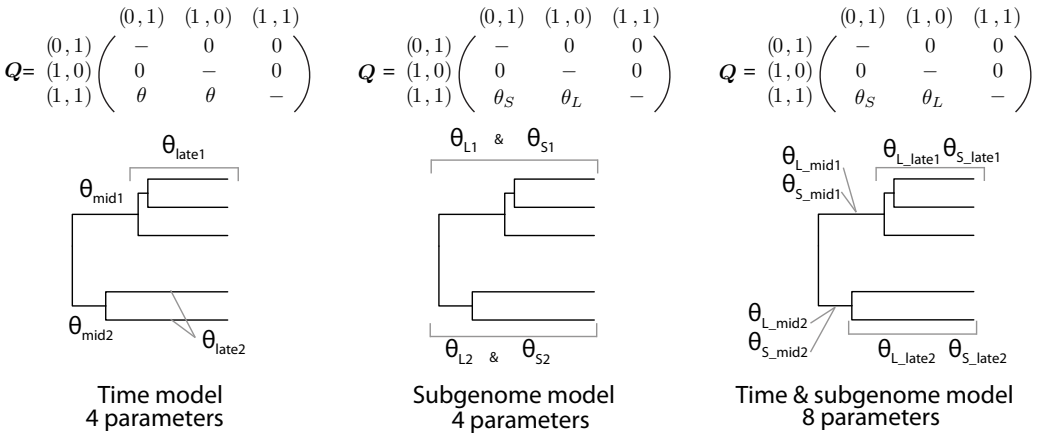


FIGURE 4.3: Model schemes for estimating the rate of pseudogenization (θ) along with the corresponding instantaneous rate matrix, indicating when one or two θs were inferred (either independently for loss of S or loss of L, or joint indicating a transition from two copies to one ignoring which subgenome copy was lost). Not shown here is a fourth model that was also assessed, wherein a single θ for each clade was estimated, with no partitioning of time or subgenome.

S or L copy was lost) for each time period. The second model assessed the effects of the subgenome on the pseudogenization rate, ignoring time, by estimating a rate for the loss of the L copy separately from the loss of the S copy, in each clade. Finally, the time-subgenome model combined these previous two models, now estimating the rate of pseudogenization for each subgenome within each time period (mid and late). See Fig. 4.3 for a visual depiction of the models and their corresponding rate matrices, and Fig. 4.1 for a visual depiction of branch groups. Comparisons of these models permit statistical evaluation of the hypothesis that there were differences in pseudogenization rates between subgenomes and between time periods. For model selection, again, we assessed model fit using BIC:

$$\text{BIC} = 2l(\hat{\boldsymbol{\Theta}}) - \log n \times m,$$

where $l(\hat{\boldsymbol{\Theta}})$ is the log-likelihood at the maximum likelihood estimates, n is the number of gene families, and m are the number of parameters estimated for the model. As above, higher BIC values are better.

TABLE 4.3: Model Support ordered by BIC (note that higher BIC is better, see Methods). BIC weights (w_i(BIC)) calculated following Wagenmakers and Farrell (2004). The ω column is the number of estimated dN/dS values in the model.

model	ω	ℓlk	BIC	w_i(BIC)
Subgenome-species	7	-4831284	-9663055	0.92
Time-Subgenome-species	11	-4831260	-9663060	0.082
Time-species	6	-4831316	-9663106	$9.0e^{-12}$
Subgenome	5	-4831349	-9663157	$6.9e^{-23}$
Ploidy	3	-4831386	-9663205	$2.4e^{-33}$
Time	4	-4831384	-9663214	$3.9e^{-35}$

4.3 Results

4.3.1 Subgenome-specific relaxation of selective constraints

Model comparisons by BIC indicated that the 'subgenome-species' model fit best and was 11.2 times more likely (BIC weight of 0.92 probability of being the best model; following Wagenmakers and Farrell 2004) than the second best model of 'time-subgenome-species' model, which was the most complex model (BIC weight of 0.082; Table 4.3). This model indicated that the diploid lineage experienced the strongest purifying selection (i.e., the lowest $\omega = 0.1245$), and the all S-subgenome lineages experience the weakest compared to their corresponding L-subgenome lineages (Fig. 4.4). For the mixed lineages, along which the whole genome duplication occurred, the S-subgenome had the weakest purifying selection detected ($\omega = 0.21$). However, this was not true for the L-subgenome lineage, where the laevis-clade had weaker purifying selection than the mixed lineages (Fig. 4.4). As well, for the laevis-clade, the two subgenomes had similar levels of purifying selection, with overlapping 95% confidence intervals. The more complicated time-subgenome-species model had a BIC that was only 5 units less, but included four more parameters, indicating that the time variable did not explain a large portion of the variance in these data (Fig. C0.2).

4.3.2 whole genome duplication duplicates differ between subgenomes in coding sequence length

Both the L- and the S-subgenome homeologs were recovered from one or more species (mean = 1.24 species per alignment) for 1417 alignments. The vast majority of these genes (1132) had only one species with both copies present and the other allotetraploids with one copy. 896 alignments had both copies only in *X. laevis* data.

On average, the S homeologs are shorter than corresponding L homeologs (-40.62 bp, 95% CI = -68.02 – -13.77). For the MCMCGLMM (see Methods) fit, effective samples

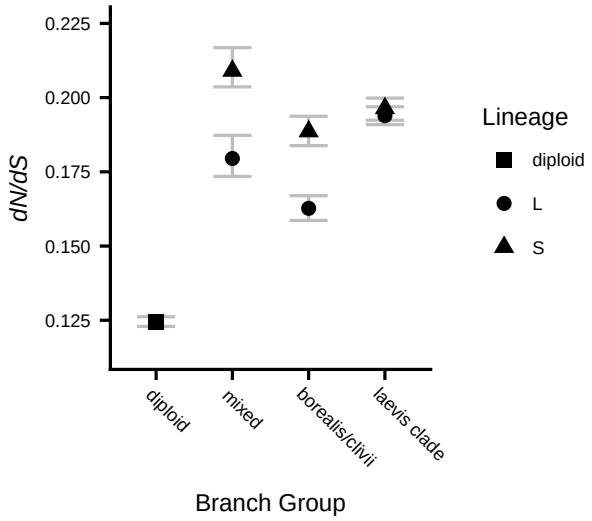


FIGURE 4.4: CODEML results from the favored 'subgenome-species' model. dN/dS estimates by CODEML 95% CIs from bootstrap replicates of the concatenated alignment.

sizes for parameter estimates were all above 1800 (and similar between parameters) and the amount of autocorrelation among samples was low (range: -0.007 – 0.04), indicating the chain had reached convergence.

4.3.3 The rates of pseudogenization differ between subgenomes of several allotetraploid *Xenopus* species

Model comparisons revealed that, similar to the CODEML analysis, the model estimating unique rates of pseudogenization for the two subgenomes was preferred over the homogeneous rates for each clade model, the model separating time periods, or the most complex model incorporating both time and subgenome (BIC: subgenome = -15,933, single rate = -15,963, time = -15,957, time-subgenome = -15,935). The results of the subgenome model indicate that the S-subgenome has a higher pseudogenization rate than the L-subgenome (borealis/clivii clade: $\hat{\theta}_S = 0.358$, $\hat{\theta}_L = 0.270$; laevis/largeni/allofraseri clade: $\hat{\theta}_S = 0.144$, $\hat{\theta}_L = 0.096$), with non-overlapping 95% confidence intervals (Fig. 4.5). The subgenome model also indicated that the borealis/clivii clade has a higher pseudogenization rate than the laevis/largeni/allofraseri clade (Fig. 4.5). As with the CODEML results, the more complex model incorporating time and subgenome had a similar fit as the subgenome model (BIC only 2 units less, compared to the other models that were over 20 BIC units less), indicating that time does not have a large effect on these estimated pseudogenization rates. But, the results of this more complex model indicated that the laevis/largeni/allofraseri clade may have an increasing rate of pseudogenization over time, with non-overlapping 95% confidence intervals within each subgenome

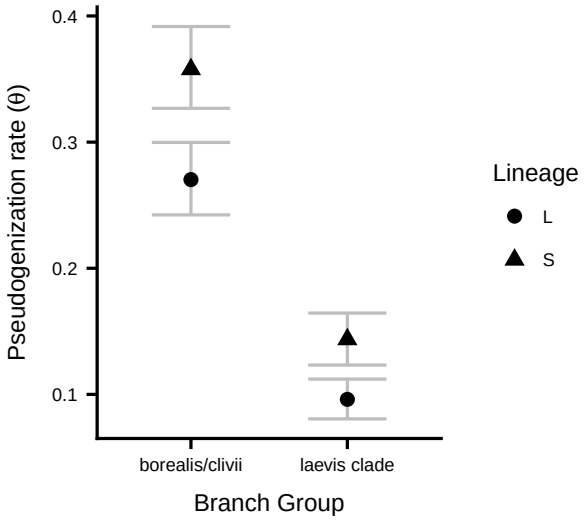


FIGURE 4.5: Estimated rates of pseudogenization from the BIC favored "subgenome model", with 95% confidence intervals based on 1000 bootstrap replicates. Missing data for non-*X. laevis* taxa varied from 0.34 – 0.55. See Table C0.1 for pseudogenization rate estimates of all models and estimated missing data proportions.

(Fig. C0.3). In the BIC-favored subgenome model, the estimated missing data proportions for each taxa ranged from 0.34 to 0.55 (all values with confidence intervals for all models are presented in Table C0.1).

Analyses on posterior distribution from *BEAST

As described above, the consensus tree used in these models (and simulations) was constructed from the posterior distribution of trees recovered from the original *BEAST analysis (Furman and Evans 2016). However, that analysis (as well as other analyses they performed and our maximum likelihood analysis above; Fig. 4.1) failed to confidently resolve the relationships among the *X. largeni*, *X. allofraseri*, and *X. laevis* clade (at least for the L-subgenome). To test the effect that alternate resolutions of these relationships have on our estimation of pseudogenization rates, we fitted the BIC preferred "subgenome model" to 1000 trees that were randomly sampled from the post-burn-in posterior distribution of the *BEAST analysis and transformed to chronograms using mcmctree, as outlined above. These trees had differences in branch lengths and tree topology (i.e., one of three possible resolutions of relationships among *X. largeni*, *X. allofraseri*, and *X. laevis*). For ease of comparison while taking into account the effect of the varying branch lengths on the estimated rates, we calculated the ratio of the estimated pseudogenization rates for the clade-specific S- to L-subgenomes. For the borealis/clivii clade, we found the median (interquartile range) ratio for S- to L-subgenome estimated

rates to be 1.323 (1.323, 1.323). Similarly, for the laevis/largeni/allofraseri clade estimated rates, we found the median (interquartile range) ratio to be 1.495 (1.488, 1.495). Thus, the S-subgenome had a higher pseudogenization rate than the L-subgenome. Similarly, we obtained ratios for the estimated pseudogenization rates for the borealis/clivii clade versus the laevis/largeni/allofraseri clade. Median (interquartile range) ratios for the S-subgenome and L-subgenome between the two clades are 2.487 (2.442, 2.490) and 2.812 (2.746, 2.814), respectively, i.e., the borealis/clivii clade had higher pseudogenization rates than the laevis/largeni/allofraseri clade. These ratios are nearly identical to those from the maximum likelihood estimates on the consensus tree, indicating that alternate resolutions of the *X. largeni*, *X. allofraseri*, and *X. laevis* clade did not affect the conclusions. This sort of sensitivity analysis shows that the phylogenetic comparative model was fairly robust to the provided phylogenetic tree.

4.4 Discussion

4.4.1 Genomic dynamics of relaxed purifying selection post-allopolyploidization

Allotetraploid species inherit each half of their genome (each subgenome) from different ancestral species. Our analyzes indicate that, after the *Xenopus* allotetraploid genome was generated, the strength of purifying selection on each subgenome differed significantly in the ancestors of some (but not all) of the species we studied. Specifically, in the lineages prior to speciation of the tetraploid ancestor and in the borealis/clivii clade, the S-subgenome experienced a greater relaxation of purifying selection than the L-subgenome. However, in the laevis/largeni/allofraseri clade, the strength of purifying selection was similar between the subgenomes. We also found that time since the speciation of the tetraploid ancestor of extant *Xenopus* did not have a large effect on the rate of purifying selection, and that gene coding regions are shorter in the S-subgenome than the homeologous coding regions in the L-subgenome. This latter result may be a consequence of more pronounced relaxation of purifying selection in this subgenome in the mixed lineage and in the borealis/clivii clade. Larger-scale subgenome-specific effects also have been observed: more rearrangements occurred since divergence of the diploid ancestors of *Xenopus* allotetraploids in the S subgenome than the L subgenome (Session et al. 2016).

That the level of purifiying selection did not vary substantially over time contrasts with the expectation that duplicates are most similar soon after allotetraploidization and thus maximally redundant, and that this redundancy would wane as homeologs diverge over time. Session et al. (2016) estimated divergence of the diploid ancestors of extant subgenus *Xenopus* species occurred about 34 million years (my) and that allotetraploidization between these diploid species then occurred roughly 15–17 my. Using our dataset and alternative methods and date calibrations, we recovered a similar estimate of 32 my (31.6–35.4 95% CI) for the divergence of the diploid ancestors (Appendix

C1). Thus, the relaxation of purifying selection has persisted and not returned to pre-duplication like levels for millions of years (Fig. 4.4; Session et al. (2016)).

These findings are consistent with those of previous studies on *Xenopus* duplicate genes (Chain and Evans 2006; Hellsten et al. 2007), and of WGD events in other taxa (Lynch and Conery 2000; Brunet et al. 2006; Scannell and Wolfe 2008), which find relaxed purifying selection post-WGD. Using a much smaller dataset, Chain et al. (2008) also explored duplicate gene evolution over time in *Xenopus*, and did not detect a difference in the level of purifying selection between the two categories in their comparison (equivalent to our "mixed" and a combined "mid" and "late" categories, similar to our subgenome model). In this analysis, we were able to assign duplicate copies to the L and S subgenomes and test for differences between the gene copies that were inherited from separate diploid ancestors, something Chain et al. (2008) were not able to do as the *X. laevis* genome sequence was not available.

4.4.2 Genomic dynamics of pseudogenization post-allopolyploidization

When an allotetraploid genome first forms, it is expected to have similar gene content in each subgenome because each one is derived from a different ancestral diploid species that carried the full complement of genes required for survival. Numerous genes can be simultaneously pseudogenized immediately after WGD, possibly due to large scale changes in the regulation of gene expression (Buggs et al. 2012; Lovell et al. 2014; Inoue et al. 2015). Here, we did not explore pseudogenization in the period of evolution that (i) immediately followed WGD but that (ii) preceded diversification of the extant allotetraploids. This is because the presence of at least one species with two homeologous sequences was required to establish orthology (Furman and Evans 2016). However, 60% of homeologous pairs are still both functional in *X. laevis* (Session et al. 2016), so presumably the most recent common allotetraploid common ancestor of the allotetraploids we studied retained at least this proportion (and probably an even higher proportion). Thus, our analysis of rates of pseudogenization focused specifically on gene pairs that (a) survived an initial period following WGD before speciation of allotetraploids, and also that (b) continue to be maintained as functional duplicates in at least one of the allotetraploid species, which describes the majority of duplicate genes in the *Xenopus* genome. Our results indicate that (i) each progenitor species contributed unequally to the functional gene content in extant allotetraploids, and that (ii) the conditions at the time of allopolyploidization (e.g., divergence between diploid ancestral species) and after allopolyploidization (e.g., species-specific population dynamics and mutation) strongly influence allopolyploid genome evolution.

After allotetraploidization in *Xenopus*, we found that the S-subgenome had a faster rate of pseudogenization than the L in several *Xenopus* allotetraploids (by about 30–50%; Fig. 4.5). The complete genome sequence of *X. laevis* reveals that the S-subgenome lost 31.5% of gene copies, where as the L had only lost 8% (Session et al. 2016). Our

findings thus extend these *X. laevis* genomic results to several other allotetraploid *Xenopus* species with 2n = 4x = 36 chromosomes, including *X. borealis*, *X. clivii*, *X. largeni*, and *X. allofraseri* (Fig. 4.5). Asymmetry in subgenome pseudogenization was observed in a smaller scale in two genes across a diversity of species in subgenus *Xenopus* (*RAG1*: Evans 2007; *DMRT1* loci: Bewick et al. 2011). Using BLAST, we assigned subgenome of origin for the results of Evans (2007) and Bewick et al. (2011), which indicated that the homeologs with the higher rates of pseudogenization are in the S-subgenome (data not shown), a finding that is consistent with the higher rate of pseudogenization in the S subgenome. This suggests the rate of pseudogenization is also higher in the S subgenomes of allo-octoploid and allo-dodecaploid *Xenopus* as well. Overall, in terms of gene copies recovered from the transcriptome data, single copy S genes were 15–30% less common than single copy L genes across the species, but this figure reflects a combination of pseudogenization and missing data (Table 4.1).

Our analyses did not detect evidence for a substantial change in the rate of pseudogenization over time since allotetraploidization in *Xenopus*. In teleosts, the rate of pseudogenization was highest soon after WGD (Inoue et al. 2015), but the period over which there was most rapid gene loss was about 60 my – greater than the time since the WGD event in *Xenopus*. Furthermore, a slowdown in teleost pseudogenization occurred only after about 80% of duplicates were lost (Inoue et al. 2015), whereas in *X. laevis* <40% of duplicates have been lost so far (Session et al. 2016). Yeast also exhibits a tempo of pseudogenization similar to teleosts, with more rapid gene loss earlier on (Scannell et al. 2006), but this pattern also played out over a longer period of time (and many more generations) than the *Xenopus* WGD analyzed here (> 60 my, Marcet-Houben and Gabaldón 2015). These results from *Xenopus*, which is a comparatively recent WGD event, indicate that in the early stages of genome restructuring post-WGD the rate of gene loss may be relatively constant until most gene copies are lost. Millions of years in the future, it is certainly possible that the rate of gene loss also will slow down in *Xenopus*.

We note that the borealis/clivii clade had a higher rate of pseudogenization in our analysis than the laevis/largeni/fraseri clade, which suggests that the rate of gene loss can be species-specific post-WGD. Similar to the species-specific relaxation of purifying selection discussed above, this could be a consequence of differences in life history, natural selection, demography, or other factors. Species-specific rates of pseudogenization after WGD also have been reported in yeast (Scannell et al. 2006). Purifying selection was stronger in the borealis/clivii clade for both subgenomes than in the laevis/largeni/allofraseri clade (Fig. 4.4), which could be because there are more singleton genes in the former clade. Indeed, in the expression data we analyzed here, the fewest number of genes in duplicate copy were recovered for the borealis/clivii clade (Table 4.1). The lack of substantial variation in the rate of pseudogenization over time coupled with pronounced variation among species in this rate, argues that aspects of the evolutionary fates of allopolyploid genomes are influenced to a great degree by species-specific phenomena (e.g., mutations, effective population size, demography). A high quality complete genome sequence for *X. borealis* and the other species will make possible further

exploration of these interpretations.

4.4.3 Asymmetric subgenome evolution

A unique implication of speciation by allopolyploidization is the merging of diverged genomes into a single species. While there may be beneficial consequences, such as higher dosage of beneficial alleles, there are also potentially negative consequences, such as poorly coordinated epistatic interactions from diverged members of a genetic network (Otto and Whitton 2000; Riddle and Birchler 2003; Comai 2005). Establishment of disomic inheritance (i.e., the formation of bivalents rather than multivalents at meiosis; Wolfe 2001) may confer greater genomic stability to allopolyploid genomes (Comai et al. 2003), and also allow for a preservation of subgenome differences that otherwise would be homogenized by recombination.

It has been well demonstrated in both old and young allopolyploids plants that the subgenome from one of the progenitors is often expressed less, and more frequently the target of pseudogenizing mutations, referred to as 'biased fractionation' (Flagel and Wendel 2010; Cheng et al. 2012; Garsmeur et al. 2013; Renny-Byfield et al. 2015). If each subgenome has a distinctive repertoire of transposible elements (TEs), a possible mechanism for differential subgenome evolution is RNA-mediated silencing of TEs that also represses adjacent genes (Woodhouse et al. 2014; Steige and Slotte 2016). Reduced gene expression is linked to weaker purifying selection and a higher rate of mutation and pseudogenization (Rocha 2006; Steige and Slotte 2016). Analysis of the extant *X. laevis* is consistent with this possibility, as the subgenomes have different TE classes and abundances, with the S-subgenome carrying subgenome-specific TE classes at high abundance (Session et al. 2016). Exploration of gene expression in this species found that L-subgenome homeologs tended to have higher expression than S-subgenome copies (mean difference of 10–25%, but a more modest median difference of 1.8% or less), along with 760 homeologous gene pairs where the homeolog with little to no expression had more relaxed purifying selection than the other (Session et al. 2016). These differences in TEs between the subgenomes were probably inherited from the diploid ancestors (Session et al. 2016), and this could have set the stage for higher pseudogenization in the S subgenome. Related to this, when the expression level of one homeolog is or evolves to be sufficient for survival, the fitness cost if the other becomes a pseudogene becomes tolerable (Freeling et al. 2012; Gout and Lynch 2015). The higher L-subgenome expression in *X. laevis* (Session et al. 2016) is thus consistent with these homeologs being less amenable to loss (Fig. 4.5). But, measuring only an extant species makes it unclear whether the differences in extant species subgenomes existed before WGD, or if the differences accumulated over time after WGD.

Though the (presumably) extinct diploid progenitors of tetraploid *Xenopus* cannot be directly assayed for differences in expression or natural selection, leveraging of multiple species and the branch specific dN/dS models provide some insight into differences at the time of WGD. We detected weaker purifying selection along the 'mixed lineages'

branches compared to the purely diploid lineage, and also significantly weaker purifying selection on the S mixed lineage than the L mixed lineage (Fig. 4.1, 4.4). The dN/dS estimates of the mixed lineages probably underestimate dN/dS immediately after WGD because a portion of the mixed lineages was diploid. Overall, however, these dN/dS estimates indicate that either S and L subgenome differences were rapidly established after WGD, or more likely were the result of divergence between the diploid progenitors, as suggested by Session et al. (2016) in the analysis of just *X. laevis*. If expression intensity is negatively correlated with purifying selection (Drummond et al. 2005; Rocha 2006), the expression differences between the L and S subgenomes may have been present soon after WGD in *Xenopus*, and persisted for many millions of years thereafter, as seen for other allopolyploids (Cheng et al. 2012; Renny-Byfield et al. 2015). Our analysis supports that S-subgenome loss has been higher than the L-subgenome, and remained consistently so over time (Fig. 4.5), supporting that the differences between the subgenome have been a persistent feature of *Xenopus* allotetraploids.

Asymmetry in subgenome evolution has several interesting consequences for genome evolution. For instance, allopolyploid cotton species, also show asymmetry in subgenome evolution with a bias in gene conversion rates (one subgenome is more frequently converted by the other subgenome), and there exists subgenome biases in gene involvement in certain phenotypes (Paterson et al. 2012). Wheat species subgenomes carry different levels of genetic diversity, with the more diverse subgenome involved in local adaptation phenotypes and the other preserving more core function genes (Feldman et al. 2012). In *Xenopus*, cyto-nuclear incompatibilities appear to be limited to one subgenome or the other (Gibeaux et al. 2018), indicating that subgenome specific evolution may additionally contribute to the origin of reproductive incompatibilities among species. Interestingly, biases in subgenome evolution may exist even though genetic exchange between subgenomes does occasionally occur. Genetic exchange between subgenomes of allotetraploid *Xenopus* is illustrated, for example, by the sex determining gene *DMW* which resides on the L-subgenome but was formed from a partial gene duplicate of the S-subgenome copy of a gene called *DMRT1* (Bewick et al. 2011).

Overall, this study paints a dynamic portrait of allopolyploid genome evolution by highlighting among several closely related allotetraploids evidence for persistent relaxed purifying selection with species-specific subgenome patterns, and ongoing pseudogenization with asymmetric rates in each subgenome. Many functional duplicate genes still remain in these and other allotetraploids, and do so for millions of years (Renny-Byfield et al. 2015, this study). As such, events that are thought to depend on rare mutational events that promote the retention of duplicate genes, such as neofunctionalization, may in fact have a protracted time-frame within which to occur.

4.5 Competing interests

The authors declare that they have no competing interests.

4.6 Author's contributions

Data and modeling were part of previous efforts by BLSF/BJE, and UJD/GBG, respectively. Here, analysis was done by BLSF and UJD, with input from BJE and GBG. All authors contributed to writing.

4.7 Acknowledgements

This work was supported by the Natural Science and Engineering Research Council of Canada (CGSD3-475567-2015 to BLSF, RGPIN/283102-2012 and RGPIN-2017-05770 to BJE, and RGPIN-2015-04477 to GBG).

4.8 Data Availability

Transcriptome sequence data are available in the NCBI short read archive, as submitted by Furman and Evans (2016) (accessions PRJNA318484, PRJNA318394, PRJNA318474, PRJNA318404). Code, data, and the phylogenetic tree for pseudogenization model analyses are available as supplemental files with the online publication.

4.9 Additional Files

Additional file 1 — Supplemental Text and Figures Additional figures, explanation of divergence estimates, and details of simulations of pseudogenization rates can be found in the supplemental file, and as Appendix C.

Part III

Speciation and Hybridization

Chapter 5

Pan-African phylogeography of a model organism, the African clawed frog *Xenopus laevis*

Benjamin L. S. Furman*, Adam J. Bewick[*1], Tia L. Harrison[**2], Eli Greenbaum[†], Václav Gvoždík[‡§], Chifundera Kusamba[¶], Ben Evans*

*Biology Department, McMaster University, Hamilton, Ontario, L8S 4K1, Canada

[†]Department of Biological Sciences, University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968

[‡]Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Kvetna 8, 603 65 Brno, Czech Republic

[§]Department of Zoology, National Museum, 193 00 Prague, Czech Republic

[¶]Laboratoire d'Herpétologie, Département de Biologie, Centre de Recherche en Sciences Naturelles, Lwiro, République Démocratique du Congo

[1]Current address: University of Georgia, Department of Plant Biology, Miller Plant Sciences Bldg., Athens, GA 30602

[2]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario M5S3B2, Canada

This paper was published in *Molecular Ecology* and is available here in its published form.

Abstract The African clawed frog *Xenopus Xenopus laevis* has a large native distribution over much of sub-Saharan Africa and is a model organism for research, a proposed disease vector, and an invasive species. Despite its prominent role in research and abundance in nature, surprisingly little is known about the phylogeography and evolutionary history of this group. Here, we report an analysis of molecular variation of this clade based on 17 loci (one mitochondrial, 16 nuclear) in up to 159 individuals sampled throughout its native distribution. Phylogenetic relationships among mitochondrial DNA (mtDNA) haplotypes were incongruent with those among alleles of the putatively female-specific sex-determining gene *DM-W*, in contrast to the expectation of strict matrilineal inheritance of both loci. Population structure and evolutionarily diverged lineages were evidenced by analyses of molecular variation in these data. These results further contextualize the chronology, and evolutionary relationships within this group, support the recognition of *X. laevis sensu stricto*, *X. petersii*, *X. victorianus* and herein revalidated *X. poweri* as separate species. We also propose that portions of the currently recognized distributions of *X. laevis* (north of the Congo Basin) and *X. petersii* (south of the Congo Basin) be reassigned to *X. poweri*.

5.1 Introduction

The African clawed frog *X. laevis sensu lato* (Kobel et al. 1996) has an unusual connection with humans, having been used in the early 20th century as a pregnancy assay (Shapiro and Zwarenstein 1934; Weisman and Coates 1941) and more recently as a model organism for research (Cannatella and Sá 1993; Gurdon 1996; Gurdon and Hopwood 2003). Also called the Common Platanna, the native range of this species spans much of sub-Saharan Africa (Tinsley et al. 1996). Established invasive colonies exist in portions of Europe, North America and South America (McCoid and Fritts 1980, 1989; Tinsley and McCoid 1996; Measey and Tinsley 1998; Lobos and Jaksic 2005), and nonpersistent populations have been reported in other localities, including parts of Asia (Measey et al. 2012). *X. laevis* has been identified as a potential vector for the amphibian pathogens *Batrachochytrium dendrobatidis* (Weldon et al. 2004) and ranavirus (Robert et al. 2007), although a causal role between *X. laevis* and the dispersal of these pathogens has not been demonstrated (Measey et al. 2012). *X. laevis* has potentially harmful consequences for *X. gilli* (Evans et al. 2004; Evans et al. 2005; Evans 2007), which is endangered (South African Frog Reassessment Group (SA-FRoG) ISASG (2013)), through ecological competition and hybridization (Tinsley 1981; Simmonds 1985; Picker et al. 1993, 1996; Evans et al. 1997, 1998; Fogell et al. 2013).

X. laevis is generally found in slow moving or stagnant water, and occasionally disperses over land (Measey and Tinsley 1998; Eggert and Fouquet 2006). It is a generalist species that does well in disturbed habitat and has a high capacity to tolerate drought conditions, salinity, starvation, anoxia and temperature fluctuations (reviewed in Measey et al. 2012). This species (and other frogs in the Family Pipidae) has unusual adaptations in adults for a mostly aquatic lifestyle, including lateral line sensory organs and a complex communication system involving a unique mechanism of sound production, context- and sex-specific vocalizations and female phonotaxis (Tobias et al. 1998; Kelley and Tobias 1999; Tobias et al. 2004; Tobias et al. 2011).

5.1.1 Tetraploidization, sex determination

An ancestor of *X. laevis* experienced genome duplication during its evolution, probably by allopolyploidization between two diploid ancestors with 18 chromosomes, to create a genome with 36 chromosomes (Tymowska 1991). However, its genome is now functionally diploidized in that, during cell division, each chromosome aligns with only one other homologous chromosome (Tymowska 1991). Tetraploidization in the ancestor of *X. laevis* duplicated essentially all genes in its genome, although many of these duplicates were reduced to a single copy as a result of pseudogenization (Chain and Evans 2006; Morin et al. 2006; Hellsten et al. 2007; Sémon and Wolfe 2008; Chain et al. 2011).

Of particular interest is the chromosome W-linked DM-domain containing gene (*DM-W*), which is present as a single allele in female *X. laevis*, but absent in males, and is the sex-determining gene in this species (Yoshimoto et al. 2008). *DM-W* originated

by partial segmental duplication of one of the two copies (paralogs) of the double sex- and mab-3-related transcription factor 1 gene (*DMRT-1*) that arose when an ancestor of *X. laevis* experienced tetraploidization (Bewick et al. 2011). If *DM-W* was strictly maternally inherited over evolutionary time, its evolutionary history is expected to match that of mtDNA, which is also thought to be maternally inherited and nonrecombining in most species.

5.1.2 Taxonomy and phylogeography of *X. laevis sensu lato*

X. laevis sensu lato (Kobel et al. 1996) comprises three currently recognized species (AmphibiaWeb 2014; Frost 2011): *X. laevis sensu stricto* from southern Africa and a disjunct population north of the Congo Basin (Daudin 1802), *X. petersii* from southern Central Africa (Bocage 1895) and *X. victorianus* from Eastern Africa (Ahl 1924). A fourth species, *X. poweri*, was described based on specimens from the area of the Victoria Falls (Zambia–Zimbabwe border) by Hewitt (1927) but considered a subspecies of *X. laevis* by some authors (Schmidt and Inger 1949; Poynton 1964), or a synonym with *X. laevis petersii* (e.g. Parker 1936b; Poynton and Broadley 1985). *X. laevis sensu lato* also includes two proposed subspecies: *X. l. bunyoniensis* (Loveridge 1932) and *X. l. sudanensis* (Perret, Jean-Luc 1966). Additional information on the taxonomic history of this clade is provided in Appendix D.

Diversity within *X. laevis sensu lato* has been explored in terms of molecular and morphological variation (Carr et al. 1987; Grohovaz et al. 1996; Evans et al. 1997; Kobel et al. 1998; Measey and Channing 2003; Evans et al. 2004; Du Preez et al. 2009) and variation in vocalization (Tobias et al. 2011). In general, these studies consistently found that populations in different parts of Africa, including populations from different portions of South Africa, are differentiated. The distribution of variation within mtDNA is perhaps best relayed in terms of four geographical zones of sub-Saharan Africa, which we will refer to as 'southern Africa' (including South Africa and Malawi), East Africa (including Tanzania, Kenya, Uganda, Burundi, Rwanda and the eastern portion of the Democratic Republic of the Congo), 'Central Africa' (including Nigeria, Cameroon, western Zambia and northern Botswana), and 'West Central Africa' (including the southern Republic of Congo, the western portion of the Democratic Republic of the Congo and Angola) (Fig. 5.1). Evans et al. (2004) analysed mtDNA sequences from *X. laevis sensu lato* from each of these zones. Their analysis recovered paraphyly of the group of mtDNA sequences from southern Africa with relatively weak support, but recovered strong support for monophyly of the group of mtDNA sequences from East and Central Africa (Evans et al. 2004). mtDNA from one sample from the Republic of Congo (West Central Africa) was closely related to a clade containing mtDNA from Central and East Africa (Evans et al. 2004). Within the country of South Africa, Grohovaz et al. (1996) and Measey and Channing (2003) found a population of *X. laevis sensu lato* sampled near the town of Niewoudtville to be distinct from populations in other parts of the country. Measey and Channing (2003) also identified a zone of admixture of mitochondrial haplotypes from Niewoudtville and haplotypes from the south-western Cape Region in the vicinity of the

town of Vredendal (not sampled in the current study), which is 100 km south-west of Niewoudtville. Du Preez et al. (2009) further identified a second zone of admixture near the town of Laingsburg (sampled in the current study), South Africa, between South African populations to the north-east and south-west of this locality based on variation in mtDNA and two autosomal genes.

The main goal of this study is to further characterize the evolutionary history of *X. laevis sensu lato* in terms of the phylogenetic relationships, divergence times and geographic distributions of diverged evolutionary lineages. We additionally evaluate support for previously proposed species designations within *X. laevis sensu lato* (*X. laevis sensu stricto*, *X. victorianus*, *X. petersii* and *X. poweri*). For evaluating support for the previously proposed species, we adopt the 'General Lineage Concept' (GLC De Queiroz 1998, 2007) of a species, which defines a species as a 'separately evolving metapopulation lineage' (De Queiroz 2007). The term 'metapopulation' refers to a set of subpopulations that are interconnected by gene flow, and 'lineage' refers to the ancestor–descendant relationship between metapopulations of the same species through time (De Queiroz 2007).

5.2 Methods

5.2.1 Samples and molecular data

A total of 183 samples of *X. laevis sensu lato* from 14 countries were used in this study, including 104 samples obtained from South Africa, 37 from Democratic Republic of the Congo (hereafter DRC), 12 from Burundi, 8 from Zambia, 7 from Cameroon, 3 from Nigeria, 3 from Uganda, 2 from Kenya, 2 from Botswana and 1 each from Rwanda, the Republic of Congo, Angola, Malawi and Tanzania (see Table D2.1, Supporting information for specific locality information). These tissue samples were obtained from field collections, tissue donations from institutional archives (California Academy of Sciences, the Museum of Comparative Zoology at Harvard University, the Natural History Museum of Geneva and the Zoological Research Museum – Alexander Koenig), a collection of live Xenopus that was at the University of Geneva, and colleagues (T. Hayes, L. Kalous, R. Tinsley and P. Wagner).

Sequences from a portion of the mitochondrial 12S and 16S rDNA genes and the intervening *tRNAVal* gene were obtained using primers from Evans et al. (2004) for 159 *X. laevis sensu lato* individuals (87% of the samples in this study), with an average of 907 base pairs (bp) per individual (range: 623–2374 bp). Sequences from the female-specific W chromosome gene *DM-W* and flanking regions were obtained using primers detailed in Bewick et al. (2011) for 96 female *X. laevis sensu lato* individuals, with an average of 1734 bp per individual (range: 1036–2049 bp). Autosomal DNA sequences were obtained from portions of the protein coding region of 15 loci ranging in length from 341 to 618 bp (Table 5.1) for 113–136 individuals per locus, using paralog-specific primers

detailed in Bewick et al. (2011). Sequence data were aligned by eye, and homologies of the aligned characters were unambiguous. Sequences of individual autosomal alleles were inferred using the 'best guess' estimates of allelic states from PHASE v.2.1.1 using default parameters (Stephens et al. 2001; Stephens and Donnelly 2003), and both alleles were analysed for the population assignment tests detailed below. DNASP v.5.10.01 (Librado and Rozas 2009) was used to quantify descriptive statistics of the sequence data, and formula 5 of (Kimura and Ohta 1972) to calculate 95% confidence intervals (95% CI) for pairwise nucleotide diversity at synonymous sites. All new sequence data are deposited in GenBank (Accession nos. KP343951–KP345838), and Accession nos. of other data in these analyses are listed in previous studies (Evans et al. 2004; Evans et al. 2005; Evans 2007; Evans et al. 2008; Bewick et al. 2011; Evans et al. 2011a).

5.2.2 Phylogenetic analyses

We used BEAST v.1.6 (Drummond and Rambaut 2007) to generate time calibrated trees for the mitochondrial and for the *DM-W* data. For each locus, we performed four independent runs, 50 million generations each, using a strict clock. Previously published orthologous sequences from *X. gilli* were used as outgroups. The timing of divergence of *X. laevis* and *X. gilli* was set to 16.7 million years (my) with a standard deviation of 3.62 my (Evans et al. 2004), to calibrate these analyses. This divergence time is based on the assumption that the separation of the South Atlantic Ocean triggered the diversification of South American from African pipid frogs ~100 my (Pitman III et al. 1993; Maisey 2000; McLoughlin 2001; Sereno et al. 2004; Ali and Aitchison 2008) and was based on analysis of data from mtDNA (Evans et al. 2004). We tested for convergence of the MCMC chains on the posterior distribution by calculating effective sample sizes (ESSs) of post-burn-in likelihoods using TRACER v.1.5 (Rambaut et al. 2014), and inspecting traces of parameter estimates. This led us to discard a burn-in of 25% of the generations from each analysis. For each analysis, the model of evolution was selected by the program MRMODELTEST version 2 (Nylander 2004) based on the Akaike information criterion (AIC). The preferred model for the mtDNA analysis was the general time reversible model (Tavaré 1986), with a proportion of invariant sites, a gamma-distributed heterogeneity in the rate of evolution and estimated base frequencies (GTR+I+Γ+bf). For the *DM-W* data set, the preferred model was the Hasegawa, Kisino and Yano model (Hasegawa et al. 1985), with a proportion of invariant sites and estimated base frequencies (HKY+I+bf).

To examine evolutionary relationships among the autosomal genes, three approaches were taken. First, phylogenetic networks were generated among phased autosomal alleles from each locus using SPLITSTREE v.4.13.1 (Huson and Bryant 2006). We used Jukes-Cantor corrected distances between alleles and the Neighbor-Net algorithm (Bryant and Moulton 2004). Support for the splits in the networks was determined with a bootstrap analysis with 1000 replicates. Second, we performed a phylogenetic analysis on concatenated autosomal data using BEAST version 1.7.4 (Drummond et al. 2012) including individuals with less than 50% missing data (i.e. the same individuals that

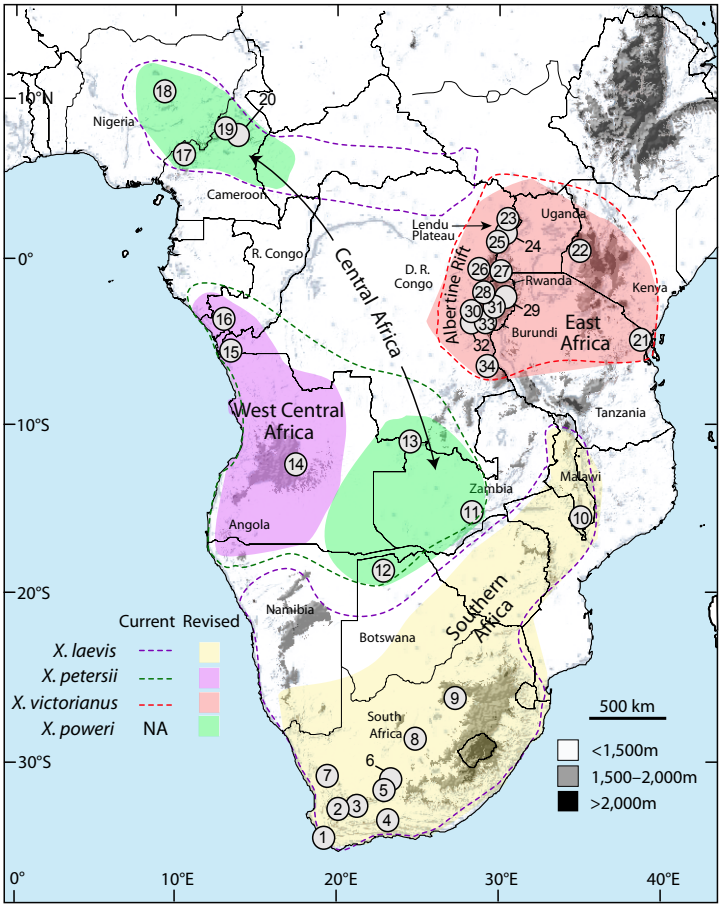


FIGURE 5.1: *X. laevis sensu lato* sampling localities, and currently recognized and revised species ranges. Numbers inside circles indicate locality numbers that correspond with samples listed in Table D2.1. Unfilled polygons with different lines indicate the currently recognized distributions of *X. laevis*, *X. petersii* (including *X. poweri* as a synonym) and *X. victorianus*, respectively (Frost 2011; *Information on Amphibian Biology and Conservation [Web Application]* 2014). Filled polygons indicate four geographical regions that are referred to in the text, each of which corresponds to the distribution of a species (named below the geographical region) that is supported by this study. No locus in this study has data from every sample depicted; mtDNA has the least missing data (see Table D2.1 for details). Additional shading refers to meters above sea level as indicated.

were included in the population assignment analyses described below). And third, we estimated a species tree using *Beast with a reduced data set of 70 individuals and 10 genes that minimized missing data across individuals and loci. For the phylogenetic analyses of concatenated autosomal data, we used a model of evolution selected by MrModeltest, with the same calibration procedure as detailed above for mtDNA and *DM-W*, and we performed four independent runs for 20 million generations each. For the *Beast analysis, we assumed a strict molecular clock with an exponentially distributed mutation rate with a mean of 4.7×10^{-10} substitutions/site/generation following Bewick et al. (2012). This mutation rate estimate is based on a multilocus analysis of data from $>$ 100 genes from pipid frogs and relied on the same assumption about the geological trigger for diversification of pipid frogs as the mtDNA analysis above. To achieve convergence, it was necessary to use a simpler model of evolution than that recommended by MrModeltest for the concatenated data set (we used HKY+Γ+bf instead of GTR+I+Γ+bf). For the *Beast analysis, we linked the model of evolution across all data partitions, and unlinked the phylogeny of each partition. *A priori* species designations were based on eight clades that were observed in the concatenated analysis of autosomal DNA, including: (i) Nigeria and Cameroon, (ii) Botswana and Zambia, (iii) Angola and western DRC, (iv) eastern DRC, Uganda and Burundi, (v) Malawi, and the South African localities Kimberly and Victoria West, (vi) the South African locality Niewoudtville, (vii) the South African localities Betty's Bay, Garden Route National Park and some individuals from Laingsburg and (viii) the South African localities Beaufort West, and other individuals from Laingsburg. The *Beast analysis was performed with four independent runs, each for 100 million generations. For Beast and *Beast analyses, orthologs from *X. gilli* were used as the outgroup, and independent runs were combined using LogCombiner version 1.7.4 (Drummond et al. 2012). Similar to the other phylogenetic analyses, 25% of the run was discarded as burn-in, and convergence was assessed based on ESSs of the parameters as calculated by Tracer version 1.6 (Drummond and Rambaut 2007). A maximum clade credibility tree with median node heights was constructed using TreeAnnotator version 1.7.4 (Drummond et al. 2012).

5.2.3 BP&P analysis

We used BP&P version 2.2 (Yang and Rannala 2010) to test for evidence of species limits. This analysis uses molecular data and a 'guide' phylogeny, which is a hypothesized relationship among populations or species, to evaluate the posterior probability of a species tree. The species tree is assumed to either be the same as the guide tree, or alternatively to be a simplified version of the guide tree that can be obtained by collapsing one or more nodes. We used a guide tree based on clusters obtained from the phylogenetic analysis of concatenated autosomal DNA, and included only the autosomal DNA sequences that were analysed in the *Beast analyses detailed above. BP&P has two different reversible jump proposal algorithms for species delimitation that influence the probability that nodes within the guide tree are expanded or collapsed during the

Markov Chain (Yang and Rannala 2010). We ran two independent chains for each algorithm using a gamma prior G (2, 1000) for both the population size and tree root age priors, with automatic adjustments of step lengths in the MCMC algorithm made by the program. In addition, we explored an alternative prior for both of these parameters in which we calculated a scale parameter (b) for the gamma distribution, by dividing 1 (a diffuse value for the shape parameter (a) of the gamma distribution) by the mutation rate used in the *Beast analysis (Bewick et al. 2012) multiplied by an estimated divergence time from the outgroup taxon *X. gilli* of 16.7 my (Evans et al. 2004), which resulted in a value of 126. We then ran two independent chains for both algorithms with this new gamma prior distribution G (1, 126) for both the ancestral population size and the tree root age. For each prior setting, the MCMC was run for 100 000 generations, and 20 000 generations were discarded as burn-in, based on visual inspection of the posterior distribution of likelihoods.

5.2.4 Population assignment

The phylogenetic analyses detailed above evaluate evolutionary relationships in the context of a bifurcating phylogeny. However, autosomal DNA relationships may reticulate or be inconsistent among loci as a result of gene flow, lineage sorting and recombination, and this is a particular concern when analysing intraspecific samples. More specifically, use of a phylogeny estimated from autosomal DNA to guide the *Beast and BP&P analyses comes with the caveat that we did not explore all possible groupings or (for BP&P) all possible relationships among these groups, and therefore the results are contingent on the *a priori* groups and guide tree that we used for *Beast and BP&P, respectively. Population assignment tests (and also the Splits-Tree analysis discussed above) therefore offer a complementary perspective on the nature of multilocus molecular variation among taxa because they do not interpret evolutionary relationships in the context of a bifurcating tree. To assess the degree of population structure and assign individual genotypes to putative populations, we used the programs Tess v.2.3 (Chen et al. 2007) and Structure v.2.3 (Pritchard et al. 2000). Both approaches estimate the probability that each individual is assigned to K populations, with an aim of minimizing Hardy-Weinberg and linkage disequilibria within the populations (Pritchard et al. 2000; Chen et al. 2007; François and Durand 2010). Unlike Structure, Tess incorporates spatial information on geographic distances between sampling points (based on GPS coordinates) into the prior distribution when calculating individual assignment probabilities (Chen et al. 2007; François and Durand 2010).

For Tess and Structure analyses, we excluded data from individuals with missing data from more than half of the loci. Data from mtDNA and *DM-W* were also excluded so that these analyses would provide a perspective on diversification independent from the analyses of the maternally inherited loci. The 135 individuals in the analysis had an average of 6.3% of the loci with missing data. We ran Tess for 100,000 generations with a burn-in of 10,000, using the conditional auto-regressive admixture model (Durand et al. 2009), starting from a neighbour-joining tree and using 10 iterations for each value of K

ranging from 2 to 10. Because some individuals were sampled from the same location, we used the 'generate spatial coordinates for individuals' option in TESS, with a standard deviation equal to 1.0. Convergence was based on inspection of post-run log-likelihood plots, and support for alternative K values was assessed by inspection of the deviance information criterion (DIC) (Spiegelhalter et al. 2002); models with lower DIC values are preferred.

For STRUCTURE analyses, we ran 20 million generations with a burn-in of 2.5 million generations for values of K equal to 2–10, with five iterations for each value of K. We specified the 'admixture model' (Falush et al. 2003) and assumed no correlation between alleles. The post-run likelihood values were stable and support for alternative K values was evaluated using the DK statistic (Evanno et al. 2005), as calculated with STRUCTURE HARVESTER WEB v.0.6.93 (Earl et al. 2012), and the ad hoc method outlined in Pritchard et al. (2000).

The samples used in the population assignment analyses comprise more from South Africa (n = 107) than from other portions of the distribution of *X. laevis sensu lato* that are not from South Africa (n = 41). To examine whether this uneven geographic sampling affected our results, we reran the TESS analysis with a random subsample of only five individuals from each sampling locality in South Africa, plus all other samples from other countries. This reduced data set had a total of 66 individuals, of which 25 were from South Africa. We ran this analysis for 100,000 generations with 10,000 discarded as burn-in, for K values ranging from 2 to 10, with 10 iterations for each value. Other analytical details were identical to those discussed above.

TESS and STRUCTURE runs were post-processed using CLUMPP v.1.1.2 (Jakobsson and Rosenberg 2007), which averages assignment probabilities across iterations. CLUMPP offers three separate algorithms that maximize similarity across all of the iterations of a given K; we selected an algorithm as recommended in the program documentation (Jakobsson and Rosenberg 2007).

5.3 Results

5.3.1 Phylogenetic incongruence between maternally inherited loci

Estimated phylogenetic relationships from mtDNA and from *DM-W* each resolve sequences into geographically clustered clades that correspond with one another, and both analyses recover strong and congruent support for paraphyly of the group of haplotypes from individuals in one pond in Laingsburg, South Africa. However, there are strongly supported inconsistencies in the estimated relationships among these clades (Fig. 5.2; see Fig. 5.1 and insert in Fig. 5.3 for sampling locations). The mtDNA phylogeny supports monophyly of the group of sequences from the following South African localities: Niewoudtville, Beaufort West, Laingsburg, De Doorns, Betty's Bay, GRNP, Hoekwil and Cape Town, whereas the *DM-W* phylogeny supports paraphyly of this group of

sequences (Fig. 5.2). Another difference with strong statistical support is seen in relationships among samples from East Africa. In the mtDNA phylogeny, all East Africa sequences that are not from or near the Lendu Plateau form a strongly supported clade. But in the *DM-W* phylogeny, this group of sequences is inferred to be paraphyletic. Strongly supported inconsistent relationships were also inferred when we restricted both analyses to include only those individuals for whom data were collected from both loci (data not shown).

5.3.2 Molecular variation, evolutionary relationships and species delimitation using autosomal loci

Table 5.1 presents polymorphism statistics for four diverged lineages of *X. laevis sensu lato* that correspond to previously proposed species within this group as redefined below. All of the loci were polymorphic within *X. laevis sensu lato*. One locus exhibited a Tajima (1989) D value that was significantly greater than zero within a geographical region depicted in Fig. 5.1, an observation that could reflect a signature of balancing selection. After weighting individual locus values by the number of synonymous sites at each locus, the largest average pairwise diversity of synonymous sites was similar for individuals from southern Africa (0.0148; 95% CI: 0.0091–0.0206), Central Africa (0.0159; 0.0100– 0.0218) and West Central Africa (0.0124; 0.0072–0.0176), but about half as large for individuals from East Africa (0.0081; 0.0039–0.0124).

Phylogenetic analysis of concatenated autosomal data from 135 *X. laevis sensu lato* individuals provided strong support for multiple diverged evolutionary lineages (Fig. 5.3), many of which correspond to those identified in the analyses of mtDNA and *DM-W*. Diverged lineages in the analysis of concatenated autosomal data include (i) individuals from southern Africa (South Africa and Malawi), (ii) individuals from East Africa (Uganda, Burundi, eastern DRC), (iii) individuals from Central Africa (Nigeria, Cameroon, Zambia, Botswana) and (iv) individuals from West Central Africa (Angola and western DRC). The topology of relationships among the geographically clustered clades was more similar to that inferred from *DM-W* than mtDNA in the sense that the group of samples from Malawi and South Africa are inferred to be monophyletic. However, it was more similar to the mtDNA phylogeny than the *DM-W* phylogeny in that the group of samples from East Africa that were not from or near the Lendu Plateau were inferred to be monophyletic. The analysis of concatenated autosomal data differs from the analyses of mtDNA and *DM-W* in that the former supports monophyly of the group of samples from East Africa (Fig. 5.2,5.3). Both maternally inherited loci supported a close relationship between haplotypes from Niewoudtville, South Africa and those from Beaufort West, South Africa and a few from Laingsburg, South Africa, but this relationship was not observed in the analysis of concatenated autosomal loci.

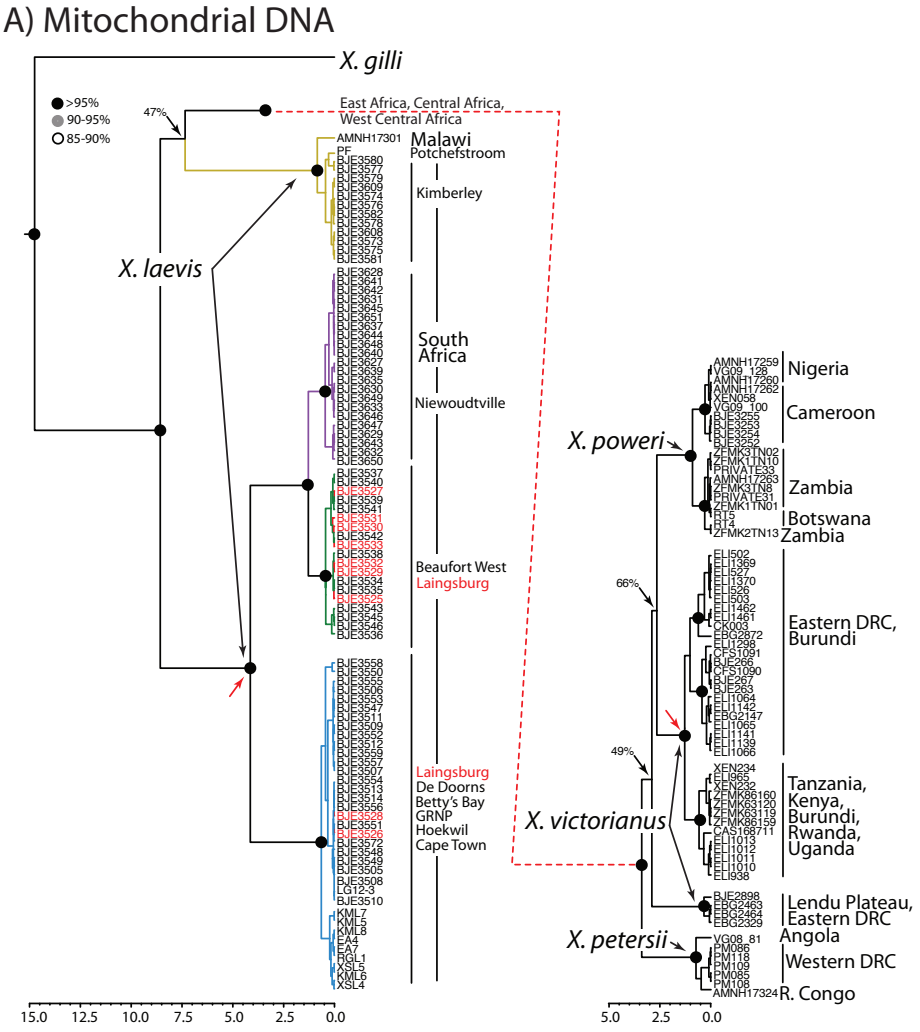


FIGURE 5.2: Chronogram among (A) mtDNA and (B) *DM-W* haplotypes in *X. laevis sensu lato*. Shaded dots over nodes indicate posterior probabilities, expressed as percentages as indicated; some terminal support values were omitted for clarity, and the posterior probabilities of various poorly supported nodes are indicated. The scale bar indicates divergence time from the present in millions of years. With the exception of the sample from Malawi, shaded branches in southern Africa correspond to sampling localities depicted in Fig. 5.3. Small arrows indicate relationships that are well supported in each phylogeny but discordant between them.

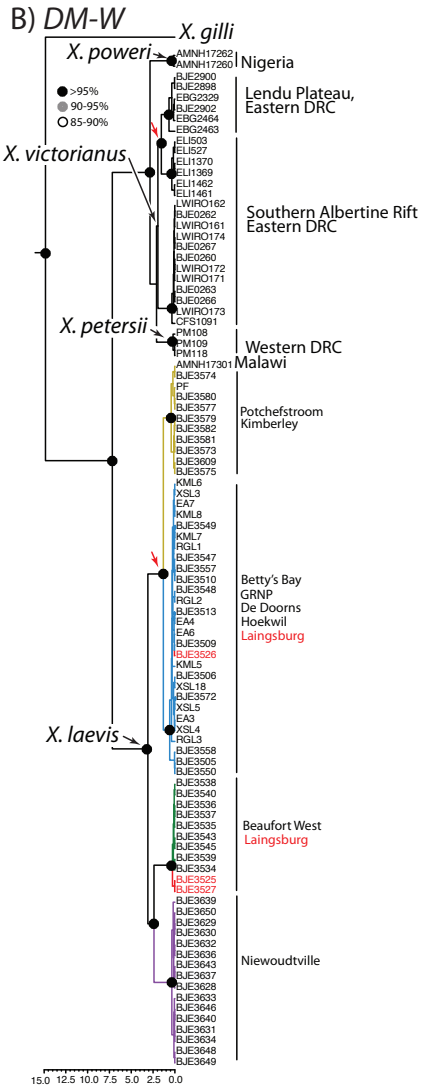


FIGURE 5.2: continued.

TABLE 5.1: Polymorphism statistics for autosomal loci for *X. laevis*, *X. petersii*, *X. poweri* and *X. victorianus*, including the gene acronym (gene), number of base pairs sequenced (bp), number of alleles sequenced (No. of allelles), number of unique haplotypes (No. of haplotypes), number of synonymous sites (SSites), the number of nonsynonymous sites (NSites), Jukes Cantor corrected pairwise nucleotide diversity for synonymous (pS) and nonsynonmous (pN) sites, the number of segregating synonymous (SS) and nonsynonymous (SN) sites, and Tajima's D based on synonymous sites (DS), with * indicating significant departure from zero. For some loci, NA indicates that Tajima's D could not be calculated due to insufficient molecular diversity or data

Gene	bp	No. of alleles	No. of haplotypes	SSites	NSites	pS	pN	SS	SN	DS
Southern Africa (*X. laevis*)										
AR	339	160	9	83	256	0.0015	0.0035	3	6	-1.28
prmt6	612	170	61	146	466	0.0237	0.0033	17	10	0.36
mogA	619	176	26	155	463	0.0031	0.0037	3	17	-0.13
c7orf25	531	184	16	119	412	0.0127	0.0005	11	4	-0.53
nfil3	534	188	21	120	415	0.0172	0.0016	12	7	-0.06
pigo	494	184	28	127	365	0.0230	0.0029	15	12	0.25
Sugp2	438	182	19	108	330	0.0103	0.0027	11	9	-1.16
mastl	537	184	46	119	418	0.0126	0.0073	11	32	-0.53
zbed4	471	170	17	110	361	0.0100	0.0029	9	9	-0.72
Rassf10	486	186	36	98	388	0.0307	0.0070	19	11	-0.29
p7e4	522	186	30	121	401	0.0274	0.0009	16	5	0.45
fem1c	474	146	24	107	367	0.0192	0.0010	17	21	-0.94
znf238.2	531	186	24	117	410	0.0030	0.0073	4	17	0.17
bcl9	489	188	18	114	372	0.0123	0.0010	10	10	-0.45
nufip2	473	156	24	105	363	0.0148	0.0019	27	29	1.97*
West Central Africa (*X. petersii*)										
AR	339	8	6	83	256	0.0065	0.0058	1	4	1.17
prmt6	612	2	2	145	467	0.0210	0.0021	3	1	NA
mogA	619	10	7	155	463	0.0000	0.0139	0	19	NA
c7orf25	531	10	9	118	413	0.0549	0.0099	18	11	-0.10
nfil3	534	10	6	120	414	0.0211	0.0013	6	1	0.72
pigo	494	10	5	127	365	0.0167	0.0006	6	1	-0.06
Sugp2	438	10	4	108	330	0.0052	0.0026	2	3	-0.69
mastl	537	10	4	119	418	0.0056	0.0018	2	3	-0.18
zbed4	471	8	2	109	357	0.0023	0.0000	1	0	-1.05
Rassf10	486	8	4	98	388	0.0118	0.0018	2	2	1.80
p7e4	522	10	3	122	400	0.0089	0.0000	3	0	0.02
fem1c	474	8	3	107	367	0.0175	0.0000	4	0	0.79
znf238.2	531	10	4	117	414	0.0052	0.0027	3	3	-1.56
bcl9	489	10	3	114	372	0.0000	0.0022	0	2	NA
nufip2	473	10	4	106	365	0.0070	0.0018	2	2	0.12
Central Africa (*X. poweri*)										
AR	339	26	5	83	256	0.0071	0.0020	4	2	-1.20
prmt6	612	22	12	145	467	0.0155	0.0017	8	6	0.03
mogA	619	24	8	155	463	0.0000	0.0036	0	10	NA
c7orf25	531	26	7	118	413	0.0043	0.0023	3	3	-0.89
nfil3	534	26	16	120	414	0.0130	0.0023	5	8	0.49
pigo	494	26	7	127	365	0.0087	0.0014	5	4	-0.48
Sugp2	438	24	5	109	329	0.0000	0.0017	0	4	NA
mastl	537	18	14	119	418	0.1052	0.0194	33	27	0.90
zbed4	471	28	7	109	355	0.0095	0.0042	3	6	0.80
Rassf10	486	18	7	95	376	0.0126	0.0021	3	3	0.99
p7e4	522	26	4	121	401	0.0028	0.0005	3	2	-1.29
fem1c	474	26	16	107	367	0.0312	0.0004	13	2	-0.10
znf238.2	531	26	9	118	413	0.0102	0.0033	4	7	0.36
bcl9	489	26	4	114	372	0.0059	0.0016	3	2	-0.36
nufip2	473	22	10	106	365	0.0135	0.0036	7	5	-0.86
East Africa (*X. victorianus*)										
AR	339	40	7	84	255	0.0007	0.0034	2	4	0.38
prmt6	612	38	10	145	467	0.0104	0.0001	8	1	-0.63
mogA	619	42	10	155	463	0.0009	0.0054	2	9	-1.3
c7orf25	531	40	5	117	413	0.0053	0.0009	3	2	-0.28
nfil3	534	40	9	120	414	0.0052	0.0016	3	5	-0.27
pigo	494	34	7	127	365	0.0052	0.0006	5	2	-1.20
Sugp2	438	42	4	108	330	0.0009	0.0004	2	1	-1.50
mastl	537	34	16	119	418	0.0173	0.0043	11	12	-0.78
zbed4	471	32	10	109	353	0.0039	0.0041	4	8	-1.50
Rassf10	486	42	15	99	387	0.0100	0.0058	4	13	0.12
p7e4	522	40	6	121	401	0.0090	0.0000	6	0	-0.62
fem1c	474	40	11	107	367	0.0137	0.0009	6	4	0.05
znf238.2	531	38	10	117	414	0.0057	0.0034	4	7	-0.73
bcl9	489	40	11	113	370	0.0157	0.0026	7	6	0.18

Gene	bp	No. of alleles	No. of haplotypes	SSites	NSites	pS	pN	SS	SN	DS
nufip2	473	38	16	106	365	0.0192	0.0021	11	10	-0.67

Geographical clustering of variation was observed in the concatenated analysis. Within Central Africa, for example, samples from Nigeria and Cameroon form a clade that is most closely related to a clade comprising samples from Botswana and Zambia. Within East Africa, samples from or near the Lendu Plateau form a clade that is most closely related to a clade containing other samples from the rest of the Albertine Rift and samples from Uganda and Burundi. Within South Africa, geographically structured clades were recovered from multiple regions, including (i) samples from Malawi and northern South Africa (Potchefstroom, Kimberley, Victoria West) and (ii) samples from Niewoudtville, South Africa, (iii) samples from Beaufort West and some samples from Laingsburg and (iv) other samples from Laingsburg plus samples from southwestern Western Cape Province.

Species tree analyses with *Beast (Heled and Drummond 2010) supported the same relationships among most clusters of sequences as the concatenated analysis (Fig. 5.4). The exception to this is that the species tree analysis infers a monophyletic relationship between the two populations that included individuals from the admixed population in Laingsburg, whereas the concatenated analysis supported a paraphyletic relationship between these two populations with respect to other populations (Fig. 5.3,5.4).

Similar to the phylogenetic analyses discussed above, the network analysis of phased *X. laevis* biparentally inherited alleles reveals strong geographic association of molecular variation. Twelve of 15 networks placed molecular variation from southern Africa and the rest of sub-Saharan Africa on distinct portions of the network (Fig. D2.1). Variation in East Africa also tended to cluster in portions of these networks that were distinct from variation in other parts of Central Africa or West Central Africa.

Using the species delimitation program BP&P, we recovered strong support (a posterior probably of ~1) for separate species statuses for all of the clusters in the guide tree, with each cluster corresponding to the terminal taxa presented in Fig. 5.4 for the *Beast analysis. Results were consistent for both species delimitation algorithms, and for both prior settings that we tried.

5.3.3 Population assignment

Tess and Structure population assignment analyses recovered similar results and support the existence of substantial population structure in *X. laevis sensu lato* (Fig. 5.5, D2.2, D2.3). The DIC plots suggest that 6–7 populations are preferred by the Tess analysis and the method of Evanno et al. (2005) supports 5 populations in the Structure analysis. As a consequence of isolation by distance (identified using partial Mantel tests, data not shown), we expected the ad hoc method of Pritchard et al. (2000) to deliver an overestimation of the number of clusters due to departure of the observed data from

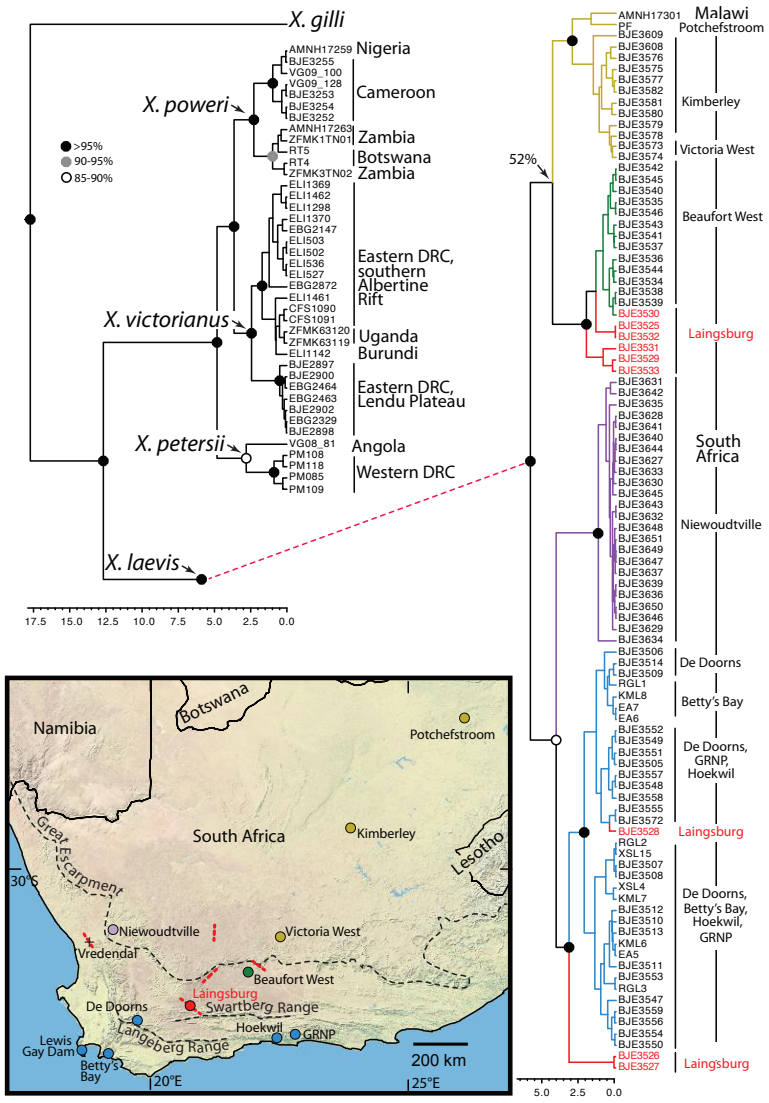


FIGURE 5.3: Phylogenetic analysis of concatenated data from up to 15 autosomal loci per individual. A map shows sampling localities with dots, a plus sign indicates a zone of admixture in Vredendal between mtDNA lineages from Niewoudtville and the south-western Western Cape Province identified by Measey and Channing (2003), and short dotted lines indicate the approximate locations of confirmed or hypothesized contact zones between *X. laevis* populations. Long dotted lines indicate major geological formations. The scale bars indicates divergence time from the present in millions of years.

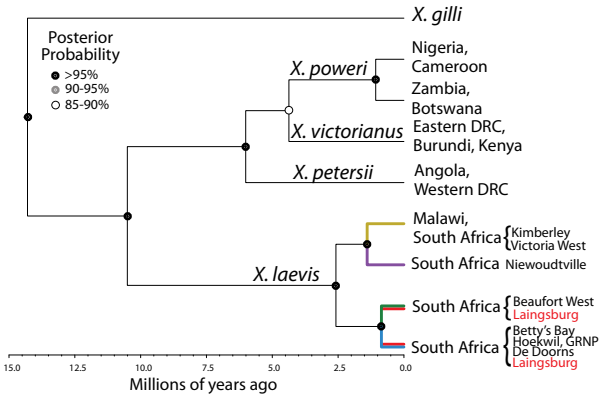


FIGURE 5.4: Species tree analysis by *BEAST. In southern Africa, shaded branches correspond to the sampling localities depicted in Fig. 5.3.

a model of multiple panmictic populations (Pritchard et al. 2000). As expected, this method supported the maximum number of clusters we tested ($K = 10$, $P < 0.001$).

Individuals assigned to each cluster were nearly identical in both analyses at most values of K. Clusters identified by TESS and STRUCTURE at higher values of K corresponded to clades identified in the phylogenetic analyses, and to the species identified by BP&P analysis. Similar to the phylogenetic analyses, these assignment tests also highlight genetic uniqueness of the *X. laevis sensu lato* population from or near the Lendu Plateau and that from Niewoudtville, and also distinguish populations in the northern and southern portions of West Central Africa, the former of which corresponds to a proposed subspecies *X. l. sudanensis* (Perret, Jean-Luc 1966).

5.4 Discussion

5.4.1 Phylogenetic incongruence among maternally inherited loci

We observed well supported, discordant relationships among lineages of two putatively maternally inherited genomic regions in the frog *X. laevis sensu lato*: mtDNA and the female-specific gene *DM-W* (Yoshimoto et al. 2008). This observation could reflect error in phylogenetic inference (that is, an incorrect phylogeny may have been inferred in one or both loci) or it could be a 'real' (biological) difference. Missing data, long-branch attraction and model misspecification, for example, may affect phylogenetic inference (Lemmon and Moriarty 2004; Kück et al. 2012; Roure et al. 2012). A biological difference in these phylogenies could arise if either of these markers was not strictly maternally inherited, or if either experienced recombination. If individuals carrying *DM-W* occasionally developed as phenotypic males, for instance, this could lead to a mode of inheritance that is not strictly maternal. Periodic phenotypic sex reversal coupled with sex specific rates of recombination (specifically, a lower recombination in the heterogametic

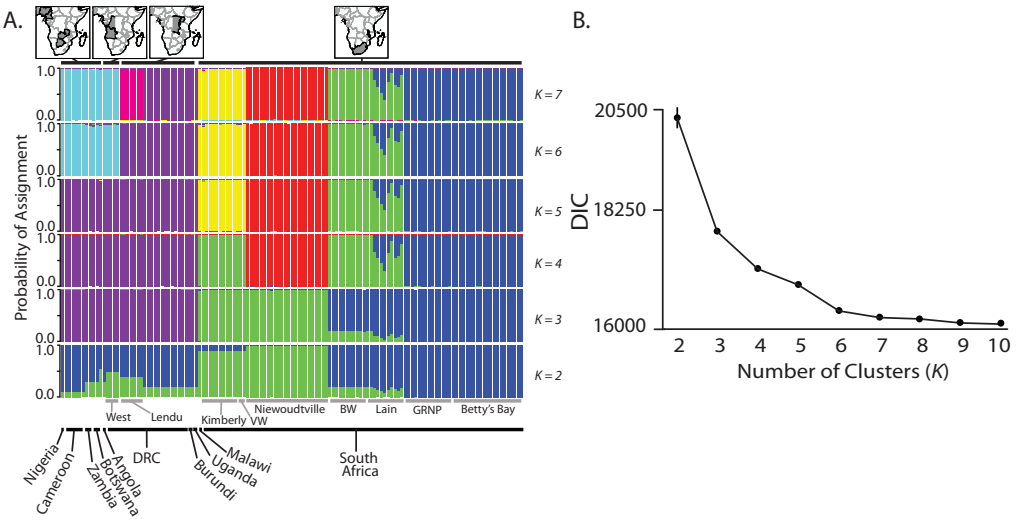


FIGURE 5.5: (A) Results of TESS analysis with population clusters (K) ranging from 2–7. (B) The deviance information criterion (DIC) for each value of K, with bars (most not visible) indicating the standard deviation of this estimate across iterations. In (A), two localities are labeled within the Democratic Republic of the Congo (DRC) including the western DRC (West DRC) and a region including or near the Lendu Plateau (Lendu). Localities in South Africa include Kimberley, Victoria West (VW), Niewoudtville, Beaufort West (BW), Laingsburg (Lain), Garden Route National Park (GRNP) and Betty's Bay.

sex) has been proposed as a mechanism for maintaining nondiverged (homomorphic) sex chromosomes in other frogs (Perrin 2009; Stöck et al. 2011), and indeed, *X. laevis* has homomorphic sex chromosomes (Tymowska 1991). Further information on rates of recombination in male and female *X. laevis* would be useful to evaluate the applicability of this hypothesis to *X. laevis*. It is also possible that *DM-W* or its flanking region exist in duplicate copies in some females, and that these copies could occasionally undergo ectopic recombination events, even if this locus were strictly maternally inherited. While evidence for recombination in mtDNA has been reported in various taxa (Piganeau et al. 2004; Tsaousis et al. 2005), many statistical approaches to detect recombination are prone to false positives (Innan and Nordborg 2002; Galtier et al. 2006; Sun et al. 2011), and we view this as an unlikely explanation for our observations.

5.4.2 Statuses of previously proposed species

Our results provide novel perspectives on the evolutionary history of *X. laevis sensu lato*, and argue for taxonomic revision from the standpoint of the GLC (De Queiroz 1998, 2007). Within *X. laevis sensu lato*, almost all of our analyses recovered support for at least four evolutionarily diverged lineages in the following geographical regions:

(i) southern Africa, including Malawi and South Africa, (ii) Central Africa, including Nigeria, Cameroon, Zambia and Botswana and (iii) West Central Africa, including the Republic of Congo, western DRC and Angola, and (iv) East Africa, including Kenya, Uganda, Rwanda, Burundi, eastern DRC and Tanzania. Each of these groups was identified as a differentiated cluster in the population assignment tests, and lineages i–iii were recovered in each of the phylogenetic analyses we performed (mtDNA, *DM-W* and concatenated and species tree analysis of autosomal DNA). Lineage iv formed a clade in the phylogenetic analyses of autosomal DNA and DM-W, but not in the analysis of mtDNA. These four lineages, respectively, correspond to four currently or previously recognized species: *X. laevis*, *X. poweri*, *X. petersii* and *X. victorianus*, but we argue for a revised distribution for two of them (*X. laevis* and *X. poweri*). A revision of the distributions of *X. laevis* and *X. poweri* is warranted because individuals from the north of the Congo Basin (Cameroon, Nigeria) are more closely related to individuals from the south of the Congo Basin (Zambia, Botswana) than they are to individuals from other parts of Africa, including southern Africa, which is where *X. laevis* occurs. Thus, we reassign the population of *X. laevis sensu lato* from Nigeria and Cameroon to *X. poweri* instead of *X. laevis*. We note that a subspecies of *X. laevis*, *X. l. sudanensis*, from the Adamawa Region in Cameroon was described by (Perret, Jean-Luc 1966). Our data potentially support the transfer of *X. l. sudanensis* to the synonymy of *X. poweri* instead of *X. laevis*, although additional data from the type localities or examination of the type specimens is needed. Similarly, another subspecies of *X. laevis*, *X. l. bunyoniensis* (Loveridge 1932), should be tentatively considered a synonym of *X. victorianus*, as evidenced by the inferred phylogeography of *X. laevis sensu lato* and by phylogenetic position of our sample from south-western Uganda. Although again we note that this study lacks samples directly from the type locality of *X. l. bunyoniensis*, which should be investigated in the future. Under our proposed taxonomy, relationships among mtDNA variants of *X. laevis*, and *X. victorianus* may be paraphyletic within each species; we note also that monophyly is not a requirement of the GLC (De Queiroz 2007).

Although the question of whether further taxonomic division is warranted is beyond the scope of this study, we do note that genetic variation within *X. laevis*, *X. victorianus*, and *X. poweri* is substantial. Within *X. laevis*, differentiated populations were identified in the following regions: (i) south-western Western Cape Province, (ii) Niewoudtville, (iii) Kimberley, Victoria West and Malawi. The south-western Western Cape Province lineage is comprised of two geographically clustered demes with admixture detected at the location of Laingsburg. Individuals from south-western and north-eastern South Africa also differ in body size and in the frequency of naturally occurring testicular oocytes (Du Preez et al. 2009). Within *X. poweri*, a population from Cameroon and Nigeria is differentiated from a population from Botswana and Zambia. Clades within *X. laevis* and within *X. poweri* were delimited from one another by the species delimitation program BP&P. However, significant evidence was recovered for isolation by distance using a partial Mantel test (data not shown), and these data therefore violate an assumption (panmixia) of the BP&P analysis. Within *X. victorianus*, the population from or near the Lendu Plateau is differentiated from other populations. The finding of

substantial genetic differentiation in these species supports the point made by Du Preez et al. (2009) that the geographic provenance of experimental animals is an important experimental variable because of among-population variation in genetic backgrounds.

Our results, which include some of the individuals from Central Africa studied by Evans et al. (2004) and Du Preez et al. (2009), but a different suite of individuals sampled in South Africa from Du Preez et al. (2009), are consistent with the findings from these and other studies (Grohovaz et al. 1996; Kobel et al. 1998; Measey and Channing 2003). Similar to Du Preez et al. (2009), we found evidence for extensive introgression between populations south-west and north-east of the locality of Laingsburg. This was evinced by (i) individuals from this locality having a diversity of evolutionary affinities in the mtDNA, *DM-W* and concatenated analysis of autosomal DNA and (ii) admixed population affinities that were identified by population assignment tests. We did not recover qualitative evidence for extensive gene flow between other populations of *X. laevis sensu lato* based on the population assignment tests. One possibility is that this could be an artefact of missing genetic information from animals in the contact zones between these lineages, for example in the Congo Basin and south of the Congo Basin, or between differentiated populations in South Africa (Fig. 5.1). Reciprocal crosses between *X. laevis sensu lato* individuals that were probably from South Africa, and individuals from Uganda or Botswana both produced fertile offspring of both sexes (Blackler et al. 1965; Blackler and Fischberg 1968). Thus, gene flow between these species is possible. Analysis of additional material from poorly sampled regions therefore could provide novel insights into the nature of gene flow among species and populations identified here.

5.4.3 Phylogeographic implications

Vegetation in sub-Saharan Africa can be broadly classified into 'savanna' habitat, which is open habitat where a C_4 carbon fixation grass layer exists, and 'non-savanna' (i.e. tropical forest) habitat, which is closed and lacks a C_4 carbon fixation grass layer, with the distribution of each habitat type being largely dependent on the extent and seasonality of rainfall (Jacobs 2004; Lehmann et al. 2011). The distributions of these habitat types cycled during climatic oscillations, with savanna habitat becoming more extensive or shifting to lower latitudes during glacial periods (Dupont 2011). Within these habitat types, there is also variation in the seasonality of rain, a factor that may have played a role in the differentiation of *X. laevis* in South Africa (Grohovaz et al. 1996). Thus, over the last 15 My or so, the evolution of *X. laevis sensu lato* took place on a varied and dynamic ecological and climatic landscape. It is also likely that geological features had an impact on population structure within *X. laevis*. In particular, the Great Escarpment (Fig. 5.3) lies between the population that ranges from Victoria West to Malawi and another population that ranges from Beaufort West to Laignsburg (Fig. 5.3). To the south-west of the Great Escarpment, the Cape Fold Belt, including the Swartberge Range and the Langeberg Range (Fig. 5.3), lie between the Beaufort West/Laignsburg population and the coastal population in the south-western Western Cape Province, South Africa. The Niewoudtville population is also on top of the Great Escarpment,

and has a zone of contact with the south-western Western Cape Province population nearby in Vredendal (Measey and Channing 2003), which is at the bottom of the Great Escarpment.

We present four molecular clock analyses (mtDNA, *DM-W*, concatenated autosomal DNA and species tree analysis of autosomal DNA) that assumed a strict molecular clock that was calibrated in two different ways (Methods). Despite these different calibration approaches, divergence times were quite similar across these analyses, although this does not necessarily indicate that these estimates are accurate. We resorted to relatively crude models of evolution in these analyses in order to achieve convergence on the posterior distribution of the parameters. Clearly, error in divergence times and evolutionary relationships could arise due to model misspecification, and other model violations. For example, because *BEAST does not account for migration, divergence times may be underestimated in the presence of migration (Leaché et al. 2013).

Although our divergence estimates for species within *X. laevis sensu lato* predate the Pleistocene, the geographic locations of proposed Pleistocene savanna refugia (see Fig. 2 in Lorenzen et al. 2012) coincide with the distributions of diverged evolutionary lineages in *X. laevis sensu lato*. A possible mechanism for these congruent areas of endemism is that diversification of many of these evolutionary lineages, including *X. laevis sensu lato*, was sculpted by the distributions and connectivity of suitable habitat, which waned and waxed over time. Being mostly aquatic, *Xenopus* presumably are particularly sensitive to ecological factors such as the abundance and seasonality of rainfall that affect opportunities for dispersal over land and time to complete metamorphosis. Regions with consistently habitable habitats potentially acted as 'lifeboats' that sustained divergent lineages that evolved before habitat contraction (Evans et al. 2004).

5.4.4 Conclusions

This study reports for the first time, evidence of phylogenetic discordance between two putatively maternally inherited genomic regions, mtDNA and *DM-W*, in the frog lineage *X. laevis sensu lato*. We do not know of a methodological explanation for this discordance, opening the possibility that there is a biological cause. Results also support the recognition of *X. laevis sensu stricto*, *X. victorianus*, *X. petersii* and newly revalidated *X. poweri*, but with the assignment of populations of *X. laevis sensu lato* from Nigeria and Cameroon to *X. poweri* instead of *X. laevis* and with the assignment of populations of *X. laevis sensu lato* from Botswana and Zambia to *X. poweri* instead of *X. petersii*. In doing so, this study clarifies the evolutionary history of one of the most intensively studied amphibian species in the context of its closely related relatives, and identifies additional differentiated populations that may themselves be meritorious of species status.

5.5 Acknowledgements

We thank four external reviewers, each of whom provided extensive and helpful comments on earlier versions of this manuscript. We thank P. Staab, D. Metzler, J. Measey, J. McGuire, and W. Conradie for helpful discussion or comments on this manuscript, Brian Golding for access to computer resources, and T. Hayes, L. Kalous, R. Tinsley, and P. Wagner for providing genetic samples. We thank M. D. Picker, Z. T. Nagy, M. M. Aristote, W. M. Moninga, M. Zigabe, A. M. Marcel, M. Luhumyo, J. F. Akuku, F. I. Alonda, A. M'Mema, F. B. Murutsi, B. Bajope, M. Manunu, M. Collet, and the Institut Congolais pour la Conservation de la Nature for hospitality, assistance with fieldwork, logistical support and permits. Major support for this study was provided by the National Science and Engineering Research Council of Canada (RGPIN/283102-2012 to B.J.E.). Additional support was from IVB institutional support (RVO: 68081766), the Ministry of Culture of the Czech Republic (DKRVO 2015/15, National Museum, 00023272), the Percy Sladen Memorial Fund, an IUCN/SSC Amphibian Specialist Group Seed Grant, K. Reed, M.D., research funds from the Department of Biology at Villanova University, a National Geographic Research and Exploration Grant (no. 8556-08), the University of Texas at El Paso and the National Science Foundation (DEB-1145459).

5.6 Data accessibility

Sequence data in this study have been deposited in GenBank (new sequences not previously in GenBank have Accession nos. KP343951–KP345838), input files and tree files have been deposited in Dryad (doi:10.5061/dryad.4n2c4) and sampling localities are available in the Table D2.1.

Chapter 6

Limited genomic consequences of hybridization between two African clawed frogs, *Xenopus gilli* and *X. laevis* (Anura: Pipidae)

Benjamin L. S. Furman*, Caroline M. S. Cauret*, Graham A. Colby*, G. John Measey†, Ben J. Evans*†

*Biology Department, Life Sciences Building room 328, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1 Canada
†Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Private Bag X1, Matieland 7602, Stellenbosch, South Africa

This paper was published in *Scientific Reports* and is available here in its published form.

Abstract The Cape platanna, *Xenopus gilli*, an endangered frog, hybridizes with the African clawed frog, *X. laevis*, in South Africa. Estimates of the extent of gene flow between these species range from pervasive to rare. Efforts have been made in the last 30 years to minimize hybridization between these two species in the west population of *X. gilli*, but not the east populations. To further explore the impact of hybridization and the efforts to minimize it, we examined molecular variation in one mitochondrial and 13 nuclear genes in genetic samples collected recently (2013) and also over two decades ago (1994). Despite the presence of F_1 hybrids, none of the genomic regions we surveyed had evidence of gene flow between these species, indicating a lack of extensive introgression. Additionally we found no significant effect of sampling time on genetic diversity of populations of each species. Thus, we speculate that F1 hybrids have low fitness and are not backcrossing with the parental species to an appreciable degree. Within *X. gilli*, evidence for gene flow was recovered between eastern and western populations, a finding

that has implications for conservation management of this species and its threatened habitat.

6.1 Introduction

Gene flow (introgression) between species may facilitate adaptive evolution through the exchange of beneficial genetic variation. This expedites the colonization of specialized ecological niches (Anderson and Hubricht 1938; Dowling and Secor 1997; Rieseberg et al. 2003), and affects future adaptive potential by increasing genetic and phenotypic variation (Anderson et al. 1949; Dowling and Secor 1997; Arnold and Martin 2009; Fitzpatrick et al. 2009; Stelkens et al. 2014). However, gene flow between species also poses risks by eroding species boundaries (Arnold 2006), disrupting adaptively evolved complexes of alleles (Rhymer and Simberloff 1996; Gilk et al. 2004), promoting the exchange of genetic variation associated with disease (Simonti et al. 2016), influencing pathogen emergence (Stukenbrock 2016), and facilitating species invasion (Figueroa et al. 2003; Blair et al. 2012). As such, hybridization has important implications for biodiversity conservation.

6.1.1 Hybridization in African clawed frogs

Hybridization features prominently in the evolutionary history of African clawed frogs (genus *Xenopus*); 28 of 29 species are polyploid, and all of these are probably allopolyploid (Evans 2008; Evans et al. 2015). When backcrossed in the laboratory, there is variation among F_1 *X. gilli-laevis* hybrid females with respect to whether or not their progeny are polyploid (Kobel 1996). Laboratory studies indicate that in some crosses (*X. gilli-X. laevis* and *X. laevis-X. muelleri*) F_1 hybrid males are sterile, but female F_1 hybrids are fertile (Kobel 1996; Malone et al. 2007). F_1 *X. gilli-X. laevis* hybrid females are capable of backcrossing with either parental species, and both sexes of the F_2 backcross generation can be fertile (Kobel 1996). Thus there exists the possibility that gene flow among *Xenopus* species could occur in nature. At least three *Xenopus* hybrid zones are thought to exist (Kobel et al. 1981; Yager 1996; Fischer et al. 2000), and hybrids in each of these zones may have the same ploidy level as the parental species (pseudotetraploid; Tymowska 1991).

6.1.2 The *X. gilli* / *X. laevis* hybrid zone

Classified by the IUCN as Endangered (*South African Frog Re-assessment Group (SA-FRoG), IUCN SSC Amphibian Specialist Group. 2010. Xenopus gilli. The IUCN Red List of Threatened Species 2010: e.T23124A9417597* n.d.), *X. gilli* (Rose and Hewitt 1926) occurs in southwestern Western Cape Provence, South Africa (Picker and Villiers 1989; Evans et al. 1997, 1998; Fogell et al. 2013). *X. gilli* is a found in seasonal ponds in lowland coastal fynbos habitat, a component of the Cape Floristic Region, which is a biodiversity hotspot (Myers et al. 2000) with an extreme level of plant endemism (Kier et al. 2009). These ponds have high concentrations of humic compounds derived from the surrounding fynbos vegetation, and a characteristic dark color and low pH

(Mitchell et al. 1986; Picker and Villiers 1989; Picker et al. 1993). The range of *X. gilli* is disjunct and includes the Cape of Good Hope section of Table Mountain National Park (CoGH), habitat near the town of Kleinmond, and habitat near the town of Pearly Beach (Picker and Villiers 1989; Evans et al. 1997, 1998; Fogell et al. 2013, Fig. 6.1). These three localities are interrupted by unsuitable, highly modified habitat that may impede contemporary gene-flow (Fogell et al. 2013). As with many amphibians (Marsh and Trenham 2001), habitat degradation is a major threat to *X. gilli* (Picker and Villiers 1989; Fogell et al. 2013).

In contrast, *X. laevis* (Daudin 1802), is found throughout southern Africa, in both natural and disturbed areas of South Africa and Malawi (Tinsley and Kobel 1996; Furman and Evans 2016). *X. laevis* is syntopic throughout the range of *X. gilli* (Picker and Villiers 1989; Evans et al. 1997, 1998; Fogell et al. 2013) and can tolerate a broad spectrum of environmental challenges including extremes of desiccation, salinity, anoxia, and temperature (Measey et al. 2012). Picker et al. (1993) proposed that there may be an ecological basis for speciation of *X. laevis* and *X. gilli* centered on higher tolerance of *X. gilli* embryos to low pH, allowing for habitat specialization.

Several aspects of external morphology readily distinguish these species, including smaller size of *X. gilli*, the presence of longitudinal dorsal mottling that does not connect over the midline in *X. gilli* only (for example, see Fig. 1 of Kobel et al. 1981), and orange and black vermiculation on the venter of *X. gilli*. F_1 hybrids between *X. gilli* and *X. laevis* are readily identified based on individuals that are morphologically intermediate with respect to size and coloration, and this identification has been confirmed by molecular tests (Kobel et al. 1981; Picker 1985; Picker et al. 1996; Evans et al. 1998). The reported abundance of F_1 hybrids varies from relatively common (Picker 1985; Picker et al. 1996; Fogell et al. 2013), to rare (Evans et al. 1997, 1998). Morphological variation of some individuals has been previously interpreted as being derived from backcrosses of F_1 hybrids with each parental species (Picker et al. 1996).

The western extent of the *X. gilli* distribution occurs within the CoGH (Picker and Villiers 1989; Fogell et al. 2013). Following reports of hybrids and expansion of *X. laevis* populations, steps were taken in the mid-1980s to minimize co-occurence of these two species within the CoGH which included removal of *X. laevis* from *X. gilli* ponds, translocation of *X. gilli* to new sites (Measey et al. 2011), and construction of a wall around a known *X. gilli* pond (Picker and Villiers 1989; Villiers et al. 2016). The hope was to minimize hybridization and resource competition, for example, if larger *X. laevis* individuals are able to outcompete *X. gilli* for food (Picker et al. 1996; Vogt et al. 2017). With some interruptions, these efforts have continued for the last 30 years in the CoGH. Similar efforts have not been made for eastern populations of *X. gilli* which are located on private property, and in some ponds in these areas where *X. gilli* had been found in the past, now only *X. laevis* are found (Fogell et al. 2013; Villiers et al. 2016).

To further investigate the effect of hybridization on gene flow between *X. gilli* and *X. laevis*, we examined DNA sequence variation in these species from one mitochondrial DNA (mtDNA) marker and 13 nuclear DNA (nDNA) markers. Genetic samples were

collected from within managed (west) and unmanaged (east) portions of the range of *X. gilli*. Samples were analyzed from both locations that were collected shortly after management began, and then in the same areas again 20 years later. We expected that if introgression was occurring in the populations during this time period, it would be more pronounced in the east population. If efforts to minimize hybridization in the west were successful, we expected more evidence of gene flow in the samples collected soon after management began as compared to more recently. However, in both localities and both sampling times, we found no evidence of shared mitochondrial haplotypes or nuclear alleles between these species, suggesting that the F_1 hybrids have low fitness and are not backcrossing with the parental species to an appreciable degree, despite potential fertility of F_1 females (Kobel 1996). Within *X. gilli*, we recovered evidence of gene flow between east and west populations, and found genetic diversity to be higher in the unprotected eastern population. These findings have implications for management and conservation of this endangered habitat specialist.

6.2 Materials and Methods

Genetic samples analyzed in this study were collected either in 1994 or in 2013. Some of the samples from the earlier collection were also analyzed in two earlier studies (Evans et al. 1997, 1998). The 2013 collection included *X. gilli* and *X. laevis* individuals from the same or geographically close (within 5 km) sites as the 1994 collection, and both sampling efforts used funnel traps. Animal sampling protocols approved by the Institutional Animal Care and Use Committee at Columbia University and work was performed in accordance with all relevant guidelines and regulations for animal experimentation, in accordance with laws for studying wildlife in South Africa and with appropriate collection permits from the Chief Directorate of Nature Conservation and Museums, and was approved by the Animal Ethics Committee at the University of Cape Town and the Stellenbosch University Research Ethics Committee: Animal Care and Use. Samples were obtained east and west of False Bay for both species and for both time periods (Fig. 6.1). We assigned individuals to species (*X. gilli* or *X. laevis*) based on dorsal and ventral patterning, shape of head, and overall size (Kobel 1996; Villiers 2004). Because this study aimed to explore genetic effects of backcrossed hybrids, for both time points, we intentionally excluded individuals whose intermediate morphology (and genetic analysis in the case of the 1994 individual (Evans et al. 1998)) indicated that they were F_1 hybrids (1 individual from 1994 and 9 from 2013).

DNA was extracted from tissue samples using Qiagen DNEasy tissue extraction kits (Qiagen, Inc), following the manufacturer's protocol, or a phenol-chloroform protocol. A fragment of the mtDNA genome was amplified and sequenced for 36 and 33 *X. gilli* and *X. laevis* individuals, respectively, using primers from Evans et al. (2004) that target a portion of the 16S ribosomal RNA gene (*16S*). Exons of 13 nDNA genes ranging from 333–770 bp in length were sequenced for 20–41 *X. gilli* and 11–31 *X. laevis* individuals using paralog specific primers (primers are from Bewick et al. 2011). These

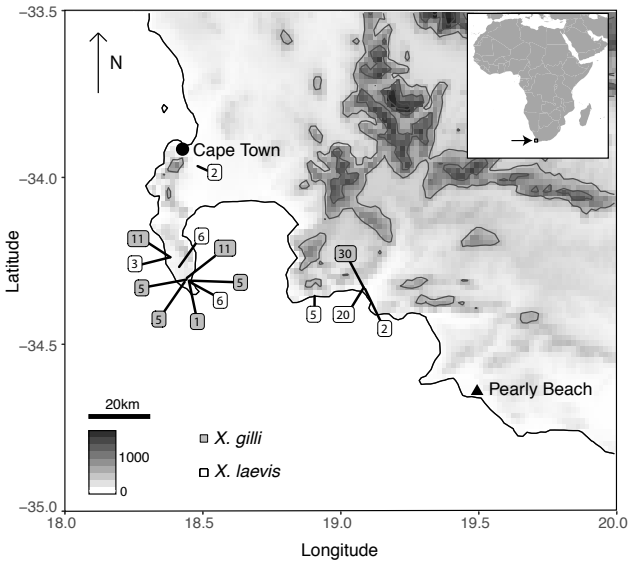


FIGURE 6.1: Sampling locations. For each species, numbers indicate the sum of number of individuals from each locality sampled in 1994 and 2013. An inset indicates the study area in southern Africa and altitude in meters is indicated on the scale. The map was made using the R package MARMAP (Pante and Simon-Bouhet 2013) using topographic data from the National Oceanic and Atmospheric Administration, USA.

exons came from the genes B-Cell CLL/Lymphoma 9 (*BCL9*), BTB domain containing 6 (*BTBD6*), Chromosome 7 Open Reading Frame 25 (*C7orf25*), Fem-1 Homolog C (*FEM1C*), Microtubule Associated Serine/Threonine Kinase Like (*MASTL*), Mannosyl-oligosaccharide glucosidase (*MOGS-1*), Nuclear Factor Interleukin 3 Regulated (*NFIL-3*), protocadherin 1 (*PCDH1*), phosphatidylinositol glycan anchor biosynthesis class O (*PIGO*), protein arginine methyltransferase 6 (*PRMT6*), Ras association domain family member 10 (*RASSF10*), SURP and G-patch domain containing 2 (*SUGP2*), and zinc finger BED-type containing 4 (*ZBED4*). A table of sample IDs and which loci were amplified for which samples is available in Table E1.1. In the phylogenetic analysis of individual genes (discussed below), we used as an outgroup a sequence from *Xenopus Silurana tropicalis* from the genome assembly version 9.0 on Xenbase (Bowes et al. 2009). When possible, we also included orthologous and homeologous sequences from *X. laevis* from the genome assembly version 9.1 on Xenbase (Bowes et al. 2009), which was identified using BLAST (Altschul et al. 1997b); this was not possible when a homeologous sequence was not identified, which could be due to gene loss or missing data in the genome sequence. Sequence data were aligned using MAFFT (Katoh and Standley 2013) and corrected by eye. Coding frame was estimated using the 'minimize stop codons' option in MESQUITE v.3.04 (Maddison and Maddison 2015), and alignments were trimmed to begin at the first position and end at the third position of the reading frame.

We calculated the phase of nDNA alleles (i.e. haplotypes) using the 'best guess'

option of PHASE (Stephens and Donnelly 2003; Stephens and Scheet 2005) with default parameters. Each individual's allelic sequences for each locus were used in subsequent population genetic, clustering, and gene tree analyses. Thus, for each nuclear locus, an individual frog was represented by two sequences, each corresponding to one allele.

6.2.1 Gene trees

Gene trees were estimated for each phased nDNA exon and the mtDNA alignment using BEAST v1.8.3 (Drummond et al. 2012). Substitution models were selected based on the Akaike Information Criterion using MRMODELTEST v.2 (Nylander 2004), and xml files were prepared for BEAST using BEAUTI (part of the BEAST package). For each nDNA locus, we ran two Markov chain Monte Carlo runs for 25 million generations. For the mtDNA, the model selected by MRMODELTEST2 (GTR+Γ) failed to converge on stable parameter estimates, and we therefore instead used the simpler HKY+Γ model, and ran two chains for 50 million generations. For each analysis, convergence of parameter estimates on the posterior distribution was assessed using TRACER v.1.555 based on an effective sample size (ESS) value > 200 and inspection of the trace of parameter estimates against the MCMC generation number. Based on this, for all phylogenetic analyses the first 25% of the posterior distribution was discarded as burn-in. Then, using TREEANNOTATOR, we produced consensus trees from the post-burn-in posterior distribution of trees.

6.2.2 Species tree

We also estimated a species tree (with the nuclear sequences used in the STRUCTURE analysis, see below) using the multi-species coalescent model of *BEAST (Heled and Drummond 2010). We trimmed the dataset to include only nDNA genes with all populations sampled (see Genetic clusters section for details) and included only individuals sampled for all genes. All *X. laevis* individuals were considered to be the same species (17 individuals), and we separated the east and west *X. gilli* populations into separate species (10 and 11 individuals, respectively), and *X. tropicalis* was considered its own species. We set a simple HKY+Γ model joined for all data partitions (so that convergence of parameter estimates could be reached), assumed a strict molecular clock joined for all data partitions, and allowed the underlying gene tree structure to vary across data partitions. We ran the 8 chains for 170 million generations and removed 50 million generations as burn-in.

6.2.3 Genetic clusters

We used STRUCTURE v.2.3.4 (Pritchard et al. 2000) to estimate individual assignment probabilities to genetic clusters using best-guess phased nDNA alleles on a subset of individuals. Three loci lacked data from the east *X. gilli* 1994 population (exons of the

genes *MOGS-1*, *PCDH1* that also lacked data from this exon for *X. laevis* east 1994 samples, and *PIGO*), so we excluded them from STRUCTURE analysis. We also excluded individuals with $> 50\%$ missing data for the remaining 10 loci. This resulted in a dataset of 13, 8, 11, and 6 *X. gilli* individuals from the following localities and years respectively: east 1994, east 2013, west 1994, and west 2013, where east and west refer to the sampling locations relative to False Bay. This analysis also included 9, 12, 6, 4 *X. laevis* individuals from east 1994, east 2013, west 1994, and west 2013 respectively. We used the admixture model of STRUCTURE and assumed no correlation between alleles at different loci. We ran the Markov chain Monte Carlo for 20 million generations, following a two million generation burn-in. We tested a number of clusters (K) ranging from 1–8, with 5 replicate analyses for each setting of K. To correct for label switching and to average assignment probabilities across runs, we used CLUMPP v.1.1.2 (Jakobsson and Rosenberg 2007). We first computed the D statistic, following recommendations in the CLUMPP manual (Jakobsson and Rosenberg 2007), to decide on the particular algorithm to employ for maximizing similarity across runs. We then used the *ad hoc* method of Pritchard et al. (2000) and the method described by Evanno et al. (2005) to evaluate the most likely number of genetic clusters (K).

6.2.4 Evolutionary models

As discussed below, our analyses did not detect mtDNA haplotypes or nuclear alleles that were shared between *X. laevis* and *X. gilli* but did detect shared haplotypes and alleles between populations of *X. gilli* and between populations of *X. laevis*. Because *X. gilli* is of conservation concern, we evaluated the fit of data from this species to three evolutionary models (Fig. 6.2). In the first (isolation) model, divergence of two *X. gilli* populations was followed by no migration between each population. Under the isolation model, all shared alleles between these populations would be due to incomplete lineage sorting (ILS). In the second (ongoing migration) model, *X. gilli* population divergence was followed by ongoing symmetrical migration between the populations. Under the ongoing migration model, shared alleles would be due to ILS or migration, and some of the alleles shared due to migration could have been exchanged millions of years ago. In the third (secondary contact) model, *X. gilli* population divergence was followed by a period of no migration and then by a period during which symmetrical migration occurred between east and west populations. Under the secondary contact model, shared alleles would again be due to ILS or migration, but alleles shared due to migration could only have been exchanged recently.

All models include a parameter T, which is the time of separation between the *X. gilli* populations and a parameter θ, which is the population polymorphism parameter of the ancestral and both descendant populations. The second and third models have an additional parameter m, which is the number of individuals in each population that are replaced per generation by individuals from the other population (east vs west), divided by the product of four times the effective population size of each population. The third model includes another parameter τ, which is the proportion of T going back in time from

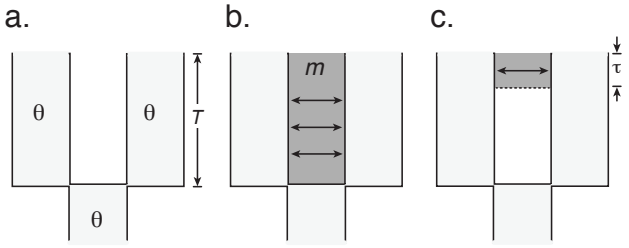


FIGURE 6.2: Evolutionary models considered for *X. gilli* sequence data from east and west populations included (a) population division without subsequent gene flow, (b) separation followed by ongoing gene flow, and (c) separation followed by secondary contact after a period of no gene flow. Model parameters include the population polymorphism parameter θ, which is assumed to be constant in the ancestral and both descendant populations, the time of speciation T, the amount of migration m, and the time of secondary contact τ.

the present that secondary contact began. Thus, the ongoing migration and isolation models are special cases of the secondary contact model in which $\tau = 1$, or $\tau = 1$ and $m = 0$, respectively. We note that several assumptions of these models are undoubtedly violated (e.g., constancy of population size over time, equivalent population size of both descendant and the ancestral populations) but we made them nonetheless so we could complete the simulations (see below) within a reasonable amount of time, and because of the relatively small size of the dataset.

The approximate likelihood of combinations of values for these parameters was estimated using rejection sampling (Weiss and Haeseler 1998). In this approach, the likelihood is approximated by the natural logarithm of the number of simulations for which the sum of four summary statistics from a simulation (discussed next) were within $\pm$ ϵ % of the sum of the observed four summary statistics from actual sequence data, divided by the number of simulations, where $\epsilon = 25$. The value of ϵ determined how close the simulations must match the observed data in order to contribute to the likelihood, and was selected based on a compromise between the computational efficiency of the likelihood estimation and the accuracy of the estimate (Weiss and Haeseler 1998). For the ongoing migration model and the secondary contact model, 40,000 simulations were performed for each combination of parameter values we considered. For the isolation model, no simulations had summary statistics within $\pm$ ϵ % of the observed; thus, 1,000,000 simulations were performed in order to achieve an upper bound for the likelihood estimate. The likelihood of the data over all combinations of the following parameter value intervals were estimated: T: every 1,000,000 generations in the interval of 0–20,000,000 generations; θ: every 0.001 units in the interval of 0.001–0.01 and every 0.01 units in the interval of 0.01–0.1; τ: every 10% in the interval of 10–100%; m every 0.1 units in the interval of 0–1 and every integer in the interval of 1–10.

We used the sum across loci of four summary statistics described by Becquet and

Przeworski (2007) for these likelihood calculations, and simulations were performed using the program mimarsim (Becquet and Przeworski 2009). These four summary statistics include the number of sites with a derived polymorphism (i) in the west population of *X. gilli* only, (ii) in the east population of *X. gilli* only, (iii) shared between the west and east populations of *X. gilli*, or (iv) fixed in either the west or in the east population of *X. gilli*. The simulations used a fixed value for the mutation rate equal to $2.69e^{-9}$ substitutions per site per generation, which was estimated based on the average synonymous divergence between a randomly selected *X. gilli* sequence and an orthologus sequence from *X. tropicalis*, and assuming a divergence time of 65 million years for the separation of these lineages (Bewick et al. 2012), and a generation time of one year. Each locus had a mutation rate scalar based on synonymous divergence to *X. tropicalis* that accommodated variation among loci in the rate of evolution. To minimize the influence of natural selection on the polymorphism data, summary statistics and likelihood calculations were based only on variation at synonymous positions. Confidence intervals were estimated using the profile likelihood method (i.e., that the 95% confidence interval is defined by the two points that are 1.92 log-likelihood (lnL) units from the maximum).

6.2.5 Population dynamics over time and space

We performed various analyses to assess whether the genetic diversity varied among these species, over time, or among populations east and west of False Bay. Pairwise F_{ST} (with significance computed by a permutation test) was quantified for the same data used in the STRUCTURE analysis using ARLEQUIN v3.5.2.2 (Excoffier and Lischer 2010). Nucleotide diversity (π) of each locus was calculated using the pegas package in R (Paradis 2010; R Core Team 2017). We then calculated a mean value of π across loci for each of the eight populations, weighting the estimate by gene length for each locus. Confidence intervals were obtained by bootstrapping the weighted π values 5000 times.

Because allelic diversity is influenced by sample size, we used the program HP-RARE to calculate rarefied estimates of allelic diversity (Kalinowski 2005), which involves down-sampling data to the smallest number of samples in each population across all nuclear loci for which there were data. This analysis was performed with the same data as used in the STRUCTURE analysis. For *X. laevis* populations there was one exception; the *PRMT6* locus had only four sampled alleles for the *X. laevis* west 1994 population, thus we did one run with all of the data (using four as the smallest number of sampled alleles) and another run excluding *PRMT6* (in which case, eight was the smallest number of sampled alleles). For all *X. gilli* populations, the smallest number of sampled alleles was eight. We generated confidence intervals by bootstrapping of the allelic diversity measurements 5000 times.

To statistically evaluate differences in genetic diversity over time, location and species, we constructed linear mixed models using the R package LME4 (Bates et al. 2015). We built models for the estimated values of nucleotide diversity (π) and allelic diversity independently with diversity values measured for each locus, using time (1994 or 2014),

location (east or west) and species (*X. gilli* or *X. laevis*) as fixed effects (all additive, no interaction terms) and considering locus as a random effect. For each parameter of both models, we also used lme4 to compute confidence intervals with the `confint` function.

6.3 Results

6.3.1 Molecular polymorphism and gene trees

In the mitochondrial and 13 nuclear gene trees, alleles from *X. gilli* and *X. laevis* clustered in reciprocally monophyletic clades (Fig. 6.3, Fig. E1.1). No individuals were found to have introgressed loci, which would have been evidenced by an allele in one species having a closer relationship to the alleles of the other species (i.e., a paraphyletic relationship). Similar to previous studies (Evans et al. 1997; Fogell et al. 2013), the mtDNA gene tree identified divergence between the east and west populations of *X. gilli* (Fig. 6.3). We identified one individual (Sample ID: XgUAE_08) from the west population of *X. gilli* that carried a mtDNA haplotype that was more closely related to haplotypes that were carried by individuals from the east population. This observation was also reported previously, from different samples (Evans et al. 1997; Fogell et al. 2013). The *Beast analysis recovered the expected species tree of these three species with posterior probabilities of one (Fig. E1.2). This analysis estimated the divergence time of *X. gilli* and *X. laevis* at about 14.05 my and divergence of the east and west *X. gilli* populations at about 1 my (0.51–1.36 my 95% HDP; when a calibration point of 65 my from *X. tropicalis* is assumed Bewick et al. 2012).

6.3.2 Genetic clusters

Structure analyses assigned each individual to groups that corresponded with species assignment (Fig. 6.4a). All *X. laevis* individuals were assigned to a single genetic cluster at $K = 2$–8, indicating a lack of allele frequency clustering, which is consistent with gene flow across the population range. The *X. gilli* samples were assigned to two clusters corresponding to sampling location (east and west) at $K = 3$–8, indicating differences in allele frequencies, which is consistent with restricted gene flow between them (Fig. 6.4a). Assignment of individuals to clusters stabilized at $K = 3$, with no new clusters being detected at higher values of K (Fig. 6.4a). The likelihood plot plateaus at $K = 3$ (Fig. 6.4b); the Evanno method (Evanno et al. 2005) supports $K = 2$ and the *ad hoc* method of Pritchard et al. (2000) supports $K = 3$.

6.3.3 Evolutionary models

Using simulations and summary statistics, we evaluated the fit of the *X. gilli* data to evolutionary models with no migration after speciation, with ongoing migration after

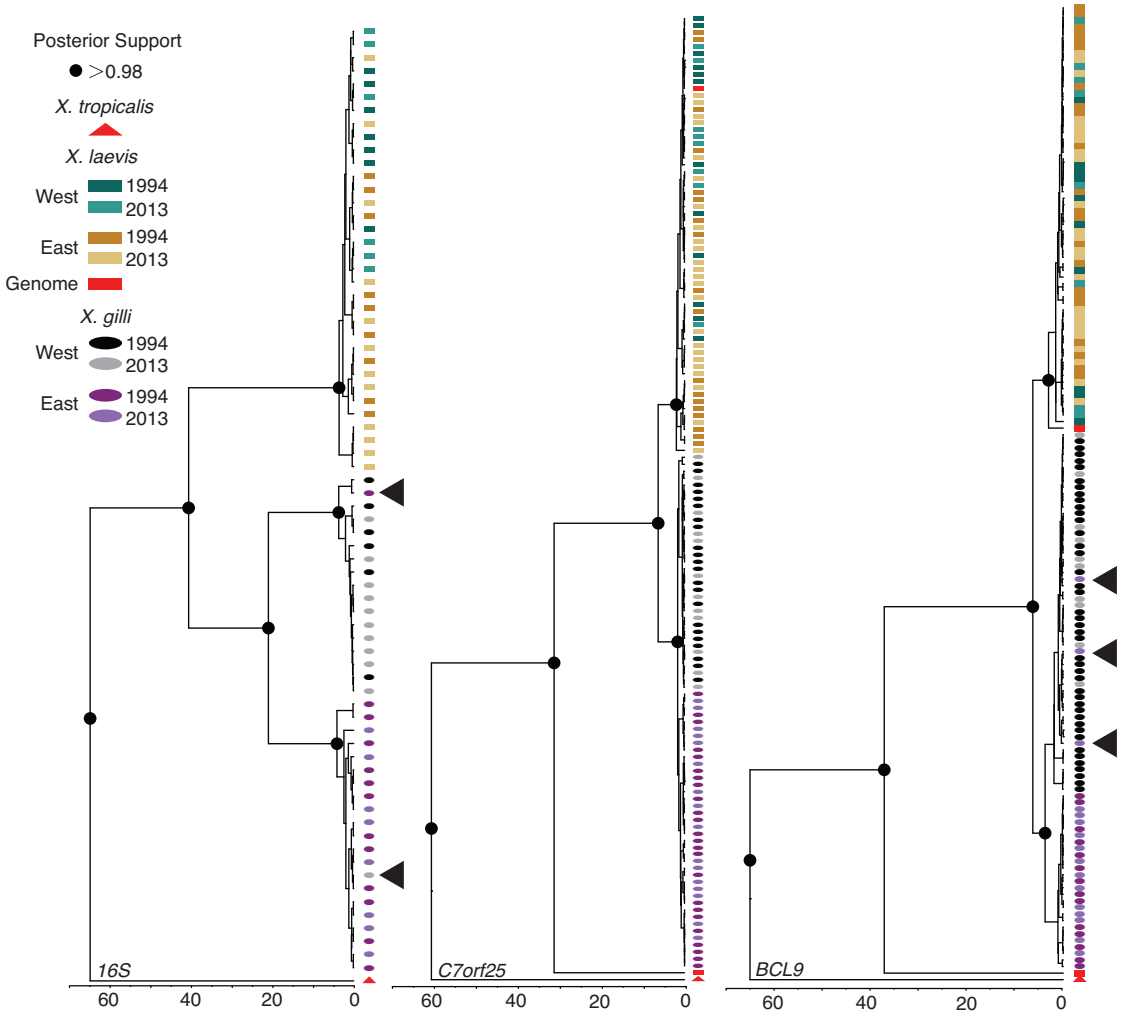


FIGURE 6.3: Representative gene trees that collectively provide no evidence of genetic exchange between *X. gilli* and *X. laevis*. The phylogeny on the left illustrates divergence between 16S rDNA mitochondrial sequences in the east and west populations of *X. gilli*, and with one shared sequence (indicated with an arrow) that occurred on both sides of False Bay. The nuclear phylogenies in the center and right provide examples of no shared alleles and shared alleles between the east and west *X. gilli* populations, respectively. Gene name acronyms are described in the Materials and Methods section. These and other phylogenies are depicted with sample labels in Fig. E1.1.

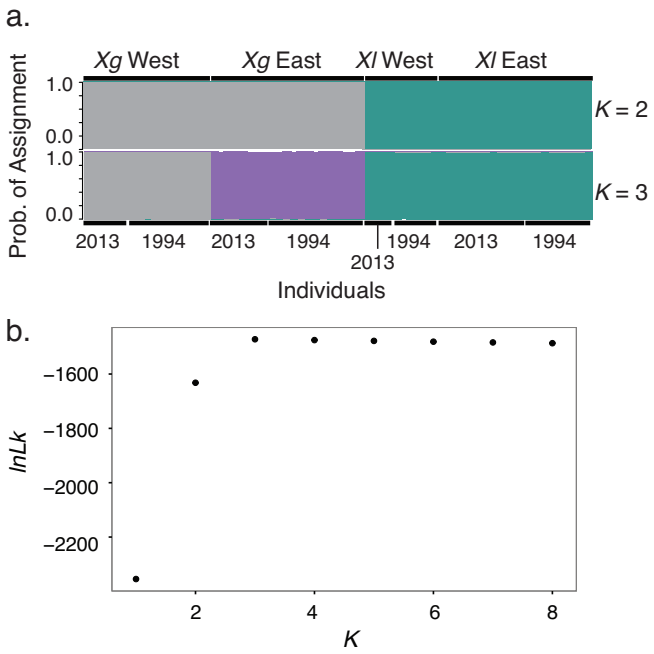


FIGURE 6.4: (a) Structure analyses for 10 loci, which had sequence data for all populations. (b) Likelihood for each value of K.)

speciation, or with secondary contact after speciation. The lnL of the secondary contact model was -8.032, the ongoing migration model was -8.987, and the isolation model was < -13.815. We were not able to more precisely estimate the likelihood of the isolation model because no simulations under this model resulted in data whose four summary statistics were within $\pm\ \epsilon$ of the observed values (see Methods).

Nested models can be compared by assuming that twice the difference between the lnL of each model follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters in each model (denoted χ^2_1 for comparison between models that differ in one parameter). However, because comparison between these successively more complex models involves a boundary condition on one parameter ($\tau = 1$ for the ongoing migration model, $m = 0$ for the isolation model), this difference in model likelihoods follows a mixture of χ^2_0 and χ^2_1 distributions (Self and Liang 1987). The secondary contact model is thus not supported over the ongoing migration model ($p = 0.08$), but the ongoing migration model is supported over the isolation model ($p = 0.009$). Overall then, these results support an inference of gene flow between *X. gilli* populations, but fail to discern substantial temporal heterogeneity in the level of gene flow.

The maximum likelihood parameter estimates and 95% confidence intervals for the ongoing migration model were θ: 0.002 (0.001–0.003) and m: 0.7 (0.1–2) individuals/generation. The maximum likelihood estimate for T was 8,500,000 generations; the 95% CI was unable to be estimated because it exceeded the boundaries we tested (1,000,000–20,000,000), suggesting low statistical power to estimate this parameter.

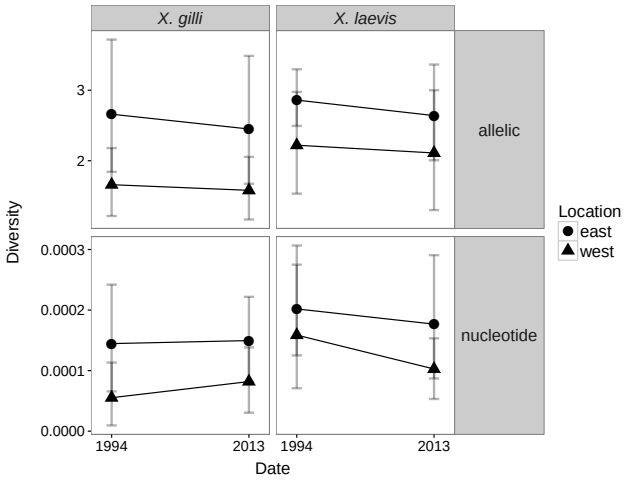


FIGURE 6.5: Genetic diversity statistics including rarefied estimates of allelic diversity (top panels) and nucleotide diversity (π) weighted by length of sequence; bottom panels). For allelic diversity, the analysis considered the same 10 loci as were analyzed by the STRUCTURE analysis (see Materials and Methods). Allelic diversity for *X. laevis* did not include *PRMT6* locus because this locus only had four alleles for the west 1994 population.

Comparisons to similar parameters estimated for African clawed frogs by other studies using other methods (Evans et al. 2011b, 2015; Furman et al. 2015) suggest that these estimates are biologically plausible. Our intuition that the shared identical alleles between east and west *X. gilli* populations are due to ongoing migration is thus supported, with caveats that several model assumptions, discussed below, are violated to some degree.

6.3.4 Population dynamics over time and space

In line with results from STRUCTURE analysis, a high F_{ST} was measured in all pairwise comparisons of the east and west populations of *X. gilli* (comparing within the same year 2013 east to 2013 west, and between years 1994 east to 2013 west and 2013 east to 1994 west; range: 0.55–0.60, $p < 0.05$). For *X. gilli*, between time points within each location (east or west), F_{ST} was not significantly different from zero (east 1994 to east 2013 and west 1994 to west 2013; $p > 0.05$, $F_{ST} < 0.02$). For *X. laevis*, pairwise comparisons of east and west populations, within the same year (1994 east to 1994 west or 2013 east to 2013 west) and between time points (1994 east to 2013 west and 2013 east to 1994 west), had intermediate F_{ST} values that departed significantly from zero ($p < 0.05$, F_{ST} = 0.07–0.16). But within locations comparing time points (1994 east to 2013 east and 1994 west to 2013 west), F_{ST} was not significantly different from zero ($p > 0.05$, F_{ST} = 0.04 for both comparisons).

Both nucleotide diversity and allelic diversity did not change drastically over time, but within species, both statistics were higher in the east population than the west (Fig. 6.5). In the linear mixed model analysis of π, the effect of species was significant with *X. laevis* higher than *X. gilli* by 0.00073 substitutions per site (95% CI: 0.00027–0.00119). The effect of location was significant with π lower in the west than the east population by 0.00091 (95% CI: 0.00045–0.00137). The effect of time of sampling was not significant, with the 2013 samples being lower by 0.00016 but the 95% CI of this difference spanning zero (-0.0006–0.00030). Similar results were recovered for allelic diversity, with *X. laevis* having higher allelic diversity than *X. gilli* (0.59, 95% CI: 0.21–0.97), the west being less diverse than the east (-0.77, 95% CI: -1.14– -0.40), and no significant effect of sampling time (-0.15, 95% CI: -0.52–0.21).

6.4 Discussion

6.4.1 Gene Flow between *X. laevis* and *X. gilli*

Previous investigations of the genetic consequences of hybridization between *X. laevis* and *X. gilli* found no evidence of widespread genetic introgression (Evans et al. 1997, 1998), a result that seemed to be at odds with the incidence of morphologically and genetically identified hybrids in this and other studies (Rau 1978; Kobel et al. 1981; Picker 1985; Picker et al. 1996; Fogell et al. 2013). In this study, we analyzed many of the samples from Evans et al. (1997, 1998) and also genetic samples that were collected more recently. Evidence of introgression between *X. laevis* and *X. gilli* was not detected in mtDNA or in any of 13 nuclear loci (Fig. 6.3;Fig. E1.1). Furthermore, each species formed separate genetic clusters with no evidence for similarities in allele frequencies (Fig. 6.4a). These findings were consistent in both sampling efforts examined here, which included targeting both populations of *X. gilli* and sampling time points separated by about two decades. Previous investigations into the extent of genetic introgression (Evans et al. 1998), used two nuclear loci that were not used in this study. Combining that study with ours brings the total number of genomic regions studied to 15, and includes variation from 6 of the 18 chromosome pairs based on gene location in the *X. laevis* version 9 genome, on Xenbase. This expanded sampling is thus consistent with the interpretation by Evans et al. (1998) that genomic introgression is not extensive.

The lack of introgression is despite the continued identification (based on morphology) of a low frequency of putative F_1 hybrids in both localities. Though there could be an adaptive benefit for hybridization because *X. gilli* embryos can tolerate ponds with higher pH levels than *X. laevis*, which perhaps could allow for invasion of *X. gilli* habitat, we found no evidence that hybridization has led to gene flow of the genetic basis of this or other ecological adaptations that evolved after these two species diverged from their most recent common ancestor. Although not the focus of this study, the relatively low abundance of F_1 hybrids argues against the possibility that a new species of hybrid origin is evolving in this zone of sympatry between *X. laevis* and *X. gilli*. Reproductive isolation

in amphibians has been shown to happen in a few million years for some lineages (Colliard et al. 2010; Dufresnes et al. 2015). In *Xenopus* species, female individuals respond to species-specific calls evoked by males (phonotaxis) (Picker 1980) and this presumably acts to some degree as a prezygotic barrier to hybridization. However, an observation is that at high densities, *Xenopus* individuals amplex indiscriminately (G. J. Measey, personal observation), potentially overriding some prezygotic barriers. In some ponds, *X. gilli* individuals can be outnumbered by *X. laevis* 3:141, and indiscriminate amplexus could mean *X. gilli* males (which are also smaller) are outcompeted for access to females. This may be why hybrids are occasionally seen, but the extended period of divergence between these species (~14 million years (my)) appears to have resulted in strong post-zygotic barriers preventing introgression.

Hybridization followed by back-crossing is expected to generate a mosaic of introgressed and non-introgressed genomic regions. Variation among genomic regions in the extent of introgression can be further augmented by natural selection favoring or disadvantaging genetic variants from one species in the genomic background of the other (Arnold and Martin 2009). In California tiger salamanders (*Ambystoma californiense*), for example, some loci are fixed for foreign alleles from the introduced barred tiger salamander (*A. mavortium*), whereas other loci exhibit no sign of introgression (Fitzpatrick et al. 2009). That the barred tiger salamander was introduced only 60 years ago suggests that mosaicism of genomic introgression arose rapidly (in ~20 generations; Fitzpatrick et al. 2009). In this study it is therefore possible that we failed to identify some introgressed regions of the genome due to the relatively sparse sampling of genomic regions. Future studies that survey variation across the entire genome, such as RADSeq (Davey et al. 2011), could more precisely quantify the extent of gene flow between these species, if it occurs.

6.4.2 Population structure in *X. gilli* and change over time

Analysis of mtDNA (Evans et al. 1997; Evans et al. 2004; Fogell et al. 2013) and skin peptides secreted by these populations (Conlon et al. 2015) support the existence of at least two distinct populations in *X. gilli* in the western and eastern portion of its range. Our mitochondrial analysis, STRUCTURE analysis, and some of the gene trees reported here (such as *MOGS-1* and *RASSF10*) also exhibit substantial geographic differences in *X. gilli* allele frequencies between these populations (Fig. 6.4,6.3,E1.1). In contrast, genetic diversity in *X. laevis* has minimal geographic structure, with most alleles occurring on both sides of False Bay, and STRUCTURE analyses assigning all *X. laevis* individuals to a single genetic cluster (Fig. 6.4a,E1.1). This is similar to findings reported by Evans et al. (1997).

When and why did population structure arise in *X. gilli*? Using mtDNA sequence data and a relaxed molecular clock, Evans et al. (2004) estimated that the divergence between *X. gilli* populations occurred 8.5 my (95% CI: 4.8–13.4), which is the same as the estimate obtained here using our coalescent modeling approach. This estimate is

older than the 1 my divergence time estimated by the *Beast analysis (Fig. E1.2), but this is not unexpected because *Beast does not incorporate gene flow after divergence in its model. Using similar data and a coalescent modeling approach, Fogell et al. (2013) recovered a somewhat more recent divergence time of 4.63 my, but with confidence intervals that overlapped with the previous estimate (95% CI: 3.17–6.38). Evans et al. (1997) proposed that the two populations split following inundation of the Cape Flats. Fogell et al. (2013) pointed out that marine inundation probably occurred multiple times in the last few million years and that cycles of aridification also likely influenced the costal fynbos habitat, on which *X. gilli* relies. Our finding of gene flow after divergence supports the idea that these populations have been periodically reconnected, allowing exchange of migrants. Therefore, whatever the cause of divergence was, it was demonstrably not a permanent barrier.

Of note is that the evolutionary models we tested are almost certainly violated by the system we explored in many ways, including variation over time and among populations in population size, mutation rate, and migration rate. Although we do not anticipate that these violations are influential enough as to negate the rejection of the isolation model, a larger dataset might provide statistical power with which to better evaluate more complex scenarios, such as the secondary contact model.

The F_{ST} and linear mixed model analyses suggest that allele frequencies have not changed substantially in the last 20 years, though there is a trend of decreasing diversity (Fig. 6.5 and from values obtained in linear mixed models indicated a non-significant decline in diversity from 1994 to 2013). If generation time is about one year or less (which is based on laboratory studies and could be an underestimate; Tinsley and Kobel 1996), this represents 20 generations. Changes in allelic diversity may signal population declines earlier than nucleotide diversity, because loss of rare alleles (which happens during population declines) would have a greater impact on count based metrics, such as allelic diversity, than they would on frequency based metrics such as π (Greenbaum et al. 2014). Thus, though not significant, a declining trend seen for allelic diversity (Fig. 6.5) may be an early indication of population declines. Linear mixed models allowing for independent changes in diversity for each locus over time revealed declining genetic diversity (except for two loci in the π models; results not shown).

6.4.3 Management

Hybridization and introgression has the potential to threaten species survival (Rhymer and Simberloff 1996). In an attempt to reduce gene flow between species, three conservation actions were implemented in the mid-1980s. A wall was built around one impoundment in CoGH (Picker and Villiers 1989), populations of pure *X. gilli* were translocated to areas without *X. laevis* (Measey et al. 2011), and *X. laevis* were manually removed from CoGH (Picker and Villiers 1989; Villiers et al. 2016; Measey et al. 2017). Removal of *X. laevis* ceased in 2000, but resumed in 2011 (Picker and Villiers 1989; Villiers et al. 2016; Measey et al. 2017). The same management efforts have not

been conducted for the population of *X. gilli* east of False Bay, most of which resides in non-protected areas (Fogell et al. 2013; Villiers et al. 2016).

Interestingly, the CoGH has greater juvenile recruitment of *X. gilli* (Villiers et al. 2016) and fewer hybrids (2.5% vs 8–27% of individuals in ponds in the west and east respectively; Fogell et al. 2013). Our results suggest that these hybrids are not producing successful offspring via backcrossing with either parental species frequently enough to produce large scale genomic impacts. These results suggest that the genomes of *X. gilli* and *X. laevis* are largely genetically distinct. Thus, the major benefits to *X. gilli* of removal of *X. laevis* from habitat shared with *X. gilli* probably stem from minimizing competition for ecological resources between these species (Picker and Villiers 1989; Villiers et al. 2016; Vogt et al. 2017).

For *X. gilli* and *X. laevis*, east populations harbor more genetic diversity than the west populations (Fig. 6.5. Allelic diversity and heterozygosity reflect a population's ability to respond to selection (Caballero and García-Dorado 2013), and thus from a genetic perspective conservation of east populations of *X. gilli* is paramount.

This study suggests that patterns of gene flow within *X. gilli* included genetic exchange between populations in the east and west. The ancestral distribution of *X. gilli* was likely patchy to begin with and has contracted considerably in the last several decades, including in locations now occupied only by *X. laevis* or *X. laevis* and hybrids (Picker and Villiers 1989; Fogell et al. 2013). Ancestral patterns of gene flow are presumably imperiled by further habitat fragmented by human activity, including habitat altering effects of invasive species such as *Acacia saligna* (Port Jackson Willow), *Acacia mearnsii* (Black Wattle), and *Hakea sericea* (Silky Hakea) (Wilson et al. 2014). Continued efforts to conserve and restore coastal fynbos habitat both inside and outside of protected areas (Villiers 2004), such as removal of invasive vegetation, restoration of native vegetation, and removal of *X. laevis*, stands to benefit *X. gilli*. This is particularly important in the east population of *X. gilli* near Kleinmond where genetic diversity is highest and the population resides on private land.

Electronic supplementary material

Supplementary Information accompanies this paper at doi:10.1038/s41598-017-01104-9 and is available in Appendix E.

Accession Codes: Alignment for each locus, along with the corresponding BEAST XML files, resulting gene trees, and data for Structure analyses have been deposited in a Dryad repository (doi:10.5061/dryad.g6g2r). Genbank accession number for the sequences include the following:

Rassf10: HQ221332–HQ221356, KY824194–KY824236, Sugp2: HQ221211–HQ221235, KY824150–KY824193, c7orf25: HQ220710–HQ220732, KY824433–KY824476, fem1c: HQ221309–HQ221331, KY824395–KY824432, mastl: HQ221046–HQ221070, KY824354–KY824394, mogs: HQ221071–HQ221092, KY824311–KY824353, nfil3: HQ221093–HQ221120, KP344016, KY852029–KY852072, BTBD6/p7e4: HQ220685–HQ220709, KY824477–KY824517, pigo: HQ221144–HQ221167, KY824237–KY824280, prmt6: HQ221168–HQ221178, HQ221180–HQ221190, zbed4: HQ221262–HQ221284, KY824107–KY824149, bcl9: KP345721, KP345621, KY851962–KY852028, pcdh1: HQ221129, KY824281–KY824310, 16s: KP345307, KP345309–KP345313, KP345315–KP345318, KY852073–KY852133

6.5 Acknowledgments

We thank Mike Picker for collaboration on early fieldwork, advice on *X. gilli* and for initiating a long-standing legacy of conservation of this species. We also thank Brian Golding for access to computational resources, Jonathan Dushoff for statistical advice, and two anonymous reviewers for comments on an earlier version of this manuscript. We thank the staff at the CoGH for the efforts in preserving *X. gilli*, both genetically and its habitat. This work was supported by support to B.J.E. from the Natural Science and Engineering Research Council of Canada (RGPIN/283102-2012) and the Museum of Comparative Zoology, Harvard University, G.J.M was funded by the National Research Foundation of South Africa (NRF Grant No. 87759) and the DST-NRF Centre of Excellence for Invasion Biology at Stellenbosch University. Research permission came from the Chief Directorate of Nature Conservation and Museums (2009/94), CapeNature (AAA007-01867), and SANParks.

Part IV

Conclusion

Chapter 7

Conclusions

More is known about *Xenopus* genomics than any other amphibian. Despite that, there is still a great deal more left to discover. This thesis asks questions focusing on basic elements of genome biology, along with questions focused more on species dynamics and the basics of gene flow between individual *Xenopus* lineages using genetics. This work has advanced our understanding of sex chromosomes in this group, the influence of the individual diploid ancestors on subgenome evolution, species boundaries, and the genetic threat to endangered *Xenopus* species.

The chapters on sex chromosomes (chapters 2,3) indicate that those of *X. borealis* contain a number of genes homologous to other, independently evolved sex chromosome systems. This observation supports recent proposals that only some regions of the genome are well suited to act as sex chromosomes, and will be continually co-opted to act as the sex chromosomes (Graves and Peichel 2010; O'Meally et al. 2012). On the other hand, the comparison of the two subgenera *Xenopus* sex chromosome systems (those of *X. laevis* and *X. borealis*), reveals that the evolutionary path sex chromosomes take is unpredictable. The long held belief that recombination suppression is progressive and inevitable clearly does not hold in this group, where the younger sex chromosomes of *X. borealis* have rapidly established widespread recombination suppression. One of the interesting findings of these investigations, in need of further exploration, is the seemingly suppressed recombination on the homeologous chromosome of the sex chromosome. This autosome shares a great deal of sequence with the sex chromosome, and suppressed recombination on this chromosome may indicate that sequence based recombination modifiers may play a large role in recombination suppression on sex chromosomes. In this case, whatever is targeting the sex chromosomes for recombination suppression may be binding to off target sequence on this autosome. Ongoing investigations are assessing signals of sexually antagonistic selection on the new sex chromosomes of *X. borealis*, as well as cytological work is being performed to determine if inversions play a role in recombination suppression.

Also present here is a subgenus *Xenopus*-wide investigation into the dynamics of genome evolution post-allopolyploidization that leverages the ability to separate the two

halves of the genome contributed by the diploid ancestors (chapter 4). We demonstrated that the higher loss of S-subgenome genes, found by the genome sequencing effort of *X. laevis* (Session et al. 2016), is a subgenus wide phenomena. We also determined that differences in the degree of purifying selection between the S and L subgenomes were likely present at the time of the whole genome duplication event (and if not, rapidly established after), and have persisted until present day. This chapter thus provides an animal perspective on allopolyploid evolution, and supports the notion of biased fractionation, most heavily reported in the plant literature.

The chapter redefining the species boundaries of *X. laevis* has provided the genetic history context for future studies that may see differences among individuals responses (like Du Preez et al. 2009; chapter 5). As well, the chapter pointed to an interesting genetic history of *X. laevis*, such as the differences between *DM-W* and mitochondrial markers, which both should be strictly maternally inherited markers, and the reasons behind the excessive genetic diversity found in the southern part of the species range. Future work will hopefully explore the contributing factors to these oddities further. And finally, this thesis demonstrated that the endangered species *X. gilli* does not show signs of genomically widespread introgression (although whole genome sequencing is needed to firmly establish this; chapter 6). Instead, the primary threat to this species by *X. laevis* may instead be through ecological competition.

Together these chapters demonstrate that *Xenopus* species are useful to address a wide array of questions important to the field of evolutionary genomics, and will remain useful models to help scientists understand the diversity of life around us.

Part V

Appendix: Sex Chromosomes

Appendix A

Supplemental Information: Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination

A1 Supplemental Methods

A1.1 Distinguishing orthologous and homeologous sequences

Our phylogenetic analyses involved tetraploid species, and it was crucial to distinguish orthologous sequences (those that diverged from one another due to speciation events) from homeologous sequences (those that diverged from one another due to genome duplication). To accomplish this, we used a phylogenetic approach. Because genome duplication preceded speciation of the tetraploid species in our ingroup, orthologs of the different species are more closely related to one another than to homeologs of any of the species. In other words, a species will have a gene sequence that is more closely related to the gene sequence of another species, than the duplicate copy of that same gene within its own genome. Thus, we retained for analysis only those genes that had two deeply diverged lineages in at least one species (corresponding to the two sets of homeologs for that species) and assumed relationships within each of these lineages were orthologous.

To generate sequence alignments, we extracted putative coding sequence for every gene in each transcriptome assembly using Get_orfs_or_cds.py, which is part of the Galaxy Tool Shed (Cock et al. 2009; Blankenberg et al. 2014). We then used a modified reciprocal BLAST (Altschul et al. 1997b) approach to collect homologous sequences. We BLASTed each transcriptome assembly and the *Xenopus Xenopus laevis* Unigene database against the *X. tropicalis* Unigene database, and saved the top hit below a threshold *e-value* ($\leq e^{-}10$). We then blasted the *X. tropicalis* Unigene database back against each transcriptome assembly, and saved all hits below this threshold *e-value* (resulting in 15,109 *X. tropicalis* sequences with matching transcriptome sequence from

the other species). We retained multiple sequences to ensure that we would capture both homeologs in the tetraploids when present, although this then required us to filter these data based on phylogenetic relationships to remove redundant sequences that were generated during transciptome assembly (see below). We retained for analysis only those BLAST matches that had at least three ingroup species with at least one species that had two sequences present (i.e., potential homeologs), leaving 7,794 sets of homologous sequences. Using a Perl script, we then generated individual sequence files for each best-BLAST result. Each file had one *X. tropicalis* ortholog and sequences from each tetraploid transcriptome matching that ortholog. We used MAFFT v.7 (Katoh and Standley 2013) to align each set of homologous sequences.

To ensure that we had enough data to make robust phylogenetic conclusions, we filtered the data sets to those that had at least 300 base pairs (bp) of ungapped alignment. However, some files did not meet this requirement because some sequences were short. Thus, we used a Perl script that, for alignments with less than 300 bp of ungapped alignment, would test for subsets of the data that a) match the previous taxon requirement (three ingroup + at least one species with two sequences), and b) meet the bp requirement of 300 ungapped sites. This script began by trying all combinations of taxa in an alignment that was one less than the total number of sequences (e.g. if 10 sequences in the total alignment, then all combinations of nine taxa would be tested). If none of these combinations met the taxa requirement, the file was discarded. If none of the combinations met the bp requirement (but could meet the taxa requirement), then all combinations of taxa that were two less than the total number of sequences were inspected (e.g. all combinations of eight taxa in the above example). The script would continue in this fashion, testing smaller and smaller combinations of taxa, until either both requirements were met (at which point a new alignment file would be generated with the particular combination of taxa), or the requirements could not be met and the file was discarded. The only exception to this execution were files with >15 sequences present. For these files only the total number of sequences minus one, and, if necessary, the total number minus two were tested; further combinations were deemed too computationally intensive to explore. If these files did not meet the requirements by the total number of sequences minus two, then they were discarded. If, at a given combination of taxa, multiple sets met the requirements, the script would select the combination that produced the longest ungapped alignment length. By doing this, we were able to salvage datasets that did not initially meet the bp requirement. This process left us with 3,781 gene alignments.

In addition to potential homeologs and orthologs, these alignments frequently included multiple slightly diverged but closely related sequences, which probably stemmed from allelic differences, sequence errors, misassembled transcripts, segmental duplications, and splice variants. For each alignment, we therefore used a phylogenetic approach to select representative orthologous and homeologous sequences. First, maximum likelihood trees were constructed using RAxML v.8.0.25 (Stamatakis 2014a) using the GTRGAMMA model for all alignments, setting *X. tropicalis* as the outgroup, and with other parameters at the default settings. We did not perform model testing for this step

because it was performed on each of two downstream phylogenetic analyses for alignments that passed an initial filter. The resulting phylogenies were then parsed using a script written in R (R Core Team 2017) and functions available in the ape (Paradis et al. 2004), phytools (Revell 2012), and phangorn (Schliep 2011) libraries (available with the Dryad repository doi:10.5061/dryad.00db7). This R script checked whether a tree had only two deeply diverged (putatively homeologous) sequences for at least one ingroup species. The script did this by comparing the age of the most recent common ancestors (MRCAs) of sequences within and between species; homeologs were expected to have a deeper MRCA than that of the orthologs. By identifying and comparing the MRCA of pairwise comparisons within and between species, we identified gene alignments in which there were only two lineages of sequences within the ingroup and where at least one species had one (putatively homeologous) sequence in each lineage. We then assumed that these homeologs were generated by one whole genome duplication (WGD) event prior to radiation of extant *Xenopus* species.

After building the maximum likelihood trees and filtering with the R script, we were left with 1,644 alignments with two deeply diverged *Xenopus* lineages for at least one species that we assumed were homeologous. We then built chronograms with each of these gene alignments using BEAST (Drummond et al. 2012). Each alignment was tested for a model of evolution using MRMODELTEST v.2 (Nylander 2004), and one MCMC chain was run for 25 million generations, using an estimated strict clock. We set the age of the root to 65 million years (my), with a standard deviation of 4.62 following (Bewick et al. 2012). Input files were prepared for BEAST using a Perl script. After inspecting the posterior distributions of parameter values from all analyses to ensure that stationarity had been reached for parameter estimates, we applied a conservative 25% burnin to all analyses and generated consensus trees using TREEANNOTATOR (part of the BEAST package). We then reinspected the trees with our R script, to confirm that there was still two deeply diverged lineages and identifiable homeologs, which left 1,600 alignments. At this point, because our best BLAST approach retained all matching sequences below a threshold e-value, some orthologs within each homeologous lineage were still represented by multiple sequences. We then selected the longest sequence for each species within each homeologous lineage using a Perl script. We then rebuilt the BEAST trees using two chains run for 75 million generations (substitution models selected as before) and inspected all posterior distributions for convergence of parameter estimates (using the R package mcmcse; Flegal et al. 2016), and re-inspected the trees with our R script to ensure homeologs were still present.

The final data set consisted of 1,585 genes, with between five and 10 taxa (each ingroup species had two, one, or zero sequences in each alignment, depending on whether there was missing data from neither, one, or both homeologs). The individual gene alignments were 382–13,911 bp (first and third quartiles = 978–1,981 bp). The total aligned length of the data set had 2,696,030 bp, and an ungapped alignment had a total length of 788,627 bp. The proportion of missing genes for each homeolog is *X. borealis* = 0.55/0.58, *X. clivii* = 0.64/0.63, *X. allofraseri* = 0.52/0.51, *X. largeni* = 0.56/0.58, *X. laevis* = 0.14/0.14.

A2 Supplemental Results and Discussion

A2.1 Multigene Phylogenetic Analyses of Nuclear DNA

Hereafter we refer to each set of orthologous sequences within subgenus *Xenopus* as the alpha and beta orthologs. The assignment of a particular set of orthologous sequences to each of these categories was arbitrary. The chronograms and maximum likelihood trees recovered from both concatenated data sets and the phylogeny recovered from the MPEST analysis all strongly supported a sister relationship of *X. borealis* and *X. clivii* in both the alpha and beta orthologs (Fig. A2.2; maximum likelihood results not shown). Similarly, the multigene *BEAST analysis also recovered strong support for this sister relationship (Fig. 2.1). This relationship had posterior/bootstrap support (chronograms and MPEST, respectively) of 1.0/100% in all analyses, which, in the case of the BEAST and RAxML analyses, may in part reflect overconfidence associated with analysis of concatenated data (Kubatko and Degnan 2007).

The relationships among *X. laevis*, *X. allofraseri*, and *X. largeni*, remain unresolved in all analyses. In fact, there was even conflict in the resolution of these three taxa between the alpha and beta orthologs within the same analysis (Fig. A2.2). For instance, in the concatenated BEAST analysis with all data, in the alpha orthologs supported a sister relationship of *X. largeni* and *X. laevis*, while the beta orthologs supported *X. largeni* and *X. allofraseri* as sister (Fig. A2.2). For the concatenated BEAST analysis with gapped sites removed, the alpha orthologs had the same resolution as the alpha orthologs in the full data set, but the beta orthologs supported *X. laevis* and *X. allofraseri* as sister taxa. For both concatenated chronograms, these various resolutions had posterior support 1.0 and bootstrap values >80% (and up to 100%). The MPEST analysis had the same results as the concatenated analysis with gapped sites removed, with alpha orthologs supporting *X. laevis* and *X. largeni* (93%) and beta orthologs supporting *X. laevis* and *X. allofraseri* (55%; Fig. A2.2). The maximum likelihood analyses also had different resolutions between the alpha and beta orthologs (results not shown). The *BEAST analysis also did not confidently resolve relationships among these three taxa, having only 0.48 posterior support for a sister relationship of *X. laevis* and *X. largeni* (Fig. 2.1).

A2.2 Phylogenetic discordance of individual gene trees

Using an R function, sister clade relationships were counted across the concatenated post burnin posterior distribution of each individual gene analysis (1,585 alignments; Analysis (i) in Methods). A sister relationship was counted only when at least one additional orthologous sequence was also present (to avoid inferring support when only two species were present and limit the search to topologically interesting relationships). Additionally, we collected the age of divergence for any inferred sister relationships for each tree in the posterior distribution of each gene tree analysis.

This analysis revealed considerable discordance among gene trees (Table A2.3). The strongest support for any clade was a monophyletic group of *X. borealis* and *X. clivii*, found in 52% and 47% of the alpha and beta orthologs, respectively, across the combined post-burnin posterior distribution of trees (with each gene having the same number of trees in the combined posterior distribution). The mean node age this group was 19.79 and 21.19 my for alpha and beta, respectively, and a large range (Table A2.3). A similar number of trees in the posterior distribution placed either *X. borealis* or *X. clivii* as the earliest branching $4x$=36 tetraploid *Xenopus* species. Topologies that did not support *X. borealis* and *X. clivii* as sister taxa, generally estimated an older age of divergence (Table A2.3), which is consistent with the hypothesis that incomplete lineage sorting contributes to the gene tree discordance. No strongly supported resolution of relationships within the clade containing *X. laevis*, *X. allofraseri*, and *X. largeni* was recovered across the gene trees, with each of the three possible sister relationships being nearly equally supported, and with similar divergence times of about 13 my (Table A2.3). Less than 10% of the trees in the combined posterior distribution of trees supported alternate topologies that lacked the clade containing *X. laevis*, *X. allofraseri*, and *X. largeni* and the clade containing *X. borealis* and *X. clivii*.

A2.3 The sex determining region of *X. laevis* is not homologous to that of *X. borealis*.

The Genotype by Sequencing (GBS) results indicated that the sex determining regions of *X. borealis* and *X. laevis* are non-homologous. To ensure that the region containing *DM-W* is not sex linked in *X. borealis*, we amplified a gene linked to *DM-W* (as previous attempts to amplify *DM-W* in *X. borealis* have been unsuccessful; Bewick et al. 2011). The gene *RAB6A* has one homolog located physically close to *DM-W* (Uno et al. 2013). We designed primers for both homeologs using the *X. laevis* genome v7.1 and amplified portions of both homeologs in our *X. laevis* cross. Molecular polymorphism in the homeolog of *RAB6A*, located on the unplaced scaffold 68,908, exhibited a pattern of inheritance that was consistent with sex linkage in the *X. laevis* family, with an insertion-deletion mutation in the mother inherited by all daughters and no sons (11 daughters, seven sons; Fig. A2.4). Conversely, parental polymorphisms in both homeologs of *RAB6A* in the *X. borealis* cross did not exhibit sex-linked inheritance; a maternal SNP on the *X. borealis* ortholog of *X. laevis* scaffold 68,908 copy was shared among daughters and sons including seven out of 11 daughters and four out of 10 sons (the father appeared to carry a null allele but, similar to *AR*, this did not affect out ability to determine a lack of sex linkage). A maternal SNP on the *X. borealis* ortholog of the other *RAB6A* homeolog in *X. laevis* (located on scaffold 19,8991) was detected in three out of six daughters and four out of eight sons, indicating that this *RAB6A* homeolog is not sex-linked. A neighbor-joining tree of these sequences confirmed the orthology and homeology of these sequences, with one homeolog in *X. borealis* grouping with the sex-linked *RAB6A* sequences of *X. laevis*, and the other *X. borealis* sequences forming their own clade.

Supplemental Figures and Tables

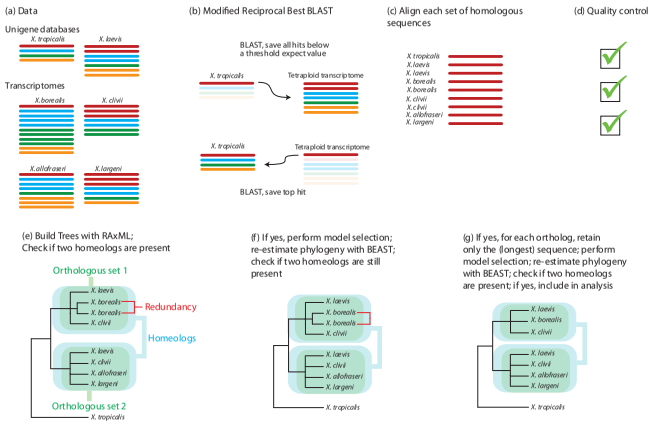


FIGURE A2.1: Our bioinformatics pipeline for identifying orthologous sequences used (a) Unigene and *de novo* transcriptome assemblies and (b) a modified reciprocal best BLAST hit approach to generate (c) sets of homologous sequences which were subjected to (d) quality control to ensure ungapped alignment length >300 bp, at least three ingroup species present, with at least one ingroup species with two (possibly homeologous) sequences present. We then used (e) RAxML to estimate a preliminary phylogeny from several thousand alignments. Phylogenies were parsed for homeologs and if present (f) model selection and BEAST analysis was performed. If homeologs were still present, the longest sequence from each ortholog was retained, and (g) another model selection and BEAST analysis was performed. If two homeologs were still present the alignment was included in downstream analyses. In (a-c) colors represent different genes; paralogs have the same color. In (a-f), redundancy includes allelic variants, splice variants, non-overlapping and overlapping gene fragments, and assembly errors. For some genes, one or both homeologs were not sequenced for some individuals.

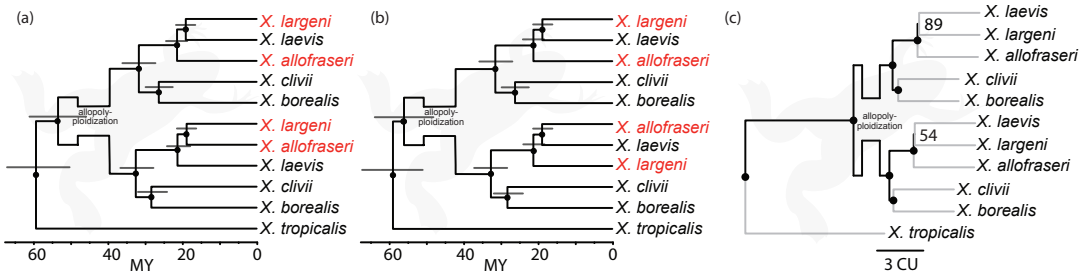


FIGURE A2.2: Analysis of nuclear data using BEAST with either (a) all gene alignments concatenated together or (b) concatenated gene alignment with gapped sites removed. Individual nuclear gene trees were also analyzed with (c) MPEST as described in the methods. In (c), grey lineages have arbitrary branch lengths, CU indicates coalescent units, and numbers indicate bootstrap support. Taxa with conflicting phylogenetic placement within each homeologous lineage are highlighted in red; other labeling follows Fig. 2.1

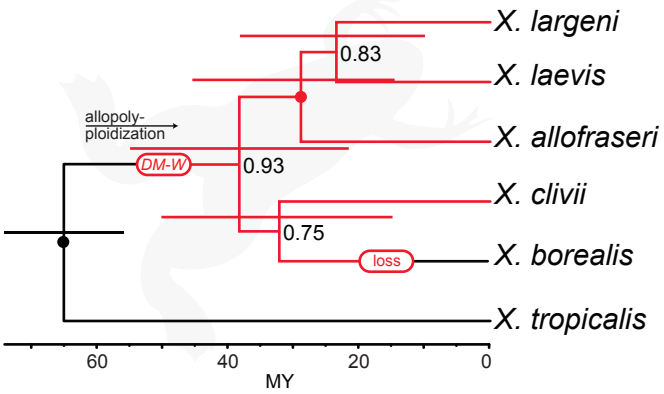


FIGURE A2.3: Bayesian analysis of mtDNA alignments (after removing poorly aligned regions, see methods) with a relaxed molecular clock. This analysis produced an identical topology to that of the *BEAST analysis using nDNA (Fig. 2.1). Labeling follows Fig. 2.1.

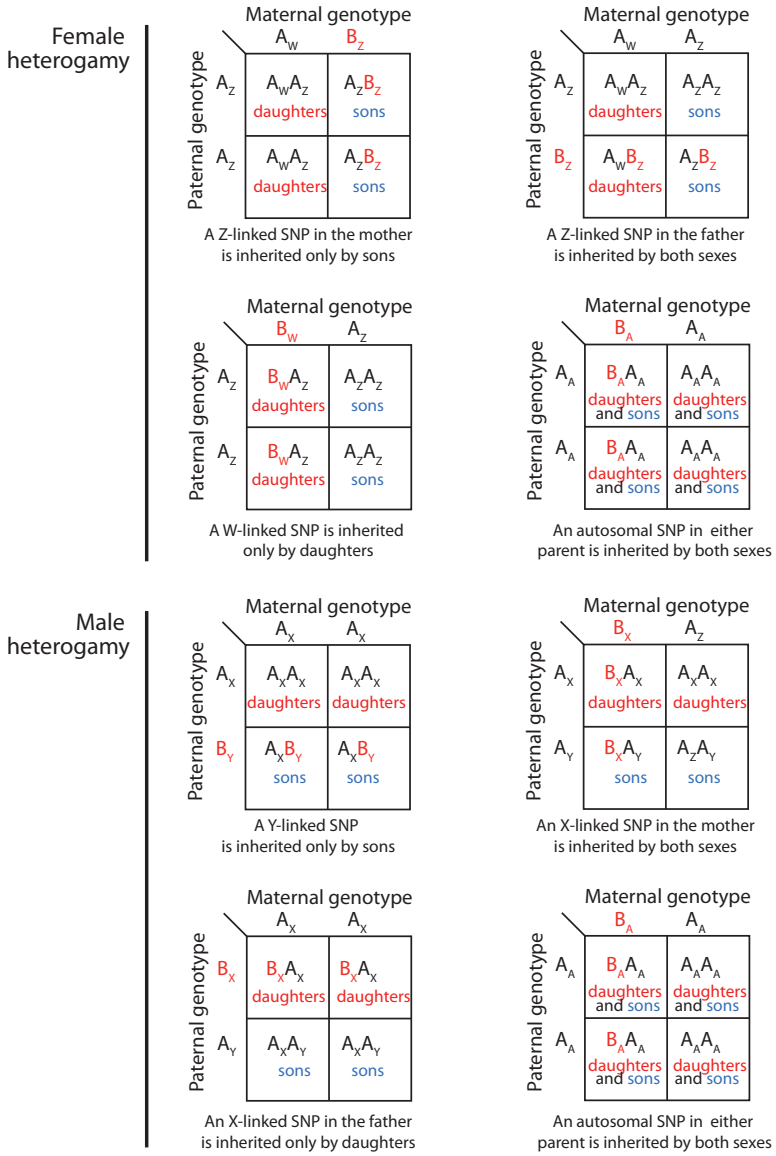


FIGURE A2.4: Not all SNPs are informative with respect to male versus female heterogamy. Diagnosis of female heterogamy requires a sex-linked SNP in the mother, and diagnosis of male heterogamy requires a sex-linked SNP in the father. For each parent, genotypes include nucleotides that are found in both parents (A) or only one (B) and that are linked to the W, Z, X, or Y chromosomes (W, Z, X, or Y subscripts respectively) or an autosome (A subscript)

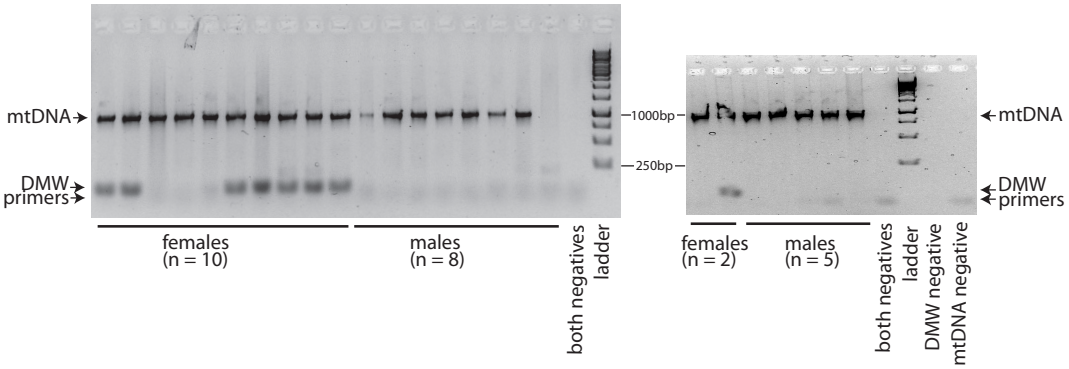


FIGURE A2.5: Attempts to amplify *DM-W* in wild *X. clivii* were only successful in females. mtDNA was used as a positive control and failed to amplify in one male.

TABLE A2.1: Transcriptome and GBS sequencing statistics. Raw sequence is number of reads; Trimmomatic is number of reads that passed our filter; Trinity is number of number of transcripts (N50 bp in brackets); Tassel values reflect reads for both lanes of sequencing (merged) and the total number of SNPs.

Species	Raw Sequence	Trimmomatic	Trinity	Tassel
X. allofraseri	20213893	18940455	96832 (1176)	-
X. borealis	19593759	19296751	81696 (1078)	-
X. largeni	19512126	19258973	82695 (1000)	-
X. clivii	18056373	17441776	72019 (885)	-
GBS	652154864	-	-	88996

TABLE A2.2: The primers used throughout the paper. Reverse primers provided in reverse compliment of the aligned sequence. The * denotes the additional primer combination used in the wild population samples (in text referred to as the alternate set of primers).

Gene	Primer name	Direction	Sequence
AR	begin	forward	ATGGCGGTGCACATAGGG
	down2	reverse	CGGGGGTCTCTTCGCTCT
SOX3	beta_for3	forward	GGTTTGGTCCCGGGGGCAGCGC
	beta_rev1	reverse	CTGAAGGGAAGAATGGTCGCC
	HiSeq_For5	forward	TAGGGAAGTTTTGTGCCGGGA
	HiSeq_Rev1	reverse	ACTCTGAAGGGAWGAGGGGTCG
	HiSeq_Rev5	internal reverse	GGGGTGGAAGCATTGCCCTTA
FMR1	alpha_f1_int	forward	TTTGTATTATGGTCTTTTCAGGTGTATT
	alpha_r1	reverse	TTATGAATGAGCTTTTTTGTTGCTGG
	alpha_f2*	forward	TGCCAACATACAACAGGCTAGGAAG
	alpha_r3*	reverse	CTAAGTTTCGTGGAACTTGTATAACATT
X. clivii DM-W	clivii_for2	forward	AATGAGGAACCATACAGCCCCAGGC
	clivii_rev2	reverse	GATTTCTGCATCGGGCACACCG
RAB6A	RAB6A_alpha_exon9_for2	forward	GCTCCTGTTAATGGCGCCCGTC
	RAB6A_alpha_exon9_rev1	reverse	CTGCTTATATATTAACAAGCCT
RAB6A	RAB6A_beta_exon9_for2	forward	GCTCCTGTTAATGTCGCCCATG
	RAB6A_beta_exon9_rev1	reverse	CTGCTTATATATTAACAAGCCC

TABLE A2.3: Conditional support and median ages with the 95% bounds for various sister relationships clades summarized across the combined post-burnin posterior distributions of individual gene tree analyses (Analysis (i), methods). The inferred sister relationship is dependent on the presence of the "condition" taxa being present (i.e. a sister relationship of *X. laevis* and *X. allofraseri* would be counted only if *X. largeni* was also in the alignment). Alpha and beta referred to the individual homeologous lineages generated by the WGD that preceded the speciation of extant *Xenopus* 4x=36 tetraploids.

homeolog	clade	condition	support	present	proportion	age	0.025	0.975
alpha	(*X. largeni*, *X. laevis*)	*X. allofraseri*	2769633	7320610	0.38	15.18	6.11	31.20
	(*X. allofraseri*, *X. laevis*)	*X. largeni*	2569313	7320610	0.35	15.10	4.46	28.47
	(*X. largeni*,*X. allofraseri*)	*X. laevis*	1806136	7320610	0.25	14.91	4.69	35.59
	(*X. clivii*, *X. borealis*)	any	3752122	7200600	0.52	22.21	7.37	40.52
	(*X. borealis*, any)	*X. clivii*	1428625	7200600	0.20	34.08	17.40	55.74
	(*X. clivii*, any)	*X. borealis*	1749065	7200600	0.25	33.64	10.54	57.36
beta	(*X. largeni*, *X. laevis*)	*X. allofraseri*	2700763	7344612	0.37	14.26	5.25	29.15
	(*X. allofraseri*, *X. laevis*)	*X. largeni*	2602804	7344612	0.35	14.90	4.87	28.99
	(*X. largeni*,*X. allofraseri*)	*X. laevis*	1899217	7344612	0.26	14.13	3.43	29.58
	(*X. clivii*, *X. borealis*)	any	3020271	6552546	0.46	23.52	6.06	41.69
	(*X. borealis*, any)	*X. clivii*	1602842	6552546	0.24	34.83	14.01	63.30
	(*X. clivii*, any)	*X. borealis*	1638245	6552546	0.25	32.65	8.85	62.56

Appendix B

Supplemental Information: Divergent evolutionary trajectories of two young, homomorphic, and closely related sex chromosome systems

B1 Supplemental Methods & Results

B1.1 Genotyping and filtering of reduced representation genome sequence data

We generated reduced representation genome sequence data from one family each from *Xenopus Xenopus laevis* and *Xenopus Xenopus borealis* as described in the main text. For the *X. borealis* Genotype by Sequencing (GBS) data, we de-multiplexed, filtered, and trimmed for sequencing quality using the processRADTags module of Stacks version 1.30 (Catchen et al. 2011) (see Furman and Evans (2016) for details). For the *X. laevis* restriction site associated DNA sequencing (RADSeq) data, we demultiplexed and trimmed using RadTools (Baxter et al. 2011) followed by Trimmomatic (Bolger et al. 2014), using default parameters. After processing, the average number of reads per *X. borealis* offspring was 5,906,315 (406,950–14,594,650, range), with the mother having 69,647,769 and the father 100,997,934. The average number of reads per *X. laevis* offspring was 6,858,340 (2,010,213–14,611,364), with the mother having 5,434,242 and the father 16,469,889.

We then aligned GBS or RADSeq data from each *X. borealis* or *X. laevis* individual to the *X. laevis* genome version 9.1 (www.xenbase.org) using BWA `mem` v 0.7.8 with default parameters (Li and Durbin 2009). To reduce memory requirements and expedite genotyping of these data, we first concatenated all *X. laevis* scaffolds that were $< 1{,}000$ bp into a "super scaffold", inserting 200 "N" bases between each scaffold when joined; this super scaffold not used for subsequent analysis. We then removed unmapped reads, retained only primary alignments, performed indel realignment of all samples using GATK version 3.4 `IndelRealigner`, and genotyped for each species separately all

individuals together with `UnifiedGenotyper` (McKenna et al. 2010). We first filtered the genotypes as the site level following recommendation in the GATK best practices (QD < 2.0, MQ < 30.0, FS > 60.0, SQR > 4.0, MQRankSum < -12.5, ReadPosRankSum < -8.0; DePristo et al. 2011). Several individuals in the *X. borealis* family had an average coverage across all sites of less than one, and were removed (this minimum coverage of one was arbitrarily chosen, and additional filters were subsequently applied to the remaining higher coverage individuals, see below). This filtering step was not necessary with *X. laevis*, because there were no individuals with very low coverage.

In order to generate a biallelic SNP set for our analyses of genetic linkage and recombination, we implemented a series of genotype filters aimed at retaining only high quality genotypes. For each variable position, we first required that both parents have genotypes, that either one parent was heterozygous and the other was homozygous, or both were heterozygous. Next, any individual genotypes with a read depth < 5 for *X. borealis*, or < 15 for *X. laevis*, or a genotype quality < 20 for individuals of either family, were set to missing. We used a less stringent cutoff for *X. borealis* because coverage was generally lower and mapping to the *X. laevis* reference genome generally poorer for the *X. borealis* GBS data. In order for a variable position to be included in our analyses, we required for both datasets that at least 80% of offspring have genotypes, and that each variable site not violate expectations for Mendelian segregation based on a χ^2 test with significance (P value) below 0.05.

Under-calling of heterozygous sites, where a heterozygous site were incorrectly genotyped as homozygous due to low coverage, is a concern with reduced representation genome sequencing data (Glaubitz et al. 2014; Andrews et al. 2016). One hallmark of this type of error is the observation of incompatible genotypes between parents and offspring (e.g., an A/A offspring genotype from parents with T/A and T/T genotypes). To cope with this, for each variable position, if $< 10\%$ of offspring had incompatible genotypes, we set these genotypes to missing data; if $> 10\%$ of offspring had incompatible genotypes, the site was discarded. Additionally, for the *X. laevis* family only, we eliminated SNPs that mapped to repetitive regions based on the annotation for the *X. laevis* v.9.1 genome sequence. We did not impose this filter for the *X. borealis* family because we had much less mapped data for that cross. We also thinned the *X. laevis* data to only include a maximum of one SNP per RADTag, because dense marker maps with many closely linked markers increased computation times needed to generate linkage maps.

The data sets for each species we studied had different characteristics. The *X. laevis* data included about 500 more maternal SNPs than *X. borealis*, and was trimmed with higher genotype quality and depth thresholds (see Methods). The *X. borealis* dataset was more sparse, with a mean of 61.8 maternal SNPs per chromosome, compared to a mean of 90 in *X. laevis*. The reason for the different number of SNPs is probably a combined consequence of limitations in mapping the *X. borealis* data to the *X. laevis* genome, generally lower coverage of the *X. borealis* GBS data than the *X. laevis* RADSeq data, and possibly a lower overall level of polymorphism in *X. borealis*. Thus, less variation

was recovered in *X. borealis* and there was greater potential for genotyping errors in this species due to lower coverage.

Nonetheless, after filtering, both data sets had high (>20X) depth and high (>50) genotype quality (Fig. B1.4). In general, the sparser panel of SNPs in *X. borealis* is expected to yield poorer coverage of the sex linked region (and of the genome overall). As well, in the sex linked region, genotyping errors could break the association between sex phenotype and genotype, leading to fewer sex linked sites and an underestimation of the size of the sex-linked region. Outside of the sex linked region, genotyping errors could also lead to a false association of sex phenotype and genotype, but for this to happen multiple individuals of the same sex would have to have the same genotype error. In contrast to this possibility, in the *X. borealis* cross, all sex linked sites were associated with sex in either 100% of offspring (with genotyping errors were evident as unlikely phase changes, see Section B1.2), or associated with sex in 45 of the 47 offspring (see Results; Fig. 3.1). This strong association is very unlikely to be a statistical anomaly associated with the lower coverage and sparser SNP density of the *X. borealis* data. In *X. laevis*, there were only a few SNPs associated with sex, and sparser data may have missed this region completely.

Lower sample sizes of individuals could lead to false associations with phenotypic sex due to sampling error of small numbers. Here the sample sizes of both families are similar (37 *X. laevis* and 39 *X. borealis*). The probability of a false positives for an individual 100% sex linked site in *X. borealis* is $0.5^{39} = 1.8 * 10^{-12}$ and in *X. laevis* is $0.5^{37} = 7.3 * 10^{-12}$. And the probability of a series of adjacent false positives 100% sex-linked sites is vastly lower than these values.

B1.2 Haplotype (phase) estimation and genotype error detection

Our methods to estimate parental haplotypes (phase) from mapped offspring genotypes and then use this information to remove putative genotype errors is presented in Fig. B1.1.

B1.3 Biological replication of recombination suppression in *X. borealis* sex chromosomes

To explore whether the sex-linked region we identified here and in Furman and Evans (2016) using GBS data was present in other *X. borealis* individuals, we Sanger sequenced two genes that occurred within the previously identified sex-linked region (*SOX3.L* and *NR5A1.L* (alternatively known as *SF-1*)) in a panel of adults and also in a second lab-generated *X. borealis* family. This allowed us to test whether the sex chromosomes and associated region of suppressed recombination that we observed in one *X. borealis* female were shared by another female. This sequencing included 18 daughters and 12 sons. PCR amplification and Sanger sequencing was performed with primers for *NR5A1.L*

(also known as *SF-1*; AAAAAAGCCTTGATCCGTGCA and AATATGTTTGGCCTGATGTGTA) that were designed using a combination of the *X. laevis* genome, v9.1 from www.xenbase.org and the HiSeq data from our *X. borealis* parents of the initial cross, outlined below. The other gene we amplified was *SOX3.L*, using primers from Furman and Evans (2016). These two genes exist around 10 Mb (*NR5A1.L*) and 35 Mb (*SOX3.L*) from the chromosome tip in the sex-linked region of chromosome 8L, according to their positions in the *X. laevis* genome, spanning a large part of the sex-linked region (Fig. 3.1). To extend this inference across more individuals, we also sequenced these two genes in six females and six males from the same supplier to assess if they contained the same sex specific SNPs as our initial cross (outlined above).

For all parents and offspring of this second *X. borealis* family, allelic variants of both loci were co-inherited by sex, and in the panel of adults we surveyed SNPs were sex associated (females possessed heterozygous sites males did not, and vice versa). We identified one female specific SNP for *NR5A1.L* (*SF-1*) in the newly surveyed adults, and it was shared with our initial *X. borealis* cross and was W-linked (Furman and Evans 2016). For *SOX3.L*, the newly surveyed adults had four sex specific SNPs, two of which were shared with our original cross and were W-linked, and two where females were heterozygous and males were not. Thus, the sex chromosomes and sex-linked region we identified in our original cross is present in multiple *X. borealis* lineages. These shared sex specific sites and co-inheritance of the two genes support chromosome 8L as the sex chromosome in an independent panel of frogs, and is consistent with recombination suppression of this large genomic region in the most recent common ancestor of all *X. borealis* surveyed here. The unique mutations in the W-haplotype of each family indicate that a single W-haplotype has not reached fixation in the species, probably owing to the young age of the system. How long ago these two families diverged, or whether fine-scale population structure exists in *X. borealis* (which could hinder fixation) is currently unknown.

B1.4 Divergence of W and Z in *X. borealis*

We evaluated sex chromosome divergence in *X. borealis* using WGS data from the mother and father of our cross. Shotgun WGS data were generated as described in the main text. We used Trimmomatic (Bolger et al. 2014) and Scythe (`https://github.com/vsbuffalo/scythe`) to clean reads and remove adapter sequences. We mapped the resulting reads to the *X. laevis* v9.1 genome with BWA `mem`, and discarded unmapped reads. We genotyped each parent using SAMTOOLS v1.3.1 `mpileup` (with the multi-allelic model) followed by BCFTOOLS v1.3-27 `call` (Li et al. 2009; Li 2011) and removed individual genotype calls with a depth < 20 or > 60 (with an expected coverage of 30x), and with a genotype quality of < 20 for variable positions. Nucleotide diversity was different for most chromosomes of the female and male, with the female having higher diversity overall (Fig. 3.4a). These inter-individual differences in levels of genome-wide polymorphism may stem from several factors including independent origins from populations of different sizes and differing degrees of inbreeding.

To explore the effect of genome-wide differences in polymorphism, we standardized the estimates of nucleotide diversity within each individual, dividing each estimate by the genome-wide average of each individual. We then ran the same linear mixed model analysis as described in the main text. This model (expectedly) now supported no difference between the two individuals (estimate of scaled female diversity compared to male: -0.002196, -0.0104–0.0061 95% CIs, t-stat = -0.522), but still supported a significant interaction of sex (female and male) and genomic location (sex chromosome and autosomes), with the female having higher diversity in the sex linked region above the male in that region and the rest of the genome (estimate of scaled female diversity in the sex linked region: 0.1175, 0.0571–0.1779, t-stat = 3.815).

B1.5 Synonymous and nonsynonymous divergence of *X. borealis* sex chromosomes

We also explored whether there was a higher rate of synonymous or nonsynonymous substitutions in genes on the non-recombining portion of the sex chromosomes compared to other genomic regions. We calculated dS and dN in the female and male using the gene coordinates for the *X. laevis* genome, as detailed in the gff file associated with v9.1 (www.xenbase.org). This analysis considered only codons with no more than one heterozygous position in order to avoid unknown phase between multiple mutations and assessed, when a SNP in a codon was present, if the two possible codons were synonymous or non-synonymous. To avoid undefined estimates of dN/dS generated by genes with no synonymous mutations, we totaled the number of synonymous SNPs and divided by the total number of synonymous sites for each region (i.e., the non-recombining portion of the sex chromosomes and the rest of the genome), and did the same for the non-synonymous SNPs and sites (similar to Mank et al. 2009). We only retained genes that had at least 200 bp measured and eliminated any genes with a $dS > 2$. We tested for significant differences between the non-recombining region of the sex chromosomes and the rest of the genome using a permutation test and 1000 replicates.

Within coding regions, dN and dS were elevated in the sex-linked region compared to the rest of the genome in the female (dS sex-linked = 0.0279, dS genome = 0.0226, $P = 0.000$; dN sex-linked = 0.0053, dN genome = 0.0041, $P = 0.005$) and in the male (dS sex-linked = 0.0271, dS genome = 0.0224, $P = 0.006$; dN sex-linked = 0.0050, dN genome = 0.0041, $P = 0.026$), suggesting rapid evolution of genes in this region. However, the overall rate of purifying selection, dN/dS, was not significantly elevated in the sex-linked region of either sex (female: dN/dS sex-linked = 0.191, genome = 0.182; male: dN/dS sex-linked = 0.185, genome = 0.184). As stated above, in order to avoid uncertainties in phase of alleles, this analysis only considered codons with one heterozygous site. Consequently, not all SNPs on the sex chromosomes and autosomes were included and this may account for the similarities of the ZZ and ZW values. As we outline in the main text, the sex linked region of chromosome 8L did not have the largest of any metric (dN, dS, or dN/dS) compared to the chromosomes individually. There may not be a signal of faster evolution on the sex chromosomes, or there was too little

data to infer the effect. Either way, it underscores that the W and Z are overall fairly similar chromosomes. We note that our more comprehensive analysis, which considered nucleotide diversity from all variable positions, did detect an elevated level of nucleotide diversity on the ZW pair, above the ZZ chromosomes and the rest of the autosomes (Fig. 3.4a-c).

B1.6 Coverage differences between female and male *X. borealis*

We tested for deletions in the sex-linked portion of the W chromosome in the WGS data, which would be consistent with degeneration of this genomic region. If deletions were present, then only reads from the maternal Z should map to *X. laevis* chromosome 8L, whereas in the father reads from both Z chromosomes should map to chromosome 8L. Thus, the mother should have substantially lower coverage in genomic regions that were deleted on the W. To test for this pattern, we calculated site depth in 100 Kb windows (created with BEDTOOLS v2.26 `makewidows`; Quinlan and Hall 2010) across each chromosome using SAMTOOLS `bedcov` with the same bam files used for genotyping the mother and father HiSeq X data (see Methods). To account for sequencing coverage differences, we then standardized depth estimates for each individual by dividing each window by the median value of coverage of each individual. We searched for windows where the mother had <70% standardized coverage of the read depth of the father. To account for possible mismapping of reads between homeologous sequences, we also reran the analysis with a threshold map quality of 30.

This analysis did not identify a strong signal of deletions on the sex-linked portion of the *X. borealis* W chromosome. Overall, *X. borealis* chromosome 8L had the most (8) 100 Kb windows where the mother had less than 70% coverage of the father, but this was relatively similar in magnitude to the other chromosomes (range = 1–6, mean = 3.5). With the higher map quality threshold of 30, we found that chromosome 8L had the same number of windows as other chromosomes (chr8L = 6, range of autosomes = 1–6). In other words, the standardized coverage of the sex-linked portion of chromosome 8L relative to that of the rest of the genome of *X. borealis* was similar for the female and male.

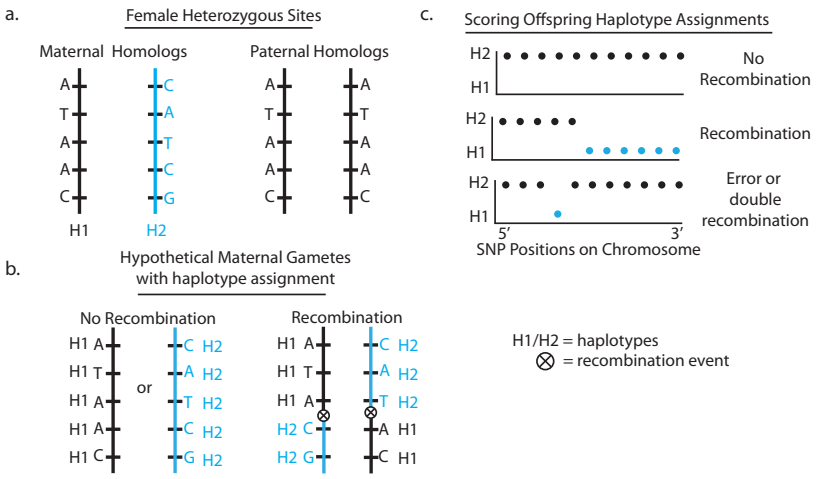


FIGURE B1.1: Using sex-specific linkage maps that were generated from sets of either female or male heterozygous sites, we determined (a) parental haplotypes. (b) A haplotype from each parent could either be inherited entirely (if no recombination happened in the genotyped region), or recombined. (c) Scoring offspring haplotypes for sex specific maps of each chromosome allows for visualization and counting of the number of recombination events and identification of genotyping errors or improbable double recombination events.

TABLE B1.1: Summary of the numbers of heterozygous and homozygous positions involved in double recombination events, as indicated by changes in assigned haplotype at sites relative to the surrounding markers. In the table, "het" refers to heterozygous sites, "hom" refers to homozygous sites. Double recombination events can involve changes of assigned haplotype for one or more markers, reflected in the table by "T". For example, if the parental haplotypes are labeled "0" and "1", a T = 1 is a double recombination event involving a single site relative to surrounding markers (e.g., 010); a T = 2 would be a double recombination event involving two markers, relative to the assigned haplotype of surrounding markers (e.g., 0110; see Fig. B1.1 for a visual explanation). In *X. laevis*, double recombination events involving a single marker overwhelmingly involved a homozygous marker being assigned to the alternate haplotype relative to surrounding markers (in the table: T = 1 hom (het)), indicating that error, in the form of undercalled truly heterozygous positions, was the likely culprit. The columns of "median (mean; count) below & above 5 Mb" reflect the numbers of double recombination events (of T > 1) that spanned either less than or more than 5 Mb of sequence (again, see Fig. B1.1 for a visual explanation). The median and mean reflect the proportion of the genotypes within the phase change that were homozygous (i.e., a 1.0 indicates that all sites in a phase change were homozygous). The count reflects the number of phase changes below/above the 5 Mb threshold. Again, in *X. laevis*, double recombination events spanning less than 5 Mb overwhelmingly involve homozygous markers. Above the 5 Mb threshold, most double recombination events involved a similar number of heterozygous and homozygous sites.

Species	sex	T = 1 hom (het)	T > 1 hom (het)	median (mean; count) < 5Mb	median (mean; count) > 5Mb	Total Genotyped Sites (missing)
X. laevis	female	1254 (101)	106 (13)	1.0 (0.91; 119)	0.55 (0.55; 176)	38973 (6167)
	male	4606 (242)	466 (29)	1.0 (0.96; 495)	0.65 (0.60; 128)	129212 (20601)
X. borealis	female	46 (45)	3 (75)	0.38 (0.42; 26)	0.50 (0.49; 52)	22304 (1837)
	male	60 (46)	6 (7)	0.50 (0.50; 10)	0.25 (0.42; 3)	17750 (1360)

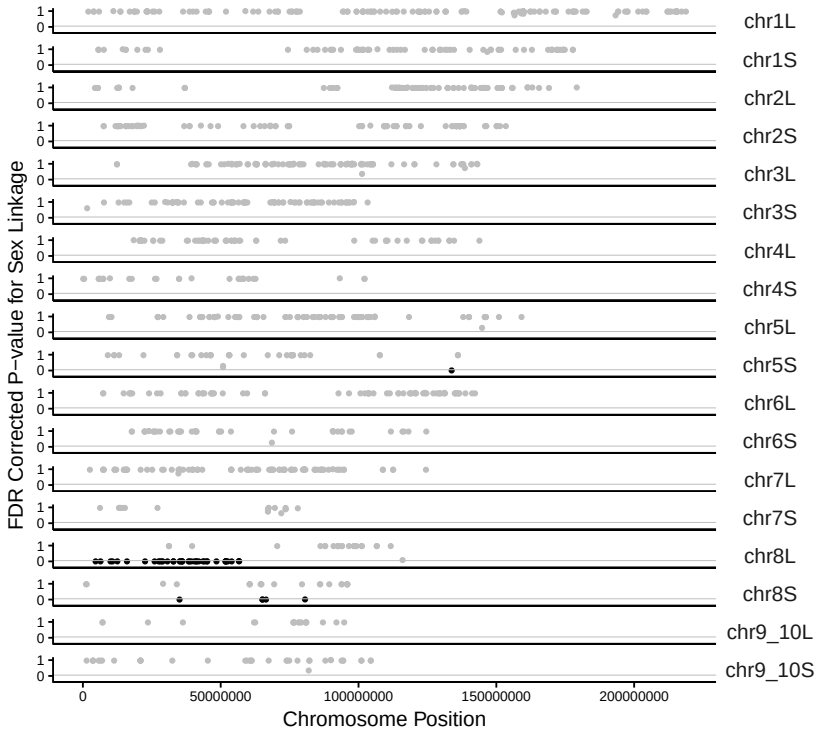


FIGURE B1.2: SNPs from the *X. borealis* family heterozygous in the mother mapped to the genome of *X. laevis*. Sex-linkage is calculated following Goudet et al. (1996), followed by an FDR correction to account for multiple testing (significant genotypes are indicated by black dots). The gray lines in each plot represent the significance threshold of 0.05. The sex-linked marker on chr5S can be mapped to chromosome 8S and 8L when a larger amount of sequence data is used (see Results).

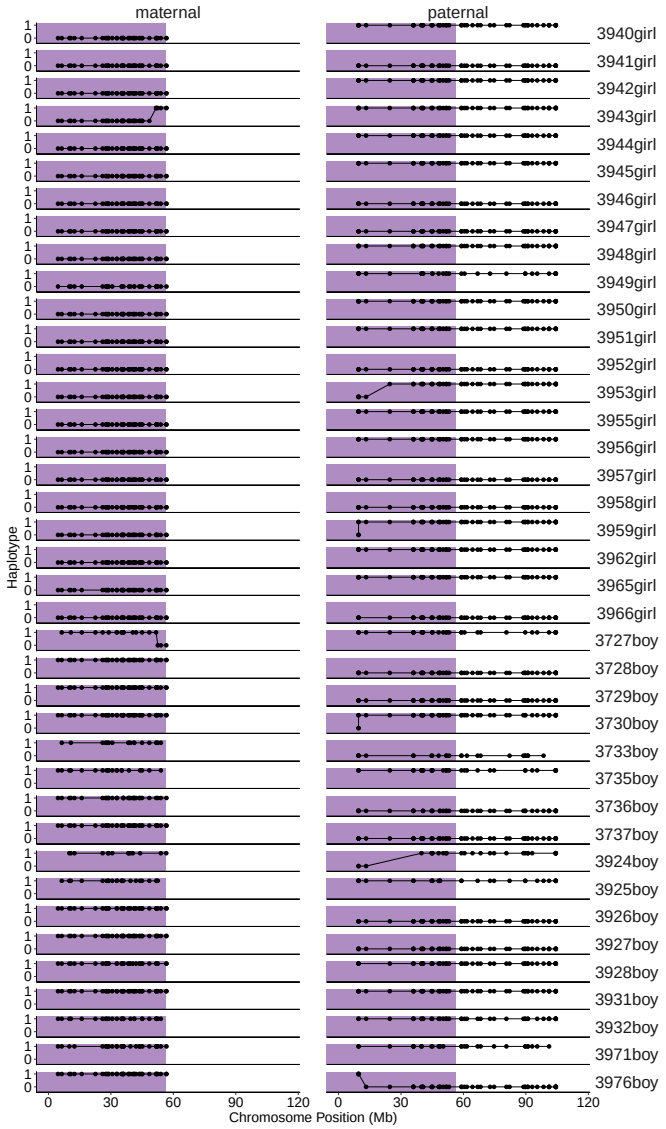


FIGURE B1.3: Phased parental haplotypes in *X. borealis* offspring for the maternal (left column) and paternal (right column) linkage groups of chromosome 8L. In the maternal haplotypes, the beginning of chromosome 8L is completely linked to sex (all daughters with the same haplotype, 0, and all sons had haplotype 1), apart from two recombination events near the end. Note, this linkage map only spans the sex-linked region (see Methods). For the paternal map, haplotypes are evenly shared between the sexes, indicating non-sex-linked inheritance of the paternal Z chromosome. In the region of the paternal Z chromosomes that is homologous to the sex-linked region of the maternal sex chromosomes, several recombination events are observed.

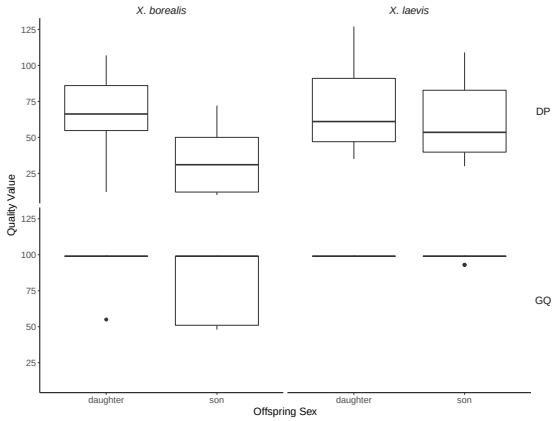


FIGURE B1.4: Median depth (DP) and genotype quality (GQ) for offspring of the *X. borealis* and *X. laevis* families for maternal heterozygous sites used in the sex linkage analysis of Fig. 3.1.

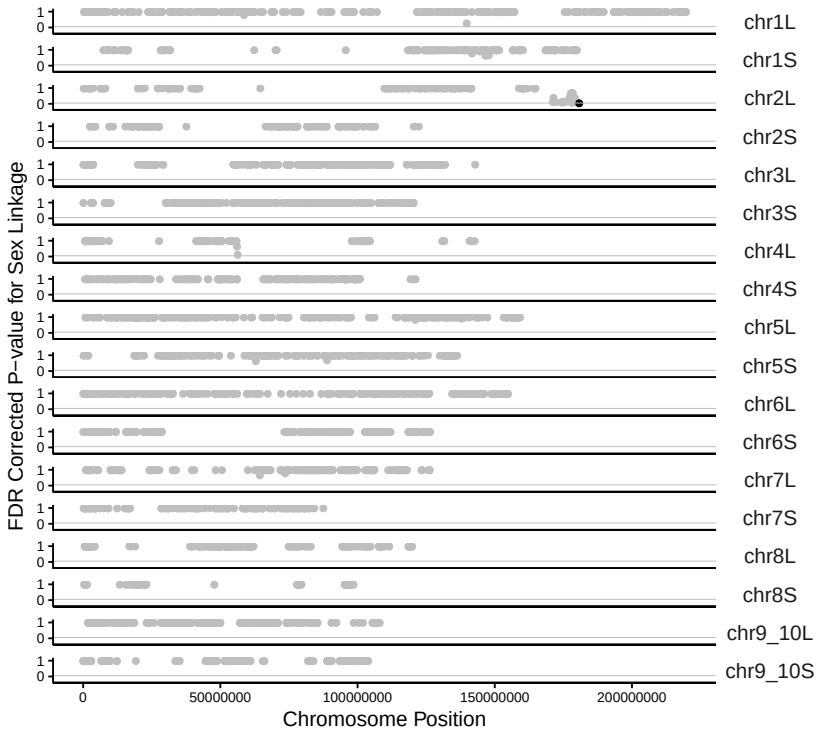


FIGURE B1.5: SNPs from the *X. laevis* family heterozygous in the mother mapped to the genome of *X. laevis*. Sex-linkage is calculated following Goudet et al. (1996), followed by an FDR correction to account for multiple testing (significant at 0.05 colored black). The gray lines in each plot represent the significance threshold of 0.05.

Part VI

Appendix: Whole Genome Duplication

Appendix C

Supplemental Information: Divergent subgenome evolution after allopolyploidization in African clawed frogs (*Xenopus*)

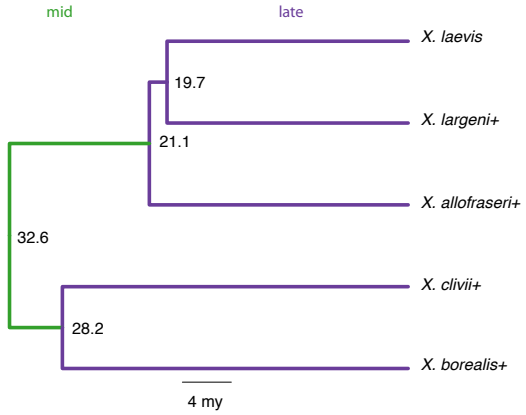


FIGURE C0.1: Chronogram for the *Xenopus* data used in the coding sequence length and pseudogenization rate analyses. The coloring of the branches corresponds to the grouping of branches for the models fit. The + signs indicate that a missing data proportion is fit for the associated taxa. The tree was constructed using *BEAST without a calibration point (see Furman and Evans (2016)), and then dates were assigned using mcmctree, assuming a divergence from *X. tropicalis* of 65 my (Bewick et al. 2012).

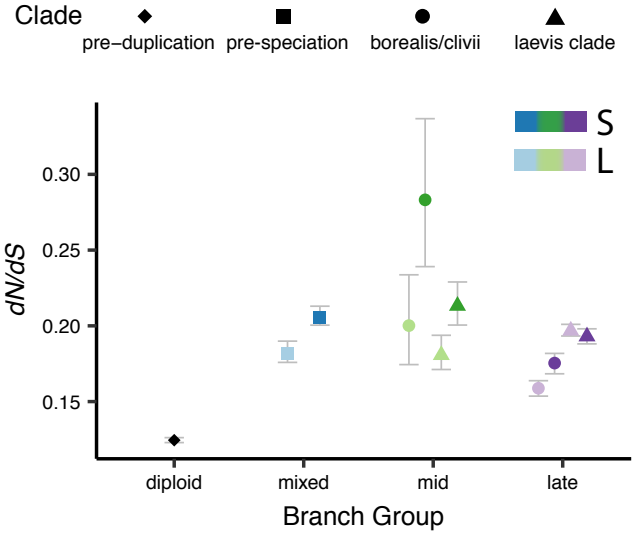


FIGURE C0.2: Results of the similarly fit time-subgenome model of dN/dS rates from the CODEML analysis (see Fig. codeml.models). This model was only 5 BIC units less than the best fit subgenome model. Error bars are 95% confidence intervals based on 100 bootstrap replicates.

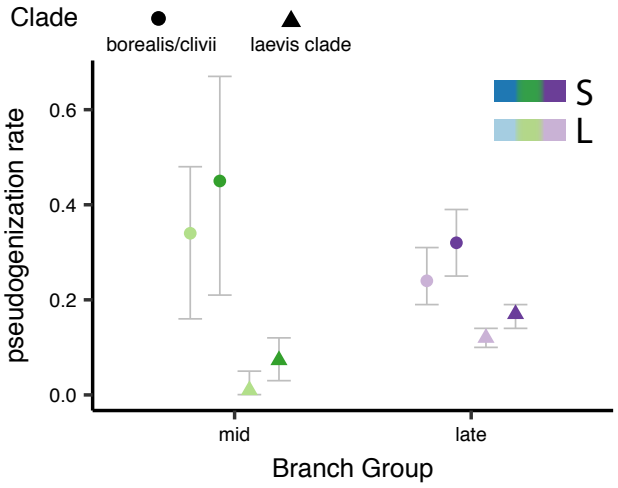


FIGURE C0.3: Results of the similarly fit time-subgenome model from the analysis of pseudogenization rates. Similar to the CODEML analyses, this model was only 2 BIC units less than the preferred subgenome model. Error bars are 95% confidence intervals are based on 1000 bootstrap replicates.

TABLE C0.1: Results of model-based analysis of temporal variation in the rate of pseudogenization. Maximum likelihood estimates (parameter), standard errors (SE), lower (LB) and upper bounds (UB) of the 95% confidence intervals, and the Bayesian information criterion values are tabulated for the four models. Pseudogenization rates are in units of genes pseudogenized per nucleotide substitution (θ, with subscripts described in the Methods section). Lastly, the proportion of missing data (δ) for each species (indicated by subscripts) is also provided. The * indicates the BIC preferred model.

model		parameter	SE	LB	UB
Homogeneous Rates	borealis/clivii clade θ	0.313	0.014	0.289	0.340
	laevis clade θ	0.118	0.008	0.104	0.134
	$\hat{\delta}_{Xenopus\ Xenopus\ borealis}$(*X. borealis*)	0.442	0.015	0.414	0.473
	$\hat{\delta}_{Xenopus\ Xenopus\ clivii}$(*X. clivii*)	0.344	0.015	0.315	0.376
$ln\ell = -7960.025$	$\hat{\delta}_{Xenopus\ Xenopus\ allofraseri}$(*X. allofraseri*)	0.501	0.012	0.479	0.524
BIC = -15963.59	$\hat{\delta}_{Xenopus\ Xenopus\ largeni}$(*X. largeni*)	0.555	0.011	0.535	0.577
Subgenome Model	borealis/clivii clade $\hat{\theta}_S$	0.358	0.017	0.327	0.392
	borealis/clivii clade $\hat{\theta}_L$	0.270	0.015	0.242	0.300
	laevis clade $\hat{\theta}_S$	0.144	0.011	0.123	0.164
	laevis clade $\hat{\theta}_L$	0.096	0.008	0.081	0.112
	$\hat{\delta}_{X.\ borealis}$	0.442	0.015	0.414	0.473
	$\hat{\delta}_{X.\ clivii}$	0.344	0.015	0.316	0.377
$ln\ell = -7937.381$	$\hat{\delta}_{X.\ allofraseri}$	0.501	0.012	0.478	0.523
BIC = -15932.81	$\hat{\delta}_{X.\ largeni}$	0.554	0.011	0.533	0.576
Time Model	borealis/clivii clade $\hat{\theta}_{\text{mid}}$	0.411	0.063	0.292	0.534
	laevis clade $\hat{\theta}_{\text{mid}}$	0.046	0.015	0.018	0.077
	borealis/clivii clade $\hat{\theta}_{\text{late}}$	0.277	0.018	0.243	0.312
	laevis clade $\hat{\theta}_{\text{late}}$	0.139	0.010	0.120	0.159
	$\hat{\delta}_{X.\ borealis}$	0.445	0.015	0.416	0.477
	$\hat{\delta}_{X.\ clivii}$	0.347	0.015	0.319	0.381
$ln\ell = -7949.44$	$\hat{\delta}_{X.\ allofraseri}$	0.494	0.012	0.470	0.517
BIC = -15956.93	$\hat{\delta}_{X.\ largeni}$	0.548	0.011	0.527	0.571
Time-Subgenome Model	borealis/clivii clade $\hat{\theta}_{\text{mid},S}$	0.450	0.123	0.203	0.669
	borealis/clivii clade $\hat{\theta}_{\text{mid},L}$	0.340	0.082	0.165	0.483
	laevis clade $\hat{\theta}_{\text{mid},S}$	0.073	0.025	0.027	0.123
	laevis clade $\hat{\theta}_{\text{mid},L}$	0.010	0.012	0.001	0.046
	borealis/clivii clade $\hat{\theta}_{\text{late},S}$	0.319	0.036	0.255	0.393
	borealis/clivii clade $\hat{\theta}_{\text{late},L}$	0.240	0.030	0.185	0.304
	laevis clade $\hat{\theta}_{\text{late},S}$	0.167	0.014	0.140	0.195
	laevis clade $\hat{\theta}_{\text{late},L}$	0.118	0.011	0.097	0.139
	$\hat{\delta}_{X.\ borealis}$	0.445	0.015	0.416	0.477
	$\hat{\delta}_{X.\ clivii}$	0.347	0.015	0.319	0.380
$ln\ell = -7924.206$	$\hat{\delta}_{X.\ allofraseri}$	0.492	0.012	0.469	0.515
BIC = -15935.49	$\hat{\delta}_{X.\ largeni}$	0.546	0.011	0.525	0.569

C1 Timing of duplication

In order to generate additional estimates of the the maximum age of whole genome duplication to complement previous estimates (e.g., Bewick et al. 2011; Furman and Evans 2016; Session et al. 2016), we evaluated synonymous divergence between homeologs in allotetraploid species from subgenus *Xenopus*, and also with respect to two outgroup sequence that each permitted us to explore the effects of different calibration points. The first ortholog was from humans and was obtained for each gene from the Human Unigene database (downloaded December 2015) and recovered orthologs using tblastx between the *X. tropicalis* sequence of each alignment and the database. We used an estimated divergence of Anurans and Mammals of 350 million years (my) (340–381), which was based on numerous calibration points and mitochondrial DNA (mtDNA) divergence Igawa et al. 2008.

Outgroup sequences were aligned to the nucleotide alignments using `Macse`, and synonymous divergence (dS) between the homeologs, when both were present in an alignment for a given species, and between each of the homeologs and the human ortholog was calculated using `yn00` Yang and Nielsen 2000; Yang 2007. When alignments had more than one species with homeologous sequences, dS was taken as an average across all species for a gene (none had both homeologs for all five species). We then used a linear extrapolation of divergence between the homeologs and an outgroup (human) to infer the age at which the homeologs separated. To calculate this, we divided the median dS of the outgroup ortholog and the homeologs by the divergence time (and both bounds of the 95% CI). We then divided the median dS between the homeologs by the previous value.

The `tblastx` search between *X. tropicalis* sequences of each locus and the Human Unigene database recovered 755 orthologous human sequences. For 16 of these alignments, dS values between homeologs were found to be above one and were excluded. Using a linear extrapolation from the divergence time of amphibians and humans (350 my, 340–381 95% CI), the median age of the homeolog divergence (i.e., speciation of the diploid progenitors of tetraploid *Xenopus* species) is estimated to be 32.5 my (31.6 – 35.4 95% CI). This date is provides independent support for the similar date of 34 my estimated by Session et al. 2016, who used different calibration points and analyses.

C2 Rate of Pseudogenization Simulations

To test whether the proposed models can recover parameters successfully for data similar to the *Xenopus* data on hand, we conducted simulations on the consensus tree (Fig. C0.1) from the *Beast Heled and Drummond 2010 analysis from Furman and Evans 2016. We simulated data using the geiger Harmon et al. 2008 package in R to emulate as closely as possible the collection mechanism for the real *Xenopus* gene membership data.

For consistency with the dataset in the paper, each simulation contains a total of 1500 gene patterns. As well, one taxa out of the five on the tree is treated as the species with complete data, similar to *Xenopus Xenopus laevis* (*X. laevis*) in the real analysis, and we constrain the root of the tree to have two homeologs for each gene family (similar to the requirements of Furman and Evans 2016). We start with an ancestral sequence of ten thousand characters with state (1,1). Using this ancestral root sequence, we simulate characters with states (0,1), (1,0), and (1,1) for all five taxa using `geiger`. Two pseudogenization rates are sampled uniformly from [0.01, 0.5] for one of the lineages, analogous to $\theta_{\text{clade1},L}$ and $\theta_{\text{clade1},S}$). "Tree stretching" (i.e., using a multiplicative factor on the branches of the tree when simulating data but using the original tree in the fitting) is used for simulation on the other lineage in the tree. The ranges of the pseudogenization rates sampled for the other clade ended up being $(0.0004, 0.495)'$ and $(0.001, 0.500)'$. Because there are multiple characters being simulated (corresponding to each of the four θ parameters for the real data analysis), a consequence of using the tree stretching procedure is that the order of the θ parameters for one clade is the same as the order of the θ parameters for the other clade. Hence, if $\theta_{\text{clade1},1}$ is greater (or lower) than $\theta_{\text{clade1},2}$, the same order is simulated for $\theta_{\text{clade2}}, 1$ and $\theta_{\text{clade2}}, 2$. While these two combinations are extensively explored in the simulation, the following combinations are not explored in the simulation: $\theta_{\text{clade1}}, 1 < \theta_{\text{clade1}}, 2$ and $\theta_{\text{clade2}}, 1 > \theta_{\text{clade2}}, 2$; and $\theta_{\text{clade1}}, 1 > \theta_{\text{clade1}}, 2$ and $\theta_{\text{clade2}}, 1 < \theta_{\text{clade2}}, 2$. This is mentioned for completeness regarding data simulation; this is unimportant for model fitting as the model can fit all combinations regardless because separate parameters are estimated and the estimation does not rely on any assumption of any order between any sets of parameters.

Proportions of missing data for four of the five taxa (analogous to the *Xenopus* data) are sampled uniformly from [0, 0.60]. Based on these missing data proportions, a corresponding number of (1,1)s are converted to (0,1), (1,0), and (0,0)s (cf. Table 4.2). Similarly, some (0,1)s and (1,0)s are converted to (0,0)s. We filtered any observed patterns that did not include at least one (1,1) or had three or more (0,0)s out of the dataset. This reflects the sampling bias of the real data as the result of constraints discussed in the Methods section. A total of $1,000$ data samples of 1500 genes patterns are run. Like the main analysis in the paper, we estimated separate pseudogenization rate parameters for the two subgenomes for the two lineages (four rate estimates in total). Thus, our simulated data closely reflects the data in the paper, and our simulations investigate the model's ability to recover known parameters. Fig. C2.4 depicts the differences between the true and estimated parameters for 1,000 simulated data sets with sampling bias similar to the *xenopus* dataset for 1500 phyletic patterns. The estimates are close to the generating parameters on average for all simulations. The range of differences between the true and estimated proportions of missing data is small, with most of the weight concentrated around zero.

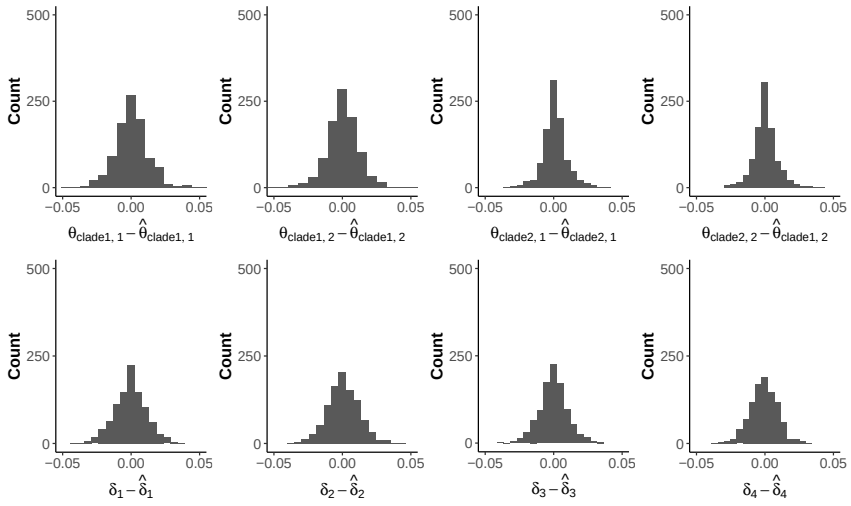


FIGURE C2.4: Histograms of the differences between the generated and estimated rate (row one) and missing proportion (row two) parameters from the simulation. The distribution is centered around zero and the range of the differences is small. For these simulations, cladei,j refers to the j^{th} subgenome for the i^{th} clade. These are analogous to the borealis/clivii and laevis clades (i) and the L- and S-subgenomes (j) in the analysis of the real data.

Part VII

Appendix: Speciation and Hybridization

Appendix D

Supplemental Information: Pan-African phylogeography of a model organism, the African clawed frog *Xenopus laevis*

D1 Supplemental Information

D1.1 Taxonomy

The Amphibian Species of the World database (Frost 2011) and AmphibiaWeb (*Information on Amphibian Biology and Conservation [Web Application]* 2014) list *Xenopus Xenopus laevis sensu lato* as comprising three species: *X. laevis* (the African clawed frog), *X. petersii* (Peters' clawed frog; including *X. poweri* as a junior synonym), and *X. victorianus* (the Lake Victoria clawed frog). As currently recognized by these two databases, *X. laevis* includes populations from southern Africa (e.g. South Africa, Namibia, Botswana, Zimbabwe, southern Zambia, Malawi) and the northwestern portion of Central Africa (e.g. Central African Republic, Cameroon, eastern Nigeria, Fig. 1), and would thus include at least two named subspecies: *X. l. laevis* and *X. l. sudanensis* (Perret, Jean-Luc 1966) in northern Central Africa.

X. petersii, originally described by Bocage (1895) based on specimens from Angola, was later classified as a subspecies of *X. laevis* (Parker 1936b). The taxon was again recognized as a distinct species by Channing (2001), and support for this classification was also argued on the basis of divergence in mitochondrial DNA (Measey and Channing 2003). *X. poweri* (Hewitt 1927) was considered a subspecies of *X. laevis* by some authors (Schmidt and Inger 1949; Poynton 1964), or a synonym with *X. (laevis) petersii* (e.g., Parker 1936b; Poynton and Broadley 1985). If *X. poweri* is a synonym of *X. petersii*, this species occurs in northern Namibia, northern Botswana, northern Zimbabwe, Zambia, Angola, the Democratic Republic of the Congo, Republic of the Congo, and southern Gabon (Fig. 5.1 Frost 2011; *Information on Amphibian Biology and Conservation [Web Application]* 2014).

X. victorianus was described by Ahl (1924), subsequently synonymized with *X. laevis* (Loveridge 1925, 1933), and was again recognized as a species by Channing and Howell (2006) and Pickersgill (2007). *X. victorianus* occurs in Tanzania, Burundi, Rwanda, eastern Democratic Republic of the Congo, Uganda, southern South Sudan, and Kenya (Fig. 5.1 Frost 2011; *Information on Amphibian Biology and Conservation [Web Application]* 2014). The type locality of the subspecies *X. laevis bunyoniensis* (the western shore of Lake Bunyonyi in southwest Uganda), described by Loveridge (1932), is also situated in this region.

There are multiple lines of evidence to support the recognition of *X. laevis*, *X. victorianus*, and *X. petersii* as separate species. For instance, *X. laevis* is larger than *X. petersii* and *X. victorianus*, with females averaging 110 mm snout–vent length (Kobel et al. 1996), although *X. victorianus* is of comparable size to *X. petersii* (females are ~62 or 65 mm respectively). *X. laevis* has flattened black claws; dorsal and ventral patterning and coloration is highly variable, and the venter is often white or grey. *X. victorianus* has more slender black claws, and highly variable coloration of the venter (Loveridge 1933), whereas *X. petersii* has narrow rounded black claws with a mottled black venter (Loveridge 1933). Additionally, *X. laevis* and *X. petersii* differ in the number of sensory organs surrounding the eye (the latter has 14 instead of 17; Channing 2001). All three of these species have distinct mating calls (Vigny 1979; Tobias et al. 2011); *X. laevis* has pronounced intraspecific variation in mating calls between animals from Malawi and the Western Cape Province, South Africa (Tobias et al. 2011).

The recognition of *X. laevis* as including populations from southern Africa and also populations from Central Africa (Fig. 5.1 Frost 2011; *Information on Amphibian Biology and Conservation [Web Application]* 2014) is inconsistent with the current understanding of evolutionary relationships within *X. laevis sensu lato*, which suggests that populations from West Central Africa are more closely related to populations from East Africa than to those from southern Africa (Evans et al. 2004; Evans et al. 2011a; Evans et al. 2011b). If taxonomy is to reflect evolutionary relationships among diverged lineages, this could be reconciled either by (i) splitting *X. laevis sensu lato* into more than one species – but not in the way that *X. laevis*, *X. victorianus*, and *X. petersii* are currently recognized by at least two taxonomic databases, or by (ii) recognizing only *X. laevis*, which would include the currently recognized *X. victorianus* and *X. petersii*.

D2 Supplemental Tables and Figures

TABLE D2.1: Locality information for genetic samples used in this study and details on sequence data collected for each sample. Please find spreadsheet file with the published paper here.

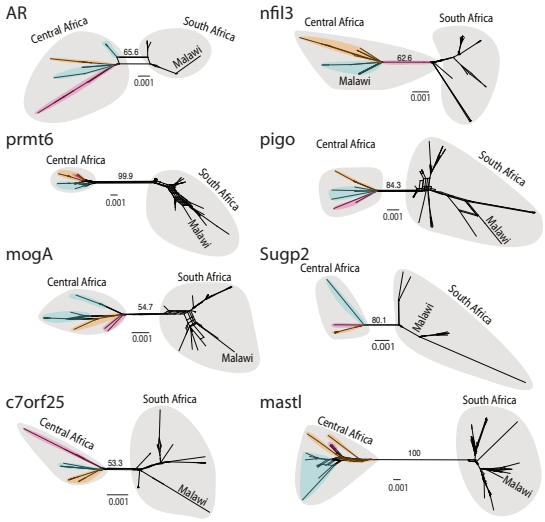


FIGURE D2.1: Gene networks of the 15 autosomal loci for *X. laevis*. Scale bars indicate substitutions/site, and values are bootstrap support for selected major lineages, which usually connect sequences from southern and Central Africa. Lineages from East Africa, Central Africa, and West Central Africa are shaded in green, orange, and pink, respectively.

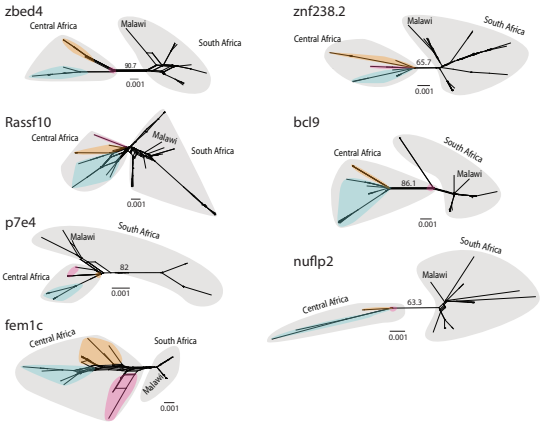


FIGURE D2.1: continued.

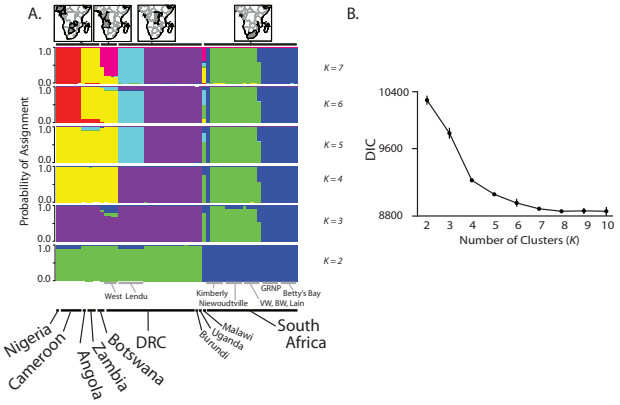


FIGURE D2.2: TESS analysis with all individuals from North and Central Africa ($n = 41$), with a reduced representation of individuals from South Africa ($n = 25$), including (A) individual assignments and (B) the deviance information criterion (DIC) begins to level off at values of K greater than 5, supporting 6–7 clusters. Labeling follows Fig. 5.5

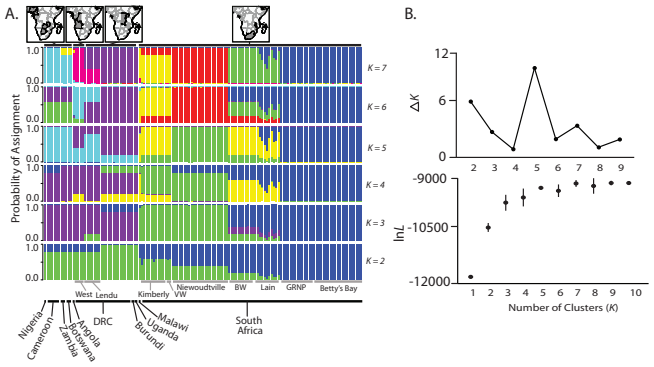


FIGURE D2.3: STRUCTURE analysis of (A) 135 *X. laevis* individuals for 15 loci with labeling following Fig. 5.5. (B) The ΔK statistic (Evanno et al. 2005) indicates support for 5 clusters, which corresponds with the value of K where the log-likelihood (lnL) begins to level off.

Appendix E

Supplemental Information: Limited genomic consequences of hybridization between two African clawed frogs, *Xenopus gilli* and *X. laevis* (Anura: Pipidae)

E1 Results

TABLE E1.1: Sample IDs and loci sequenced for each individual in this study. Locus acronyms are defined in the main text. A 1 or a 0 indicates that the gene was or was not sequenced respectively

Sample ID	16s	btbd6	c7orf25	fem1c	mastl	mogs	nfil1	pcdh1	prmt6	rassf10	sugp2	zbed4	pigo	bcl9
XgUAE_01_gilli_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
XgUAE_05_gilli_WestCape_2013	1	0	1	1	0	1	1	1	1	1	1	1	0	1
XgUAE_02_gilli_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
XgUAE_08_gilli_WestCape_2013	1	1	1	1	1	0	1	1	1	1	1	1	1	1
XgUAE_66_gilli_WestCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_04_gilli_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
XgUAE_07_gilli_WestCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	0
XgUAE_11_gilli_WestCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_06_gilli_WestCape_2013	1	1	1	1	1	0	1	1	0	1	1	1	1	1
XgUAE_03_gilli_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
new_XgUAE_05_gilli_WestCape_2013	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Xb_2_3_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_1_gilli_WestCape_1994	0	1	0	0	0	1	1	0	1	0	1	1	0	0
Rg_0_3_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_2_gilli_WestCape_1994	0	1	1	0	0	1	1	0	1	0	1	1	0	0
Sd_3_2_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Rg_3_1_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_0_2_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sd_2_0_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Rg_1_3_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Rg_2_3_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sd_0_1_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_0_1_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xb_12_1_gilli_WestCape_1994	0	1	1	1	1	1	1	0	1	1	1	1	1	1
Jd_0_2_gilli_WestCape_1994	0	1	1	0	0	1	1	1	1	1	1	1	1	1
Jd_1_2_gilli_WestCape_1994	0	1	1	1	1	1	1	1	0	1	1	1	1	1
Jd_3_0_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Jd_1_0_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sd_3_1_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_1_2_0_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Sd_1_3_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Xb_1_2_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Xg_123_3_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Xb_12_13_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_2_1_gilli_WestCape_1994	0	1	1	1	1	1	1	0	1	1	1	1	1	1
Xg_23_1_gilli_WestCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Rg_1_1_gilli_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xb_3_0_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Xg_3_2_gilli_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_132_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_133_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_130_gilli_EastCape_2013	1	1	1	0	1	1	1	1	1	1	1	1	1	1
XgUAE_131_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_128_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_129_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	0	1	1
XgUAE_127_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_134_gilli_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Xm_3_0_gilli_EastCape_1994	0	1	1	1	1	0	1	0	1	1	1	0	0	1
Xm_0_1_gilli_EastCape_1994	0	1	1	1	1	0	1	0	1	1	1	1	0	0
Xr_2_2_gilli_EastCape_1994	0	1	1	1	1	0	1	0	0	1	1	1	0	0
Xr_16_gilli_EastCape_1994	1	1	1	1	1	0	1	0	1	1	1	1	0	1
Xm_2_0_gilli_EastCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Xr_3_12_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Sample ID	16s	btbd6	c7orf25	fem1c	mastl	mogs	nfil1	pcdh1	prmt6	rassf10	sugp2	zbed4	pigo	bcl9
Xm_0_2_gilli_EastCape_1994	0	0	1	1	1	0	1	0	1	1	1	1	0	0
Xr_12_1_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xr_3_1_gilli_EastCape_1994	0	1	1	1	1	0	1	0	1	1	1	1	0	0
Xr_3_2_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Km_1_gilli_EastCape_1994	0	1	1	1	1	0	1	0	1	1	1	0	0	0
Xr_1_1_gilli_EastCape_1994	1	1	0	1	1	0	1	0	1	1	1	1	0	1
Xm_2_1_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xr_1_2_gilli_EastCape_1994	0	1	1	1	1	0	1	0	1	1	1	1	0	1
Xs_92_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xr_31_2_gilli_EastCape_1994	0	1	1	1	1	0	1	0	0	1	1	1	0	0
Xr_2_1_gilli_EastCape_1994	1	1	1	1	1	0	1	0	1	1	1	1	0	1
Xr_0_3_gilli_EastCape_1994	1	1	1	1	1	0	1	0	1	1	1	1	0	1
Xr_2_12_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Km_2_gilli_EastCape_1994	1	1	1	1	1	0	0	0	1	1	1	1	0	0
Km_01_3_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xr_12_3_gilli_EastCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
XgUAE_63_laevis_WestCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_62_laevis_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
XgUAE_69_laevis_WestCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_65_laevis_WestCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_68_laevis_WestCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_60_laevis_WestCape_2013	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Xl_2_0_laevis_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Rgl_1_laevis_WestCape_1994	1	1	1	1	1	0	1	0	0	1	1	1	1	1
Lg_12_0_laevis_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Lg_2_1_laevis_WestCape_1994	0	0	1	0	1	1	1	0	1	1	1	1	1	0
Xl_28_laevis_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Lg_2_3_laevis_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Rgl_2_laevis_WestCape_1994	0	1	1	1	1	1	1	1	0	1	1	1	1	1
Lg_12_38_laevis_WestCape_1994	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Lg_1_1_laevis_WestCape_1994	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Rgl_3_laevis_WestCape_1994	0	1	1	1	1	0	1	0	0	1	1	1	1	1
Lg_2_2_laevis_WestCape_1994	0	1	1	1	1	0	0	0	0	1	1	1	1	1
'XgUAE_124_laevis_EastCape_2013	0	0	0	0	0	0	0	1	0	0	0	0	0	0
XgUAE_103_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_109_laevis_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_105_laevis_EastCape_2013	1	0	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_106_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_123_laevis_EastCape_2013	1	0	1	0	1	1	1	0	1	1	0	1	1	1
XgUAE_124_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_108_laevis_EastCape_2013	1	1	1	0	1	1	1	0	1	1	1	1	1	1
XgUAE_104_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	0	1	1
XgUAE_100_laevis_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
XgUAE_102_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_101_laevis_EastCape_2013	1	1	1	1	1	1	1	0	1	1	1	1	1	1
XgUAE_107_laevis_EastCape_2013	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ea_12_1_laevis_EastCape_1994	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Kml_5_laevis_EastCape_1994	1	0	0	0	0	1	0	0	1	0	0	0	0	1
Kml_7_laevis_EastCape_1994	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Xsl_5_laevis_EastCape_1994	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Xsl_15_laevis_EastCape_1994	0	1	1	0	1	1	1	0	1	1	1	1	1	1
Kml_8_laevis_EastCape_1994	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Ea_5_laevis_EastCape_1994	0	1	1	1	1	0	1	0	0	1	1	1	1	1
Xsl_3_laevis_EastCape_1994	1	0	0	0	0	1	0	0	1	0	0	0	0	1
Xsl_7_laevis_EastCape_1994	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Kml_6_laevis_EastCape_1994	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Ea_6_laevis_EastCape_1994	0	1	1	1	1	1	1	0	1	1	1	0	1	1

Sample ID	16s	btbd6	c7orf25	fem1c	mastl	mogs	nfil1	pcdh1	prmt6	rassf10	sugp2	zbed4	pigo	bcl9
Xsl_18_laevis_EastCape_1994	0	1	0	0	0	0	1	0	0	0	1	1	0	0
Ea_7_laevis_EastCape_1994	1	1	1	0	1	1	1	0	1	1	1	1	1	0
Xsl_4_laevis_EastCape_1994	1	1	1	0	1	0	1	0	0	1	1	1	1	1
Ea_4_laevis_EastCape_1994	1	0	0	0	0	1	0	0	1	0	0	0	0	0

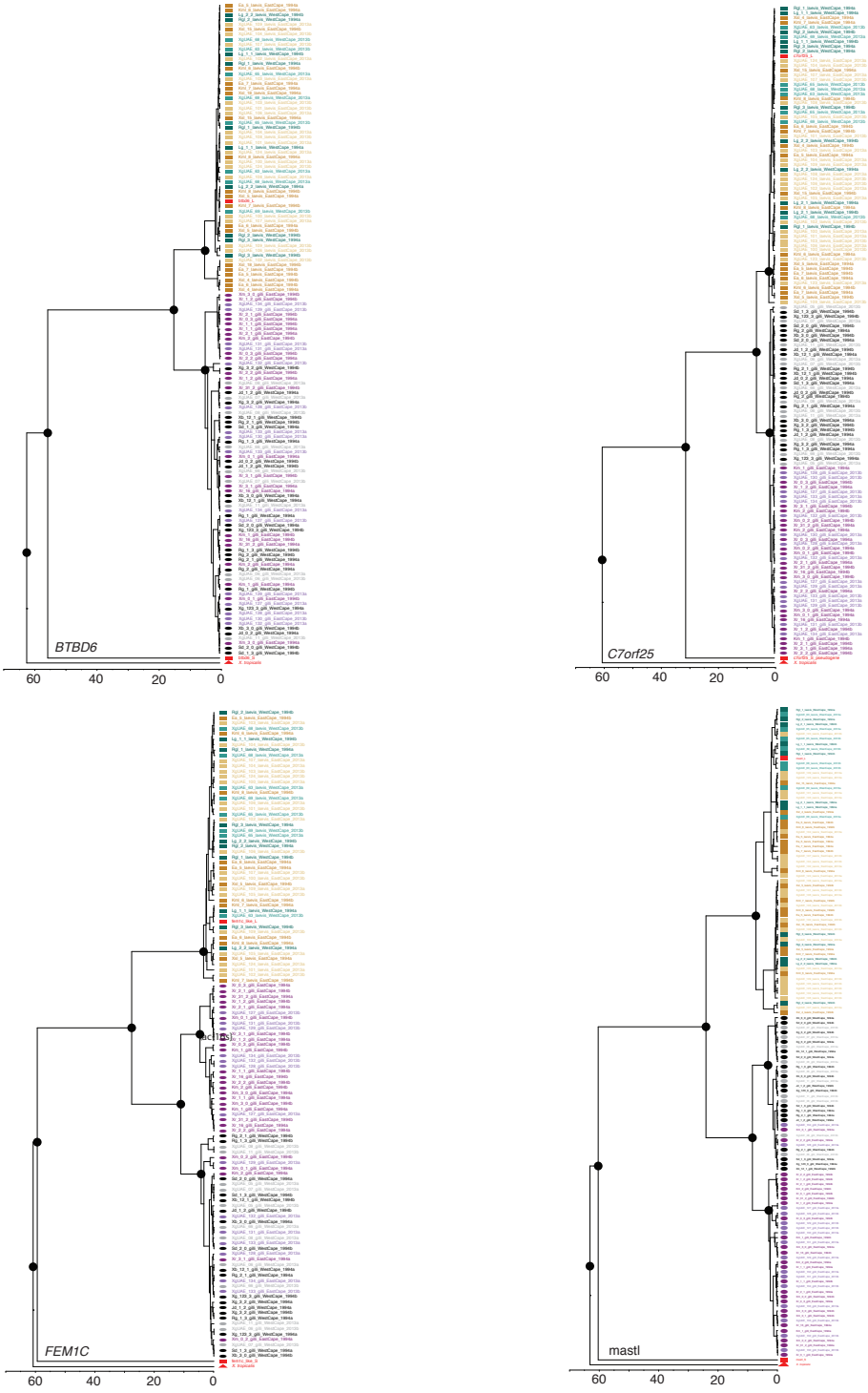


FIGURE E1.1: Phylogenetic trees for all loci included in this study. Sample names included at tip labels.

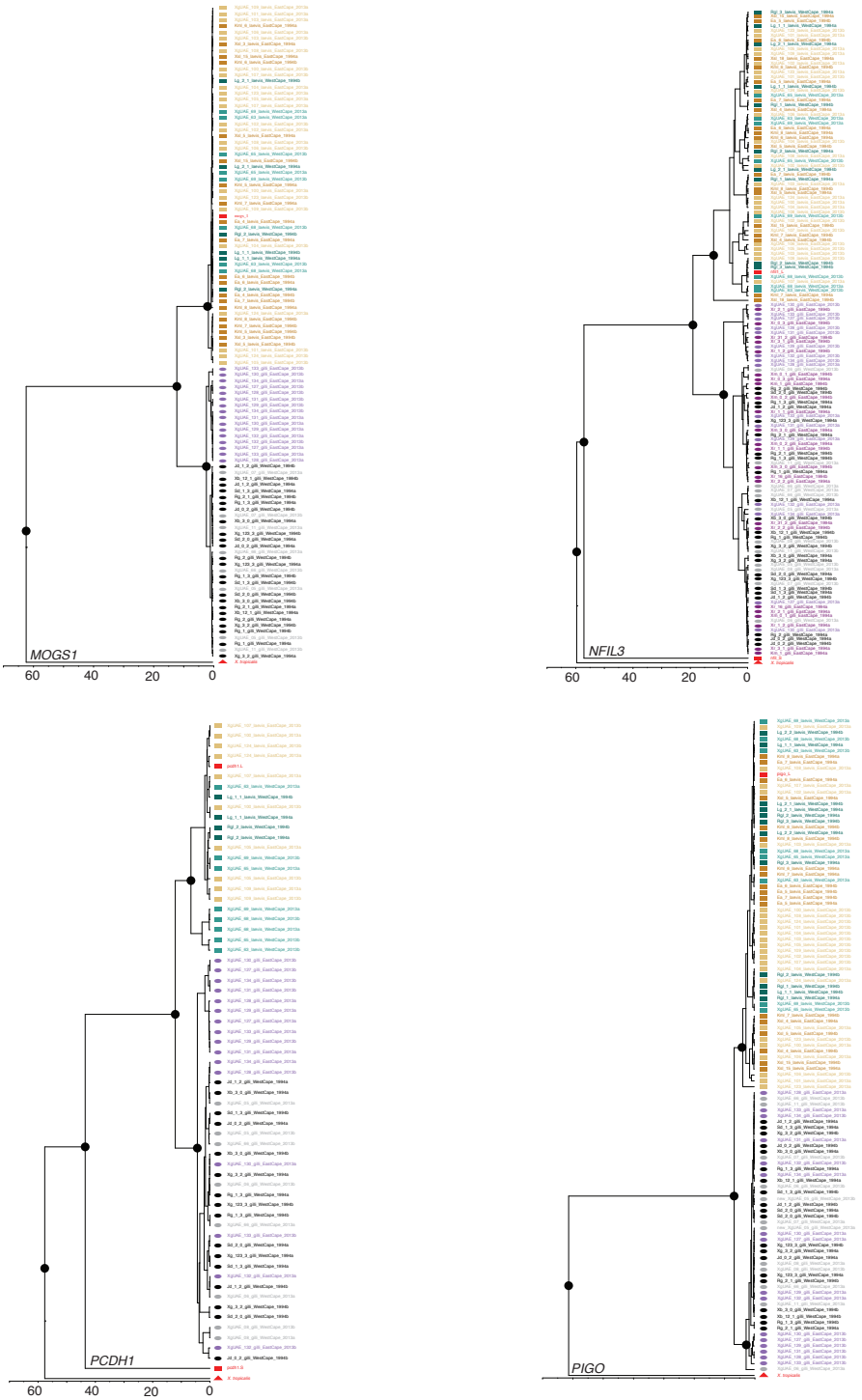


FIGURE E1.1: continued.

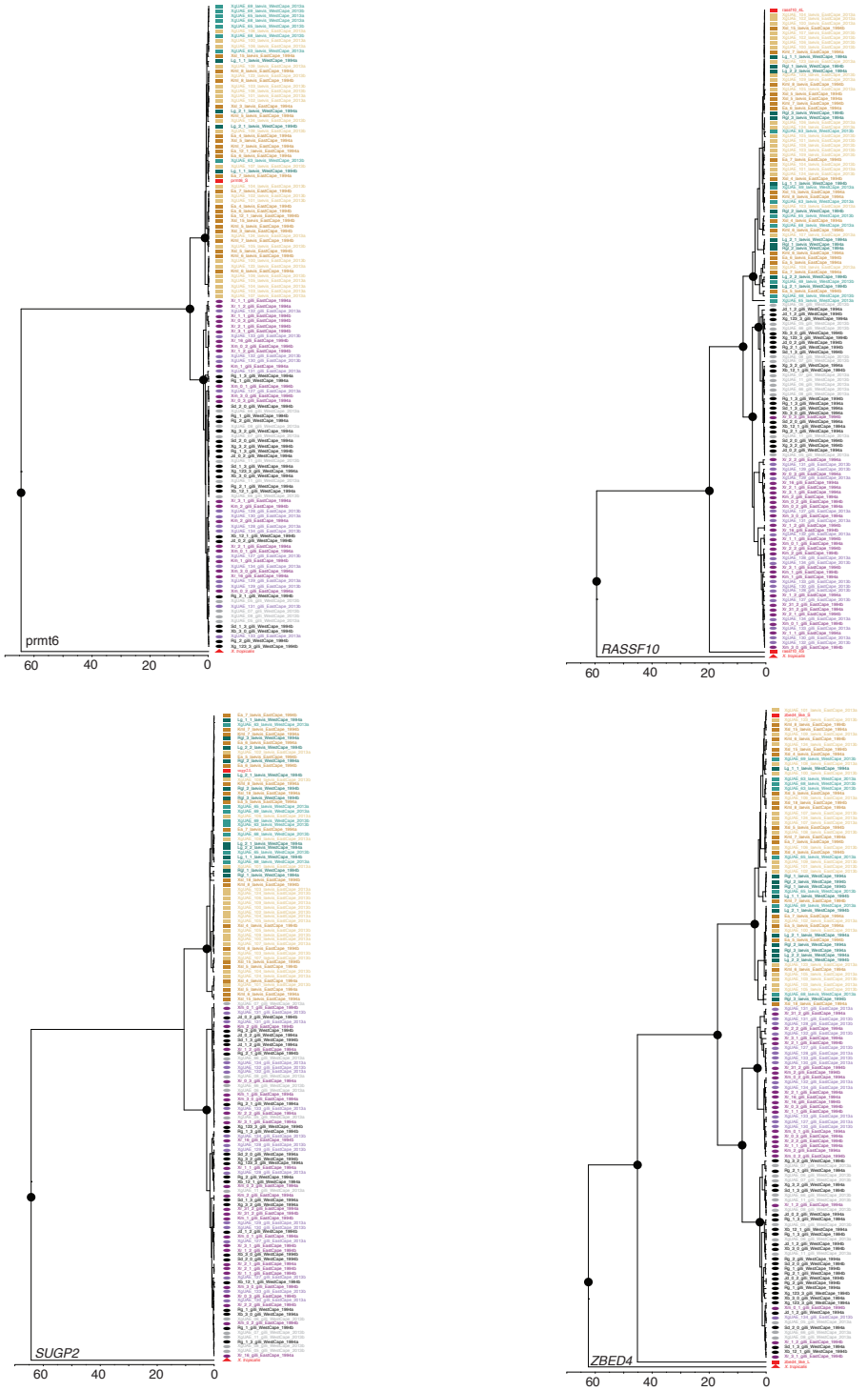


FIGURE E1.1: continued.

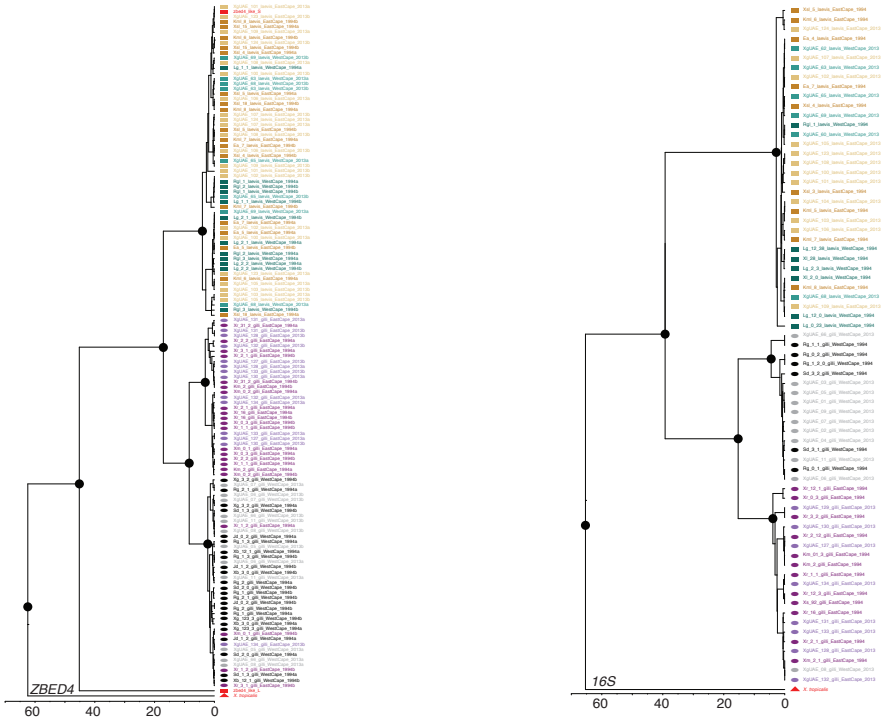


FIGURE E1.1: continued.

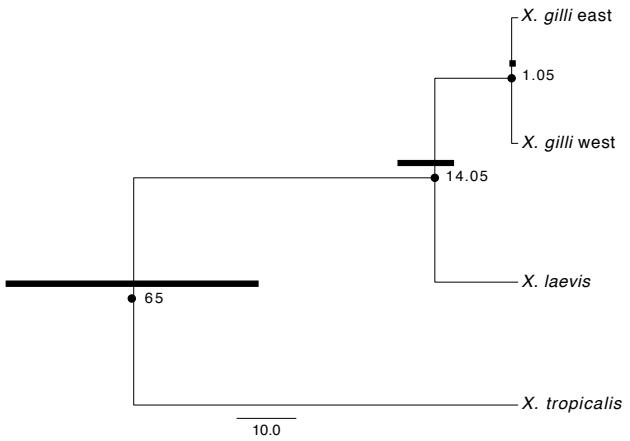


FIGURE E1.2: *BEAST analysis of the same 10 loci used in the Structure analysis (see Materials and Methods). The root of the tree was scaled to a 65 my divergence time from *X. tropicalis* (Bewick et al. 2012). Error bars represent 95% HDP on the height of a node.

Bibliography

Adams, K. L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. *Journal of Heredity* 98(2), 136–141.

Adolfsson, S. and Ellegren, H. (2013). Lack of dosage compensation accompanies the arrested stage of sex chromosome evolution in ostriches. *Molecular Biology and Evolution* 30(4), 806–810.

Ahl, E. (1924). *Über eine froschsammlung aus Nordost-Afrika und Arabien*. Vol. 11. 1.

Ali, J. R. and Aitchison, J. C. (2008). Gondwana to Asia: plate tectonics, paleogeography and the biological connectivity of the Indian sub-continent from the Middle Jurassic through latest Eocene (166–35 Ma). *Earth-Science Reviews* 88(3-4), 145–166.

Allendorf, F. W. and Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary genetics of fishes*. Springer, 1–53.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997a). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), 3389–3402.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997b). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17), 3389–3402.

Amster, G. and Sella, G. (2016). Life history effects on the molecular clock of autosomes and sex chromosomes. *Proceedings of the National Academy of Sciences* 113(6), 1588–1593.

Anderson, E. et al. (1949). Introgressive hybridization. *Introgressive hybridization.*

Anderson, E. and Hubricht, L. (1938). Hybridization in Tradescantia. III. The evidence for introgressive hybridization. *American Journal of Botany*, 396–402.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17(2), 81–92.

Arnold, M. L. (2006). *Evolution through genetic exchange*. Vol. 3. Oxford University Press Oxford.

Arnold, M. L. and Martin, N. H. (2009). Adaptation by introgression. *Journal of Biology* 8(9), 1.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444(7116), 171–178.

Bachtrog, D. (2013). Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nature Reviews Genetics* 14(2), 113.

Bachtrog, D. and Charlesworth, B. (2002). Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* 416(6878), 323–326.

Bachtrog, D., Hom, E., Wong, K. M., Maside, X., and Jong, P. de (2008). Genomic degradation of a young Y chromosome in Drosophila miranda. *Genome Biology* 9(2), R30.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., et al. (2014). Sex determination: why so many ways of doing it? *PLoS Biology* 12(7), e1001899.

Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., Webster, M. T., Öst, T., Schneider, M., Kempenaers, B., et al. (2010). The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research* 20(4), 485–495.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* 3(10), e3376.

Barton, N. H. and Charlesworth, B. (1998). Why sex and recombination? *Science* 281(5385), 1986–1990.

Batada, N. N. and Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics* 39(8), 945–949.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and De Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967), 836–840.

Baxter, S. W., Davey, J. W., Johnston, J. S., Shelton, A. M., Heckel, D. G., Jiggins, C. D., and Blaxter, M. L. (2011). Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One* 6(4), e19315.

Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research* 17(10), 1505–1519.

Becquet, C. and Przeworski, M. (2009). Learning about modes of speciation by computational approaches. *Evolution* 63(10), 2547–2562.

Bergero, R. and Charlesworth, D. (2009). The evolution of restricted recombination in sex chromosomes. *Trends in Ecology and Evolution* 24(2), 94–102.

Bergero, R., Charlesworth, D., Filatov, D. A., and Moore, R. C. (2008). Defining regions and rearrangements of the *Silene latifolia* Y chromosome. *Genetics* 178(4), 2045–2053.

Bergero, R., Forrest, A., Kamau, E., and Charlesworth, D. (2007). Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes. *Genetics* 175(4), 1945–1954.

Berset-Brändli, L., Jaquiéry, J., Broquet, T., Ulrich, Y., and Perrin, N. (2008). Extreme heterochiasmy and nascent sex chromosomes in European tree frogs. *Proceedings of the Royal Society of London B: Biological Sciences* 275(1642), 1577–1585.

Berta, P., Hawkins, J. B., Sinclair, A. H., Taylor, A., Griffiths, B. L., Goodfellow, P. N., and Fellous, M. (1990). Genetic evidence equating *SR-Y* and the testis-determining factor. *Nature* 348(6300), 448.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications* 5, 3657.

Bewick, A. J., Chain, F. J. J., Heled, J., and Evans, B. J. (2012). The Pipid Root. *Systematic Biology* 61(6), 913–926.

Bewick, A. J., Anderson, D. W., and Evans, B. J. (2011). Evolution of the closely related, sex-related genes *DM-W* and *DMRT1* in African clawed frogs (*Xenopus*). *Evolution* 65(3), 698–712.

Bewick, A. J., Chain, F. J., Zimmerman, L. B., Sesay, A., Gilchrist, M. J., Owens, N. D., Seifertova, E., Krylov, V., Macha, J., Tlapakova, T., et al. (2013). A large pseudoautosomal region on the sex chromosomes of the frog *Silurana tropicalis*. *Genome Biology and Evolution* 5(6), 1087–1098.

Birchler, J. A. and Newton, K. J. (1981). Modulation of protein levels in chromosomal dosage series of maize: the biochemical basis of aneuploid syndromes. *Genetics* 99(2), 247–266.

Blackler, A. W. and Fischberg, M. (1968). Hybridization of *Xenopus laevis petersi* (*poweri*) and *X laevis*. *Revue suisse de zoologie; annales de la Societe zoologique suisse et du Museum d'histoire naturelle de Geneve* 75(4), 1023.

Blackler, A. W., Fischberg, M., and Newth, D. R. (1965). Hybridization of two subspecies of *Xenopus laevis*. *Revue Suisse de Zoologie* 72, 841–857.

Blair, A. C., Blumenthal, D., and Hufbauer, R. A. (2012). Hybridization and invasion: an experimental test with diffuse knapweed (*Centaurea diffusa Lam.*) *Evolutionary Applications* 5(1), 17–28.

Blanc, G. and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell* 16(7), 1679–1691.

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., Taylor, J., Nekrutenko, A., et al. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biology* 15(2), 403.

Bocage, J. V. B. du (1895). *Herpétologie d'Angola et du Congo: ouvrage publié sous les auspices du minstére de la marine et des colonies*. Imprimerie Nationale.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

Bowes, J. B., Snyder, K. A., Segerdell, E., Gibb, R., Jarabek, C., Noumen, E., Pollet, N., and Vize, P. D. (2008). Xenbase: a *Xenopus* biology and genomics resource. *Nucleic Acids Research* 36(suppl 1), D761–D767.

Bowes, J. B., Snyder, K. A., Segerdell, E., Jarabek, C. J., Azam, K., Zorn, A. M., and Vize, P. D. (2009). Xenbase: gene expression and improved integration. *Nucleic Acids Research*, gkp953.

Brelsford, A., Dufresnes, C., and Perrin, N. (2016). High-density sex-specific linkage maps of a European tree frog (*Hyla arborea*) identify the sex chromosome without information on offspring sex. *Heredity* 116(2), 177–181.

Brelsford, A., Stöck, M., Betto-Colliard, C., Dubey, S., Dufresnes, C., Jourdan-Pineau, H., Rodrigues, N., Savary, R., Sermier, R., and Perrin, N. (2013). Homologous sex chromosomes in three deeply divergent Anuran species. *Evolution* 67(8), 2434–2440.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution* 23(9), 1808–1816.

Bryant, D. and Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2), 255–265.

Buggs, R. J., Chamala, S., Wu, W., Tate, J. A., Schnable, P. S., Soltis, D. E., Soltis, P. S., and Barbazuk, W. B. (2012). Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology* 22(3), 248–252.

Bull, J. J. (1983). *Evolution of sex determining mechanisms*. The Benjamin/Cummings Publishing Company, Inc.

Butlin, R. K. (2005). Recombination and speciation. *Molecular Ecology* 14(9), 2621–2635.

Caballero, A. and García-Dorado, A. (2013). Allelic diversity and its implications for the rate of adaptation. *Genetics* 195(4), 1373–1384.

Canestro, C., Albalat, R., Irimia, M., and Garcia-Fernàndez, J. (2013). Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. In: *Seminars in cell & developmental biology*. Vol. 24. 2. Elsevier, 83–94.

Cannatella, D. (2015). *Xenopus* in space and time: fossils, node calibrations, tip-dating, and paleobiogeography. *Cytogenetic and Genome Research* 145(3-4), 283–301.

Cannatella, D. C. and Sá, R. O. de (1993). *Xenopus laevis* as a Model Organism. *Systematic Biology* 42(4), 476–507.

Cannatella, D. C. and Sa, R. O. de (1993). *Xenopus laevis* as a Model Organism. *Systematic Biology* 42(4), 476–507.

Cannatella, D. C. and Trueb, L. (1988). Evolution of pipoid frogs: intergeneric relationships of the aquatic frog family Pipidae (Anura). *Zoological Journal of the Linnean Society* 94(1), 1–38.

Carr, S. M., Brothers, A. J., and Wilson, A. C. (1987). Evolutionary inferences from restriction maps of mitochondrial DNA from nine taxa of *Xenopus* frogs. *Evolution* 41(1), 176–188.

Castresana, J. (Apr. 2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17(4), 540–52.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes/Genomes/Genetics* 1(3), 171–182.

Chain, F. J. J. and Evans, B. J. (2006). Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. *PLoS Genetics* 2(4), e56.

Chain, F. J., Dushoff, J., and Evans, B. J. (2011). The odds of duplicate gene persistence after polyploidization. *BMC Genomics* 12(1), 599.

Chain, F. J., Ilieva, D., and Evans, B. J. (2008). Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evolutionary Biology* 8(1), 43.

Channing, A. (2001). *Amphibians of central and southern Africa*. Comstock Pub. Associates.

Channing, A. and Howell, K. (2006). *Amphibians of East Africa*. Comstock Pub. Associates/Cornell University Press.

Charlesworth, B. (1991). The evolution of sex chromosomes. *Science* 251(4997), 1030–1033.

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10(3), 195.

Charlesworth, B. and Charlesworth, D. (1973). Selection of new inversions in multi-locus genetic systems. *Genetics Research* 21(2), 167–183.

Charlesworth, B. and Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 355(1403), 1563–1572.

Charlesworth, B., Morgan, M., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4), 1289–1303.

Charlesworth, D., Charlesworth, B., and Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95(2), 118–128.

Chen, C., Durand, E., Forbes, F., and François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Resources* 7(5), 747–756.

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., An, N., Chalopin, D., Volff, J.-N., et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* 46(3), 253–260.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PloS one* 7(5), e36442.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11), 1422–1423.

Cohen, O. and Pupko, T. (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* 27(3), 703–713.

Colliard, C., Sicilia, A., Turrisi, G. F., Arculeo, M., Perrin, N., and Stöck, M. (2010). Strong reproductive barriers in a narrow hybrid zone of West-Mediterranean green toads (*Bufo viridis* subgroup) with Plio-Pleistocene divergence. *BMC Evolutionary Biology* 10(1), 232.

Comai, L. (2000). Genetic and epigenetic interactions in allopolyploid plants. *Plant Molecular Biology* 43(2-3), 387–399.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6(11), 836.

Comai, L., Madlung, A., Josefsson, C., and Tyagi, A. (2003). Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358(1434), 1149–1155.

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. *New Phytologist* 200(1), 251–260.

Conlon, J. M., Mechkarska, M., Coquet, L., Leprince, J., Jouenne, T., Vaudry, H., and Measey, G. J. (2015). Evidence from peptidomic analysis of skin secretions that allopatric populations of *Xenopus gilli* (Anura: Pipidae) constitute distinct lineages. *Peptides* 63, 118–125.

Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. *Nature Reviews Genetics* 8, 23–34.

Cui, Z., Hui, M., Liu, Y., Song, C., Li, X., Li, Y., Liu, L., Shi, G., Wang, S., Li, F., et al. (2015). High-density linkage mapping aided by transcriptomics documents ZW sex determination system in the Chinese mitten crab *Eriocheir sinensis*. *Heredity* 115(3), 206.

Dang, U. J., Devault, A. M., Mortimer, T. D., Pepperell, C. S., Poinar, H. N., and Golding, G. B. (2016). Estimation of Gene Insertion/Deletion Rates with Missing Data. *Genetics* 204(2), 513–529.

Dang, U. J. and Golding, G. B. (2016). markophylo: Markov Chain Analysis on Phylogenetic Trees. *Bioinformatics* 32(1), 130–132.

Daudin, F. M. (1802). *Histoire naturelle des rainettes, des grenouilles et des crapauds. Ouvrage orné de 38 planches représentant 54 espèces peintes d'après nature.* Levrault.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12(7), 499–510.

De Queiroz, K. (1998). The general lineage concept of species, species criteria, and the process of speciation. In: *Endless Forms: Species and Speciation.* Oxford Press, New York, 57–75.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology* 56(6), 879–886.

Dehal, P. and Boore, J. L. (Oct. 2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3(10), e314.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* 30(11), 2478–2483.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5), 491–498.

Devlin, R. H. and Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* 208(3), 191–364.

Dowling, T. E. and Secor, C. L. (1997). The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics*, 593–619.

Drummond, A. J. and Rambaut, A. (Jan. 2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7(1), 214.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (Aug. 2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8), 1969–73.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences* 102(40), 14338–14343.

Du Preez, L. H., Kunene, N., Hanner, R., Giesy, J. P., Solomon, K. R., Hosmer, A., and Van Der Kraak, G. J. (2009). Population-specific incidence of testicular ovarian follicles in *Xenopus laevis* from South Africa: A potential issue in endocrine testing. *Aquatic Toxicology* 95(1), 10–16.

Dufresnes, C., Bertholet, Y., Wassef, J., Ghali, K., Savary, R., Pasteur, B., Brelsford, A., Rozenblut-Kościsty, B., Ogielska, M., Stöck, M., et al. (2014). Sex-chromosome differentiation parallels postglacial range expansion in European tree frogs (*Hyla arborea*). *Evolution* 68(12), 3445–3456.

Dufresnes, C., Brelsford, A., Crnobrnja-Isailović, J., Tzankov, N., Lymberakis, P., and Perrin, N. (2015). Timeframe of speciation inferred from secondary contact zones in the European tree frog radiation (*Hyla arborea* group). *BMC Evolutionary Biology* 15(1), 155.

Dupont, L. (2011). Orbital scale vegetation change in Africa. *Quaternary Science Reviews* 30(25-26), 3589–3602.

Durand, E., Jay, F., Gaggiotti, O. E., and François, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution* 26(9), 1963–1973.

Earl, D. A. et al. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4(2), 359–361.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., and Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40(8), 1–18.

Eggert, C. and Fouquet, A. (2006). A preliminary biotelemetric study of a ferai invasive *Xenopus laevis* population in France. *Alytes* 23(3-4), 144–149.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6(5), e19379.

Eulenstein, O., Huzurbazar, S., and Liberles, D. A. (2010). Reconciling Phylogenetic Trees. In: *Evolution after Gene Duplication*. Ed. by K. Dittmar and D. Liberles. Hoboken, NJ, USA: John Wiley & Sons, Inc. Chap. 10, 185–206.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14(8), 2611–2620.

Evans, B. J., Greenbaum, E., Kusamba, C., Carter, T. F., Tobias, M. L., Mendel, S. A., and Kelley, D. B. (Apr. 2011a). Description of a new octoploid frog species (Anura: Pipidae: *Xenopus*) from the Democratic Republic of the Congo, with a discussion of the biogeography of African clawed frogs in the Albertine Rift. *Journal of Zoology* 283(4), 276–290.

Evans, B. J. (June 2007). Ancestry influences the fate of duplicated genes millions of years after polyploidization of clawed frogs (*Xenopus*). *Genetics* 176(2), 1119–1130.

Evans, B. J., Carter, T. F., Tobias, M. L., Kelley, D. B., Hanner, R., and Tinsley, R. C. (2008). A new species of clawed frog (genus *Xenopus*) form the Itombwe Massif, Democratic Republic of the Congo: implication for DNA barcodes and biodiversity conservation. *Zootaxa* 1780, 55–68.

Evans, B. J., Pyron, R. A., and Weins, J. J. (2012). Polyploidy and sex chromosome evolution in amphibians. In: *Polyploidy and Genome Evolution*. Ed. by P. S. Soltis and D. E. Soltis. Springer, 385–410.

Evans, B. J., Kelley, D. B., Melnick, D. J., and Cannatella, D. C. (2005). Evolution of *RAG-1* in polyploid clawed frogs. *Molecular Biology and Evolution* 22(5), 1193–1207.

Evans, B. J., Bliss, S. M., Mendel, S. A., and Tinsley, R. C. (2011b). The Rift Valley is a major barrier to dispersal of African clawed frogs (*Xenopus*) in Ethiopia. *Molecular Ecology* 20(20), 4216–4230.

Evans, B. J., Carter, T. F., Greenbaum, E., Gvoždík, V., Kelley, D. B., McLaughlin, P. J., Pauwels, O. S., Portik, D. M., Stanley, E. L., Tinsley, R. C., et al. (2015). Genetics, Morphology, Advertisement Calls, and Historical Records Distinguish Six New Polyploid Species of African Clawed Frog (*Xenopus*, Pipidae) from West and Central Africa. *PloS one* 10(12), e0142823.

Evans, B. J., Kelley, D. B., Tinsley, R. C., Melnick, D. J., and Cannatella, D. C. (Oct. 2004). A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution* 33(1), 197–213.

Evans, B. J. and Kwon, T. (2015). Molecular Polymorphism and Divergence of Duplicated Genes in Tetraploid African Clawed Frogs (*Xenopus*). *Cytogenetic and Genome Research* 145(3-4), 243–252.

Evans, B., Morales, J., Picker, M., Kelley, D., and Melnick, D. (1997). Comparative molecular phylogeography of two *Xenopus* species, *X. gilli* and *X. laevis*, in the southwestern Cape Province, South Africa. *Molecular Ecology* 6(4), 333–343.

Evans, B., Morales, J., Picker, M., Melnick, D., and Kelley, D. (1998). Absence of extensive introgression between *Xenopus gilli* and *Xenopus laevis laevis* (Anura: Pipidae) in southwestern Cape Province, South Africa. *Copeia* 1998(2), 504–509.

Evans Ben, J. (Jan. 2008). Genome evolution and speciation genetics of clawed frogs (*Xenopus* and *Silurana*). en. *Fronteirs in Bioscience* Volume(13), 4687.

Excoffier, L. and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10(3), 564–567.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164(4), 1567–1587.

Fawcett, J. A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences* 106(14), 5737–5742.

Feldman, M., Levy, A. A., Fahima, T., and Korol, A. (2012). Genomic asymmetry in allopolyploid plants: wheat as a model. *Journal of Experimental Botany* 63 (14), 5045–5059.

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Biology* 22(3), 240–249.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* 78(2), 737–756.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. English. *Journal of Molecular Evolution* 17(6), 368–376.

Felsenstein, J. (1992). Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution* 46 (1), 159–173.

Felsenstein, J. (2004). *Inferring Phylogenies*. Vol. 2. Sunderland, Massachusetts: Sinauer Associates.

Figueroa, M., Castillo, J., Redondo, S., Luque, T., Castellanos, E., Nieva, F., Luque, C., Rubio-Casal, A., and Davy, A. (2003). Facilitated invasion by hybridization of *Sarcocornia* species in a salt-marsh succession. *Journal of Ecology* 91(4), 616–626.

Fischer, W., Koch, W., and Elepfandt, A. (2000). Sympatry and hybridization between the clawed frogs *Xenopus laevis laevis* and *Xenopus muelleri* (Pipidae). *Journal of Zoology* 252(01), 99–107.

Fitzpatrick, B. M., Johnson, J. R., Kump, D. K., Shaffer, H. B., Smith, J. J., and Voss, S. R. (2009). Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders. *BMC Evolutionary Biology* 9(1), 176.

Flagel, L. E. and Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist* 186(1), 184–193.

Flegal, J. M., Hughes, J., and Vats, D. (2016). *mcmcse: Monte Carlo Standard Errors for MCMC*. R package version 1.2-1. Riverside, CA and Minneapolis, MN.

Fogell, D. J., Tolley, K. A., and Measey, G. J. (2013). Mind the gaps: investigating the cause of the current range disjunction in the Cape Platanna, *Xenopus gilli* (Anura: Pipidae). *PeerJ* 1, e166.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4), 1531–1545.

Ford, L. S. and Cannatella, D. C. (1993). The major clades of frogs. *Herpetological Monographs*, 94–117.

François, O. and Durand, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources* 10(5), 773–784.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology* 60, 433–453.

Freeling, M., Woodhouse, M. R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J. C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* 15(2), 131–139.

Frost, D. (2011). *Amphibian Species of the World: an Online Reference. Version 6 (31 January, 2014). American Museum of Natural History, New York, USA*.

Fujii, J., Kodama, M., Oike, A., Matsuo, Y., Min, M.-S., Hasebe, T., Ishizuya-Oka, A., Kawakami, K., and Nakamura, M. (2014). Involvement of Androgen Receptor in Sex Determination in an Amphibian Species. *PLoS One* 9(5), e93655.

Furman, B. L. S., Bewick, A. J., Harrison, T. L., Greenbaum, E., Gvoždík, V., Kusamba, C., and Evans, B. J. (2015). Pan-African phylogeography of a model organism, the African clawed frog *Xenopus laevis*. *Molecular Ecology* 24(4), 909–925.

Furman, B. L. S. and Evans, B. J. (2016). Sequential turnovers of sex chromosomes in African clawed frogs (*Xenopus*) suggest some genomic regions are good at sex determination. *G3: Genes|Genomes|Genetics* 6(11), 3625–3633.

Gallach, M., Domingues, S., and Betrán, E. (2011). Gene duplication and the genome distribution of sex-biased genes. *International Journal of Evolutionary Biology* 2011.

Galtier, N., Enard, D., Radondy, Y., Bazin, E., and Belkhir, K. (2006). Mutation hot spots in mammalian mitochondrial DNA. *Genome Research* 16(2), 215–222.

Gamble, T., Coryell, J., Ezaz, T., Lynch, J., Scantlebury, D. P., and Zarkower, D. (2015). Restriction site-associated DNA sequencing (RAD-seq) reveals an extraordinary number of transitions among gecko sex-determining systems. *Molecular Biology and Evolution*, msv023.

Garamszegi, L. Z. (2014). *Modern phylogenetic comparative methods and their application in evolutionary biology. Concepts and Practice*. London, UK: Springer.

Garsmeur, O., Schnable, J. C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2013). Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31(2), 448–454.

Gay, D. M. (1990). *Usage summary for selected optimization routines*. AT&T Bell Laboratories. Murray Hill, New Jersey.

Geraldes, A., Rambo, T., Wing, R. A., Ferrand, N., and Nachman, M. W. (Nov. 2010). Extensive gene conversion drives the concerted evolution of paralogous copies of the SRY gene in European rabbits. *Molular Biology and Evolution* 27(11), 2437–2440.

Gibeaux, R., Acker, R., Kitaoka, M., Georgiou, G., Kruijsbergen, I. van, Ford, B., Marcotte, E. M., Nomura, D. K., Kwon, T., Veenstra, G. J. C., et al. (2018). Paternal chromosome loss and metabolic crisis contribute to hybrid inviability in *Xenopus*. *Nature*.

Gilk, S. E., Wang, I. A., Hoover, C. L., Smoker, W. W., Taylor, S., Gray, A. K., and Gharrett, A. (2004). *Outbreeding depression in hybrids between spatially separated pink salmon, Oncorhynchus gorbuscha, populations: marine survival, homing ability, and variability in family size*. Springer.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS One* 9(2), e90346.

Gordo, I. and Charlesworth, B. (2001). The speed of Müller's ratchet with background selection, and the degeneration of Y chromosomes. *Genetics Research* 78(2), 149–161.

Goudet, J., Raymond, M., Meeüs, T. de, and Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* 144(4), 1933–1940.

Gout, J.-F., Kahn, D., Duret, L., and Consortium, P. P.-G. (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genetics* 6(5), e1000944.

Gout, J.-F. and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology and Evolution* 32 (8), 2141–2148.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7), 644–652.

Graves, J. A. M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell* 124(5), 901–914.

Graves, J. A. M. and Peichel, C. L. (2010). Are homologies in vertebrate sex determination due to shared ancestry or to limited options? *Genome Biology* 11(4), 205.

Gray, J. E. (1864). Notice of a new genus (*Silurana*) of frogs from West Africa. *Journal of Natural History* 14(82), 315–316.

Greenbaum, G., Templeton, A. R., Zarmi, Y., and Bar-David, S. (2014). Allelic richness following population founding events–A Stochastic modeling framework incorporating gene flow and genetic drift. *PLoS One* 9(12), e115203.

Groenen, M. A., Wahlberg, P., Foglio, M., Cheng, H. H., Megens, H.-J., Crooijmans, R. P., Besnier, F., Lathrop, M., Muir, W. M., Wong, G. K.-S., et al. (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research* 19(3), 510–519.

Grohovaz, G. S., Harley, E., and Fabian, B. (1996). Significant mitochondrial DNA sequence divergence in natural populations of *Xenopus laevis* (Pipidae) from Southern Africa. *Herpetologica*, 247–253.

Gurdon, J. B. (1996). Introductory comments: *Xenopus* as a laboratory animal. In: *The Biology of Xenopus*. 68. Zoological Society of London, 3–8.

Gurdon, J. B. and Hopwood, N. (2003). The introduction of Xenopus laevis into developmental biology: of empire, pregnancy testing and ribosomal genes. *International Journal of Developmental Biology* 44(1), 43–50.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8(8), 1494–1512.

Hackett, C. and Broadfoot, L. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 90(1), 33.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* 33(2), 1–22.

Haldane, J. B. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics* 12(2), 101–109.

Han, M. V., Thomas, G. W., Lugo-Martinez, J., and Hahn, M. W. (2013). Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Molecular Biology and Evolution* 30(8), 1987–1997.

Handley, L.-J. L., Ceplitis, H., and Ellegren, H. (2004). Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics* 167(1), 367–376.

Hansson, B., Ljungqvist, M., Dawson, D., Mueller, J., Olano-Marin, J., Ellegren, H., and Nilsson, J.-Å. (2010). Avian genome evolution: insights from a linkage map of the blue tit (*Cyanistes caeruleus*). *Heredity* 104(1), 67.

Hao, W. and Golding, G. B. (2006). The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* 16(5), 636–643.

Harland, R. M. and Grainger, R. M. (Dec. 2011). *Xenopus* research: metamorphosed by genetics and genomics. *Trends Genetics* 27(12), 507–515.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., and Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics* 24(1), 129–131.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22(2), 160–174.

Hassold, T., Judis, L., Chan, E., Schwartz, S., Seftel, A., and Lynn, A. (2004). Cytological studies of meiotic recombination in human males. *Cytogenetic and Genome Research* 107(3-4), 249–255.

Heled, J. and Drummond, A. J. (Mar. 2010). Bayesian inference of species trees from multilocus data. *Molecular Biology Evolution* 27(3), 570–80.

Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N. H., Shu, S., Taher, L., et al. (2010). The genome of the Western clawed frog *Xenopus tropicalis*. *Science* 328(5978), 633–636.

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

Henrici, A. C. and Báez, A. M. (2001). First occurrence of *Xenopus* (Anura: Pipidae) on the Arabian Peninsula: a new species from the Upper Oligocene of Yemen. *Journal of Paleontology* 75(4), 870–882.

Hewitt, J. (1927). Further descriptions of reptiles and batrachians from South Africa. *Records of the Albany Museum* 3, 371–415.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research* 8(3), 269–294.

Hogan, M. C. (2001). Fall in intracellular PO 2 at the onset of contractions in Xenopus single skeletal muscle fibers. *Journal of Applied Physiology* 90(5), 1871–1876.

Hughes, J. F. and Rozen, S. (2012). Genomics and genetics of human and primate Y chromosomes. *Annual review of genomics and human genetics* 13, 83–108.

Hunt, P. A. and Hassold, T. J. (2002). Sex matters in meiosis. *Science* 296(5576), 2181–2183.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2) (1), 254–267.

Huxley, J. (1928). Sexual difference of linkage in *Gammarus chevreuxi*. *Journal of Genetics* 20(2), 145–156.

Igawa, T., Kurabayashi, A., Usuki, C., Fujii, T., and Sumida, M. (2008). Complete mitochondrial genomes of three neobatrachian anurans: a case study of divergence time estimation using different data and calibration settings. *Gene* 407(1), 116–129.

Information on Amphibian Biology and Conservation [Web Application] (2014). `https://amphibiaweb.org/`.

Innan, H. and Nordborg, M. (2002). Recombination or mutational hot spots in human mtDNA? *Molecular Biology and Evolution* 19(7), 1122–1127.

Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences* 112(48), 14918–14923.

Irisarri, I., Vences, M., San Mauro, D., Glaw, F., and Zardoya, R. (2011). Reversal to air-driven sound production revealed by a molecular phylogeny of tongueless frogs, family Pipidae. *BMC Evolutionary Biology* 11(1), 114.

Jackson, J. A. and Tinsley, R. C. (2001). Host-specificity and distribution of cephalochlamydid cestodes: correlation with allopolyploid evolution of pipid anuran hosts. *Journal of Zoology* 254(3), 405–419.

Jacobs, B. F. (2004). Palaeobotanical studies from tropical Africa: relevance to the evolution of forest, woodland and savannah biomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 359(1450), 1573–1583.

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate protokaryotype. *Nature* 431(7011), 946.

Jakobsson, M. and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23(14), 1801–1806.

Ji, Y., Stelly, D. M., De Donato, M., Goodman, M. M., and Williams, C. G. (1999). A candidate recombination modifier gene for *Zea mays* L. *Genetics* 151(2), 821–830.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., and Soltis, D. E. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345), 97–100.

Johnston, S. E., Bérénos, C., Slate, J., and Pemberton, J. M. (2016). Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay sheep (*Ovis aries*). *Genetics* 203(1), 583–598.

Just, W., Rau, W., Vogel, W., Akhverdian, M., Fredga, K., Graves, J. A., and Lyapunova, E. (Oct. 1995). Absence of *Sry* in species of the vole *Ellobius*. *Nature Genetics* 11(2), 117–118.

Kalinowski, S. T. (2005). hp-rare 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes* 5(1), 187–189.

Kamiya, T., Kai, W., Tasumi, S., Oka, A., Matsunaga, T., Mizuno, N., Fujita, M., Suetake, H., Suzuki, S., Hosoya, S., et al. (2012). A trans-species missense SNP in *Amhr2* is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *PLoS Genetics* 8(7), e1002798.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* 90(430), 773–795.

Katoh, K. and Standley, D. M. (Apr. 2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30(4), 772–80.

Kawai, A., Ishijima, J., Nishida, C., Kosaka, A., Ota, H., Kohno, S., and Matsuda, Y. (Feb. 2009). The ZW sex chromosomes of *Gekko hokouensis* (Gekkonidae, Squamata) represent highly conserved homology with those of avian species. *Chromosoma* 118(1), 43–51.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Olason, P., and Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology* 23(16), 4035–4058.

Kelley, D. and Tobias, M. (1999). The vocal repertoire of *Xenopus laevis*. In: *The Design of Animal Communication*. MIT Press, Cambridge, 9–35.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., Mutke, J., and Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences* 106(23), 9322–9327.

Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* 2(1), 87–90.

Kitano, J., Ross, J. A., Mori, S., Kume, M., Jones, F. C., Chan, Y. F., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., et al. (2009). A role for a neo-sex chromosome in Stickleback speciation. *Nature* 461(7267), 1079–1083.

Kobel, H. R. (1981). Evolutionary trends in *Xenopus* (Anura, Pipidae). *Monitore Zoologico Italiano. Supplemento* 15(1), 119–131.

Kobel, H. R. and Du Pasquier, L. (1986). Genetics of polyploid *Xenopus*. *Trends in Genetics* 2, 310–315.

Kobel, H. R., Pasquier, L. D., and Tinsley, R. C. (1981). Natural hybridization and gene introgression between *Xenopus gilli* and *Xenopus laevis laevis* (Anura: Pipidae). *Journal of Zoology* 194(3), 317–322.

Kobel, H. (1996). Reproductive capacity of experimental *Xenopus gilli x X. l. laevis* hybrids. In: *The Biology of Xenopus*. Zoological Society of London, 73–80.

Kobel, H., Barandun, B., and Thiébaud, C. H. (1998). Mitochondrial rDNA phylogeny in *Xenopus*. *Herpetological Journal* 8(1), 13–17.

Kobel, H., Loumont, C., and Tinsley, R. (1996). The extant species. In: *The Biology of Xenopus*. 68. Zoological Society of London.

Kondo, M., Nanda, I., Hornung, U., Asakawa, S., Shimizu, N., Mitani, H., Schmid, M., Shima, A., and Schartl, M. (Mar. 2003). Absence of the candidate male sex-determining gene *dmrt1b(Y)* of medaka from other fish species. *Current Biology* 13(5), 416–420.

Kondo, M., Nanda, I., Hornung, U., Schmid, M., and Schartl, M. (Sept. 2004). Evolutionary origin of the medaka Y chromosome. *Current Biology* 14(18), 1664–1669.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1991). Male development of chromosomally female mice transgenic for *Sry*. *Nature* 351(6322), 117–121.

Kosambi, D. (1943). The estimation of map distances from recombination values. *Annals of Human Genetics* 12(1), 172–175.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56(1), 17–24.

Kück, P., Mayer, C., Wägele, J.-W., and Misof, B. (2012). Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One* 7(5), e36593.

Lahn, B. T. and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science* 286(5441), 964–967.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2013). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology* 63(1), 17–30.

Lehmann, C. E., Archibald, S. A., Hoffmann, W. A., and Bond, W. J. (2011). Deciphering the distribution of the savanna biome. *New Phytologist* 191(1), 197–209.

Lemmon, A. R. and Moriarty, E. C. (2004). The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology* 53(2), 265–277.

Lenormand, T. (2003). The evolution of sex dimorphism in recombination. *Genetics* 163(2), 811–822.

Lenormand, T. and Dutheil, J. (2005). Recombination difference between sexes: a role for haploid selection. *PLoS Biology* 3(3), e63.

Lewis, P. O. (2001). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology* 50(6), 913–925.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21), 2987–2993.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16), 2078–2079.

Librado, P. and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11), 1451–1452.

Liu, L., Yu, L., and Edwards, S. V. (Jan. 2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* 10(1), 302.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A., Zhao, M., Ma, J., Yu, J., Huang, S., et al. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications* 5, 3930.

Lloyd, R. E., Foster, P. G., Guille, M., and Littlewood, D. T. J. (2012). Next generation sequencing and comparative analyses of *Xenopus* mitogenomes. *BMC Genomics* 13(1), 496.

Lobos, G. and Jaksic, F. M. (2005). The ongoing invasion of African clawed frogs (*Xenopus laevis*) in Chile: causes of concern. *Biodiversity and Conservation* 14(2), 429–439.

Lorenzen, E., Heller, R., and Siegismund, H. R. (2012). Comparative phylogeography of African savannah ungulates. *Molecular Ecology* 21(15), 3656–3670.

Lovell, P. V., Wirthlin, M., Wilhelm, L., Minx, P., Lazar, N. H., Carbone, L., Warren, W. C., and Mello, C. V. (2014). Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biology* 15(12), 565.

Loveridge, A. (1932). New races of a skink (*Siaphos*) and frog (*Xenopus*) from the Uganda Protectorate. *Proceedings of the Biological Society of Washington* 45, 113–116.

Loveridge, A. (1925). Notes on East African Batrachians, collected 1920–1923, with the Description of four new Species. *Journal of Zoology* 95(2), 763–791.

Loveridge, A. (1933). *Reports on the Scientific Results of an Expedition to the Southwestern Highlands of Tanganyika Territory: Herpetology. VII.* Museum.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., Buckler, E. S., and Costich, D. E. (2013). Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics* 9(1), e1003215.

Lynch, M. and Conery, J. S. (Nov. 2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science* 290(5494), 1151–1155.

Mable, B. (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society* 82(4), 453–466.

Mable, B., Alexandrou, M., and Taylor, M. (2011). Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology* 284(3), 151–182.

Maddison, W. P. and Maddison, D. R. (2015). *Mesquite a modular system for evolutionary analysis*. `http://mesquiteproject.org`. Version 3.04.

Maisey, J. G. (2000). Continental break up and the distribution of fishes of Western Gondwana during the Early Cretaceous. *Cretaceous Research* 21(2-3), 281–314.

Makino, T. and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences* 107(20), 9270–9274.

Makova, K. D. and Li, W.-H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416(6881), 624–626.

Malone, J. H., Chrzanowski, T. H., and Michalak, P. (2007). Sterility and gene expression in hybrid males of *Xenopus laevis* and *X. muelleri*. *PLoS One* 2(8), e781.

Mank, J. E. (2009a). The evolution of heterochiasmy: the role of sexual selection and sperm competition in determining sex-specific recombination rates in eutherian mammals. *Genetics Research* 91(5), 355–363.

Mank, J. E. (2009b). The W, X, Y and Z of sex-chromosome dosage compensation. *Trends in Genetics* 25(5), 226–233.

Mank, J. E., Nam, K., and Ellegren, H. (2009). Faster-Z evolution is predominantly due to genetic drift. *Molecular Biology and Evolution* 27(3), 661–670.

Mank, J. E., Promislow, D. E., and Avise, J. C. (2006). Evolution of alternative sex-determining mechanisms in teleost fishes. *Biological Journal of the Linnean Society* 87(1), 83–93.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* 13(8), e1002220.

Margarido, G. R. A., de Souza, A. P., and Garcia, A. A. F. (2007). OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144, 78–79.

Marri, P. R., Hao, W., and Golding, G. B. (2006). Gene Gain and Gene Loss in *Streptococcus*: Is It Driven by Habitat? *Molecular Biology and Evolution* 23(12), 2379–2391.

Marsh, D. M. and Trenham, P. C. (2001). Metapopulation dynamics and amphibian conservation. *Conservation Biology* 15(1), 40–49.

Matsubara, K., Tarui, H., Toriba, M., Yamada, K., Nishida-Umehara, C., Agata, K., and Matsuda, Y. (Nov. 2006). Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proceedings of the National Academy of Science* 103(48), 18190–18195.

Matsuda, M., Nagahama, Y., Shinomiya, A., Sato, T., Matsuda, C., Kobayashi, T., Morrey, C. E., Shibata, N., Asakawa, S., Shimizu, N., Hori, H., Hamaguchi, S., and Sakaizumi, M. (May 2002). *DMY* is a Y-specific DM-domain gene required for male development in the medaka fish. en. *Nature* 417(6888), 559–563.

Matsuda, Y., Uno, Y., Kondo, M., Gilchrist, M. J., Zorn, A. M., Rokhsar, D. S., Schmid, M., and Taira, M. (2015). A new nomenclature of *Xenopus laevis* chromosomes based on the phylogenetic relationship to *Silurana/Xenopus tropicalis*. *CytoGenetics Genome Research* 145(3-4), 187–191.

Matzuk, M. M. and Lamb, D. J. (Nov. 2008). The biology of infertility: research advances and clinical challenges. *Nature Medicine* 14(11), 1197–1213.

Mawaribuchi, S., Takahashi, S., Wada, M., Uno, Y., Matsuda, Y., Kondo, M., Fukui, A., Takamatsu, N., Taira, M., and Ito, M. (2016). Sex chromosome differentiation and the W-and Z-specific loci in *Xenopus laevis*. *Developmental Biology*.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226 (4676), 792–801.

McCoid, M. J. and Fritts, T. H. (1980). Observations of feral populations of *Xenopus laevis* (Pipidae) in southern California. *Bulletin of the Southern California Academy of Sciences* 79(2), 82–86.

McCoid, M. J. and Fritts, T. H. (1989). Growth and fatbody cycles in feral populations of the African clawed frog, *Xenopus laevis* (Pipidae), in California with comments on reproduction. *The Southwestern Naturalist*, 499–505.

McGrath, C. L., Gout, J.-F., Johri, P., Doak, T. G., and Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research* 24(10), 1665–1675.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9), 1297–1303.

McLoughlin, S. (2001). The breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Journal of Botany* 49(3), 271–300.

Measey, G. J. and Tinsley, C. (1998). Feral *Xenopus laevis* in south Wales. *Herpetological Journal* 8, 23–27.

Measey, G. J., Villiers, A. L. de, and Soorae, P. (2011). Conservation introduction of the Cape platanna within the Western Cape, South Africa. *Global Re-introduction Perspectives*, 91–93.

Measey, G. and Channing, A. (2003). Phylogeography of the genus *Xenopus* in southern Africa. *Amphibia-Reptilia* 24(3), 321–330.

Measey, G., Rödder, D., Green, S., Kobayashi, R., Lillo, F., Lobos, G., Rebelo, R., and Thirion, J.-M. (2012). Ongoing invasions of the African clawed frog, *Xenopus laevis*: a global review. *Biological Invasions* 14(11), 2255–2270.

Measey, J., Davies, S., Vimercati, G., Rebelo, A., Schmidt, W., and Turner, A. A. (2017). Invasive amphibians in southern Africa: a review of invasion pathways. *Bothalia-Applied Biodiviversity and Conservation* 47(2), a2117.

Mikamo, K. and Witschi, E. (1966). The mitotic chromosomes in *Xenopus laevis* (Daudin): normal, sex reversed and female WW. *Cytogenetics* 5(1), 1–19.

Mitchell, D., Coley, P., Webb, S., and Allsopp, N. (1986). Litterfall and decomposition processes in the coastal fynbos vegetation, south-western Cape, South Africa. *The Journal of Ecology*, 977–993.

Miura, I. (2008). An evolutionary witness: the frog *Rana rugosa* underwent change of heterogametic sex from XY male to ZW female. *Sex Dev* 1(6), 323–331.

Morin, R. D., Chang, E., Petrescu, A., Liao, N., Griffith, M., Kirkpatrick, R., Butterfield, Y. S., Young, A. C., Stott, J., Barber, S., et al. (2006). Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. *Genome Research* 16(6), 796–803.

Muller, H. J. (1932). Some genetic aspects of sex. *The American Naturalist* 66(703), 118–138.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1(1), 2–9.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* 403(6772), 853–858.

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., and Sakaizumi, M. (2012). Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*. *Genetics* 191(1), 163–170.

Myosho, T., Takehana, Y., Hamaguchi, S., and Sakaizumi, M. (2015a). Turnover of sex chromosomes in celebensis group medaka fishes. *G3: Genes|Genomes|Genetics* 5(12), 2685–2691.

Myosho, T., Takehana, Y., Hamaguchi, S., and Sakaizumi, M. (2015b). Turnover of sex chromosomes in celebensis group medaka fishes. *G3: Genes|Genomes|Genetics* 5(12), 2685–2691.

Natri, H. M., Shikano, T., and Merilä, J. (2013). Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Molecular Biology and Evolution* 30(5), 1131–1144.

Navarro, A. and Ruiz, A. (1997). On the fertility effects of pericentric inversions. *Genetics* 147(2), 931–933.

Nietlisbach, P., Camenisch, G., Bucher, T., Slate, J., Keller, L. F., and Postma, E. (2015). A microsatellite-based linkage map for song sparrows (*Melospiza melodia*). *Molecular Ecology Resources* 15(6), 1486–1496.

Nylander, J. (2004). *MrModeltest v2 Distributed by Author. Evolutionary Biology Center, Uppsala University.*

O'Meally, D., Ezaz, T., Georges, A., Sarre, S. D., and Graves, J. A. (Jan. 2012). Are some chromosomes particularly good at sex? Insights from amniotes. *Chromosome Research* 20(1), 7–19.

Ohno, S. (1970). *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Ohno, S. (1966). *Sex chromosomes and sex-linked genes.* Vol. 1. Springer-Verlag Berlin Heidelberg.

Olmstead, A. W., Lindberg-Livingston, A., and Degitz, S. J. (June 2010). Genotyping sex in the amphibian, *Xenopus (Silurana) tropicalis*, for endocrine disruptor bioassays. *Aquatic Toxicology* 98(1), 60–6.

Ortiz-Barrientos, D., Engelstädter, J., and Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends in Ecology & Evolution* 31(3), 226–236.

Otto, S. P. and Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nature Reviews Genetics* 3(4), 252.

Otto, S. P. and Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics* 34(1), 401–437.

Ottolini, C. S., Newnham, L. J., Capalbo, A., Natesan, S. A., Joshi, H. A., Cimadomo, D., Griffin, D. K., Sage, K., Summers, M. C., Thornhill, A. R., et al. (2015). Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nature Genetics* 47(7), 727–735.

Pante, E. and Simon-Bouhet, B. (2013). Marmap: a package for importing, plotting and analyzing bathymetric and topographic data in R. *PLoS One* 8(9), e73051.

Papp, B., Pal, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945), 194–197.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26, 419–420.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290.

Parker, H. W. (1936a). Reptiles and Amphibians collected by the Lake Rudolf Rift Valley Expedition, 1934. *Journal of Natural History* 18(108), 594–609.

Parker, H. W. (1936b). Reptiles and Amphibians collected by the Lake Rudolf Rift Valley Expedition, 1934. *Journal of Natural History* 18(108), 594–609.

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K. C., Shu, S., Udall, J., et al. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* 492(7429), 423–427.

Perret, Jean-Luc (1966). Les amphibiens du Cameroun. *Zoologische Jahrbücher. Abteilung für Systematik, Ökologie und Geographie*, 289–464.

Perrin, N. (2009). Sex reversal: a fountain of youth for sex chromosomes? *Evolution* 63(12), 3043–3049.

Peters, W. C. H. (1844). Über einige neue Fische und Amphibien aus Angola und Mozambique. *Monatsberichte der Königlichen Preussische Akademie des Wissenschaften zu Berlin* 1844, 32–37.

Picker, M. (1980). *Xenopus laevis* (Anura: Pipidae) mating systems: a preliminary synthesis with some data on the female phonoresponse. *African Zoology* 15(3), 150–158.

Picker, M., Harrison, J., and Wallace, D. (1996). Natural hybridization between *Xenopus laevis laevis* and *X. gilli* in the south-western Cape Province, South Africa. In: *The Biology of Xenopus*. Zoological Society of London.

Picker, M., McKenzie, C., and Fielding, P. (1993). Embryonic tolerance of *Xenopus* (Anura) to acidic blackwater. *Copeia*, 1072–1081.

Picker, M. D. (1985). Hybridization and habitat selection in *Xenopus gilli* and *Xenopus laevis* in the south-western Cape Province. *Copeia*, 574–580.

Picker, M. D. and Villiers, A. L. de (1989). The distribution and conservation status of *Xenopus gilli* (Anura: Pipidae). *Biological Conservation* 49(3), 169–183.

Pickersgill, M. (2007). Frog Search, Results of Expeditions to Southern and Eastern Africa. *Edition Chimaira, Frankfurt am Main*.

Piganeau, G., Gardner, M., and Eyre-Walker, A. (2004). A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution* 21(12), 2319–2325.

Pitman III, W., Cande, S., LaBrecque, J., and Pindell, J. (1993). Fragmentation of Gondwana: the separation of Africa from South America. In: *Biological relationships between Africa and South America*, 15–34.

Pokorna, M. and Kratochvíl, L. (2009). Phylogeny of sex-determining mechanisms in squamate reptiles: are sex chromosomes an evolutionary trap? *Zoological Journal of the Linnean Society* 156(1), 168–183.

Poynton, J. C. and Broadley, D. G. (1985). Amphibia Zambesiaca 1. Scolecomorphidae, Pipidae, Microhylidae, Hemisidae, Arthroleptidae. *Annals of the Natal Museum* 26(2), 503–553.

Poynton, J. C. (1964). Amphibia of southern Africa; a faunal study. *Annals of the Natal Museum* 17, 1–334.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155(2), 945–959.

Qian, W., Liao, B.-Y., Chang, A. Y.-F., and Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics* 26(10), 425–430.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6), 841–842.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J. (2014). *Tracer v1.6*. `http://beast.bio.ed.ac.uk/Tracer`.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One* 6(9), e22594.

Rau, R. E. (1978). The development of *Xenopus gilli* Rose & Hewitt (Anura, Pipidae). *The Annals of the South African Museums* 76(2), 247–263.

Renny-Byfield, S., Gong, L., Gallagher, J. P., and Wendel, J. F. (2015). Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Molecular Biology and Evolution* 32(4), 1063–1071.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3, 217–223.

Rhymer, J. M. and Simberloff, D. (1996). Extinction by hybridization and introgression. *annualreviews Reviews of Ecology and Systematics*, 83–109.

Rice, W. R. (1987). The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution* 41(4), 911–914.

Rice, W. R. et al. (1994). Degeneration of a nonrecombining chromosome. *Science* 263(5144), 230–231.

Riddle, N. C. and Birchler, J. A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends in Genetics* 19(11), 597–600.

Rieseberg, L. H. (2001). Polyploid evolution: Keeping the peace at genomic reunions. *Current Biology* 11(22), R925–R928.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J. L., Schwarzbach, A. E., Donovan, L. A., and Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* 301(5637), 1211–1216.

Robert, J., Abramowitz, L., Gantress, J., and Morales, H. D. (2007). *Xenopus laevis*: a possible vector of ranavirus infection? *Journal of Wildlife Diseases* 43(4), 645–652.

Roberts, R. B., Ser, J. R., and Kocher, T. D. (Nov. 2009). Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. *Science* 326(5955), 998–1001.

Rocha, E. P. (2006). The quest for the universals of protein evolution. *Trends in Genetics* 22(8), 412–416.

Roco, Á. S., Olmstead, A. W., Degitz, S. J., Amano, T., Zimmerman, L. B., and Bullejos, M. (2015). Coexistence of Y, W, and Z sex chromosomes in *Xenopus tropicalis*. *Proceedings of the National Academy of Sciences* 112(34), E4752–E4761.

Roelants, K., Gower, D. J., Wilkinson, M., Loader, S. P., Biju, S., Guillaume, K., Moriau, L., and Bossuyt, F. (2007). Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences* 104(3), 887–892.

Rose, W. and Hewitt, J. (1926). Description of a new species of *Xenopus* from the Cape peninsula. *Transscriptions of the Royal Society of South Africa* 14(3), 343–346.

Ross, J. A., Urton, J. R., Boland, J., Shapiro, M. D., and Peichel, C. L. (Feb. 2009). Turnover of sex chromosomes in the stickleback fishes (gasterosteidae). *PLoS Genetics* 5(2), e1000391.

Roure, B., Baurain, D., and Philippe, H. (2012). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution* 30(1), 197–214.

Rovatsos, M., Vukić, J., and Kratochvíl, L. (2016). Mammalian X homolog acts as sex chromosome in lacertid lizards. *Heredity*.

Rovatsos, M., Vukić, J., Lymberakis, P., and Kratochvíl, L. (2015). Evolutionary stability of sex chromosomes in snakes. In: *Proceedings of the Royal Society of London B: Biological Sciences*. Vol. 282. 1821. The Royal Society, 20151992.

Ruan, X., Wang, W., Kong, J., Yu, F., and Huang, X. (2010). Genetic linkage mapping of turbot (*Scophthalmus maximus L.*) using microsatellite markers and its application in QTL analysis. *Aquaculture* 308(3), 89–100.

Salmon, A., Ainouche, M. L., and Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular ecology* 14(4), 1163–1175.

Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S., and Wolfe, K. H. (Mar. 2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440(7082), 341–345.

Scannell, D. R. and Wolfe, K. H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* 18(1), 137–147.

Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics* 27(4), 592–593.

Schmid, M., Sims, S., Haaf, T., and Macgregor, H. (1986). Chromosome banding in amphibia. X: 18S and 28S ribosomal RNA genes, nucleolus organizers and nucleoli in Gastrotheca riobambae. *Chromosoma* 94(2), 139–145.

Schmid, M., Evans, B. J., and Bogart, J. P. (2015). Polyploidy in amphibia. *Cytogenetic and genome research* 145(3-4), 315–330.

Schmidt, K. and Inger, R. (1949). Amphibians exclusive of the genera *Afrixalus* and *Hyperolius*. Exploration du Parc National de l'Upemba, Mission G.F. de Witte, en Collaboration avec W. Adam, A. Janssens, L. Van Meel, et R. Verheyen. *Institut des Parcs Nationnaux du Congo Belge* 56, 1–264.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82(398), 605–610.

Sémon, M. and Wolfe, K. H. (June 2008). Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. *Proceedings of the National Academy of Sciences* 105(24), 8333–8.

Seo, T.-K. (2008). Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molular Biology and Evolution*.

Sereno, P. C., Wilson, J. A., and Conrad, J. L. (2004). New dinosaurs link southern landmasses in the Mid–Cretaceous. *Proceedings of the Royal Society of London B: Biological Sciences* 271(1546), 1325–1330.

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538(7625), 336–343.

Shapiro, H. and Zwarenstein, H. (1934). A rapid test for pregnancy on *Xenopus laevis*. *Nature* 133(3368), 762.

Shaw, T. I., Ruan, Z., Glenn, T. C., and Liu, L. (July 2013). STRAW: Species TRee Analysis Web server. *Nucleic Acids Research* 41(Web Server issue), W238–41.

Shen, J. J., Dushoff, J., Bewick, A. J., Chain, F. J., and Evans, B. J. (2013). Genomic dynamics of transposable elements in the western clawed frog (*Silurana tropicalis*). *Genome Biology and Evolution* 5(5), 998–1009.

Simmonds, M. (1985). Interactions between *Xenopus* species in the southwestern Cape Province, South-Africa. In: *South African Journal of Science*. Vol. 81. 4. Academy of South Africa, 200–200.

Simonti, C. N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D. S., Chisholm, R. L., Crosslin, D. R., Hebbring, S. J., Jarvik, G. P., Kullo, I. J., Li, R., Pathak, J., Ritchie, M. D., Roden, D. M., Verma, S. S., Tromp, G., Prato, J. D., Bush, W. S., Akey, J. M., Denny, J. C., and Capra, J. A. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. *Science* 351(6274), 737–741.

Smith, C. A., Roeszler, K. N., Ohnesorg, T., Cummins, D. M., Farlie, P. G., Doran, T. J., and Sinclair, A. H. (2009). The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. *Nature* 461(7261), 267–271.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* 23(1), 23–35.

South African Frog Re-assessment Group (SA-FRoG), IUCN SSC Amphibian Specialist Group. 2010. Xenopus gilli. The IUCN Red List of Threatened Species 2010: e.T23124A9417597 (n.d.). `http://dx.doi.org/10.2305/IUCN.UK.2004.RLTS.T23124A9417597.en`. Accessed: 2016-09-29.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4), 583–639.

Stamatakis, A. (2014a). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9), 1312–1313.

Stamatakis, A. (May 2014b). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9), 1312–3.

Steige, K. A. and Slotte, T. (2016). Genomic legacies of the progenitors and the evolutionary consequences of allopolyploidy. *Current Opinion in Plant Biology* 30, 88–93.

Stelkens, R., Brockhurst, M., Hurst, G., Miller, E., and Greig, D. (2014). The effect of hybrid transgression on environmental tolerance in experimental yeast crosses. *Journal of Evolutionary Biology* 27(11), 2507–2519.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics* 73(5), 1162–1169.

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics* 76(3), 449–462.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* 68(4), 978–989.

Stevens, N. M. (1905). *Studies in spermatogenesis with special reference to the accessory chromosome.*

Stevison, L. S., Hoehn, K. B., and Noor, M. A. (2011). Effects of inversions on within- and between-species recombination and divergence. *Genome Biology and Evolution* 3, 830–841.

Stöck, M., Savary, R., Betto-Colliard, C., Biollay, S., Jourdan-Pineau, H., and Perrin, N. (2013). Low rates of X-Y recombination, not turnovers, account for homomorphic sex chromosomes in several diploid species of Palearctic green toads (*Bufo viridis* subgroup). *Journal of Evolutionary Biology* 26(3), 674–682.

Stöck, M., Horn, A., Grossen, C., Lindtke, D., Sermier, R., Betto-Colliard, C., Dufresnes, C., Bonjour, E., Dumas, Z., Luquet, E., et al. (2011). Ever-young sex chromosomes in European tree frogs. *PLoS Biology* 9(5), e1001062.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* 49(2), 169–181.

Stukenbrock, E. H. (2016). The Role of Hybridization in the Evolution and Emergence of New Fungal Plant Pathogens. *Phytopathology* 106(2), 104–112.

Sun, S., Evans, B., and Golding, G. (2011). Heterotachy and false positives for recombination in animal mtDNA. *Molecular Biology and Evolution* 28, 2549–2559.

Sun, Y., Svedberg, J., Hiltunen, M., Corcoran, P., and Johannesson, H. (2017). Large-scale suppression of recombination predates genomic rearrangements in *Neurospora tetrasperma*. *Nature Communications* 8(1), 1140.

Sutou, S., Mitsui, Y., and Tsuchiya, K. (2001). Sex determination without the Y Chromosome in two Japanese rodents *Tokudaia osimensis osimensis* and *Tokudaia osimensis* spp. *Mammalian Genome* 12(1), 17–21.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3), 585–595.

Takehana, Y., Matsuda, M., Myosho, T., Suster, M. L., Kawakami, K., Shin-I, T., Kohara, Y., Kuroki, Y., Toyoda, A., Fujiyama, A., Hamaguchi, S., Sakaizumi, M., and Naruse, K. (2014). Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish *Oryzias dancena*. *Nature Communications* 5, 4157.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* 17(2), 57–86.

Tinsley, R. C. and Kobel, H. R., eds. (1996). *The Biology of Xenopus*. Zoological Society of London, 440.

Tinsley, R., Loumont, C., and Kobel, H. (1996). Geographical distribution and ecology. In: *The Biology of Xenopus*. 68. Zoological Society of London.

Tinsley, R. and McCoid, M. (1996). Feral populations of *Xenopus* outside Africa. In: *The Biology of Xenopus*. 68. Zoological Society of London, 1960–1999.

Tinsley, R. C. (1981). Interactions between *Xenopus* species (Anura Pipidae). *Monitore Zoologico Italiano. Supplemento* 15(1), 133–150.

Tobias, M. L., Barnard, C., O'Hagan, R., Horng, S. H., Rand, M., and Kelley, D. B. (2004). Vocal communication between male *Xenopus laevis*. *Animal Behaviour* 67(2), 353–365.

Tobias, M. L., Viswanathan, S. S., and Kelley, D. B. (1998). Rapping, a female receptive call, initiates male–female duets in the South African clawed frog. *Proceedings of the National Academy of Sciences* 95(4), 1870–1875.

Tobias, M., Evans, B. J., and Kelley, D. B. (2011). Evolution of advertisement calls in African clawed frogs. *Behaviour* 148(4), 519–549.

Tsaousis, A. D., Martin, D., Ladoukakis, E., Posada, D., and Zouros, E. (2005). Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution* 22(4), 925–933.

Tymowska, J. (1991). Polyploidy and cytogenetic variation in frogs of the genus *Xenopus*. In: *Amphibian Cytogenetics and Evolution*. Ed. by D. M. Green and S. K. Sessions. Academic Press San Diego, CA, 259–297.

Tymowska, J. and Fischberg, M. (1973). Chromosome complements of the genus *Xenopus*. *Chromosoma* 44(3), 335–342.

Uno, Y., Nishida, C., Oshima, Y., Yokoyama, S., Miura, I., Matsuda, Y., and Nakamura, M. (2008). Comparative chromosome mapping of sex-linked genes and identification of sex chromosomal rearrangements in the Japanese wrinkled frog (*Rana rugosa*, Ranidae) with ZW and XY sex chromosome systems. *Chromosome Research* 16(4), 637–647.

Uno, Y., Nishida, C., Takagi, C., Ueno, N., and Matsuda, Y. (2013). Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. *Heredity* 111(5), 430–6.

Uno, Y., Nishida, C., Takagi, C., Igawa, T., Ueno, N., Sumida, M., and Matsuda, Y. (2015). Extraordinary Diversity in the Origins of Sex Chromosomes in Anurans Inferred from Comparative Gene Mapping. *CytoGenetics Genome Research* 145(3-4), 218–229.

Van Doorn, G. S. and Kirkpatrick, M. (Oct. 2007). Turnover of sex chromosomes induced by sexual conflict. *Nature* 449(7164), 909–912.

Venn, O., Turner, I., Mathieson, I., Groot, N. de, Bontrop, R., and McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science* 344(6189), 1272–1275.

Veyrunes, F., Waters, P. D., Miethke, P., Rens, W., McMillan, D., Alsop, A. E., Grutzner, F., Deakin, J. E., Whittington, C. M., Schatzkamer, K., Kremitzki, C. L., Graves, T., Ferguson-Smith, M. A., Warren, W., and Marshall Graves, J. A. (June 2008). Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Research* 18(6), 965–973.

Vicoso, B. and Bachtrog, D. (2009). Progress and prospects toward our understanding of the evolution of dosage compensation. *Chromosome Research* 17(5), 585–602.

Vicoso, B., Kaiser, V. B., and Bachtrog, D. (2013). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proceedings of the National Academy of Sciences* 110(16), 6453–6458.

Vigny, C. (1979). The mating calls of 12 species and sub-species of the genus *Xenopus* (Amphibia: Anura). *Journal of Zoology* 188(1), 103–122.

Villiers, A. de (2004). Species account: *Xenopus gilli* (Rose & Hewitt, 1927). In: *Atlas and red data book of the frogs of South Africa, Lesotho and Swaziland.* Ed. by L. Minter, M. Burger, J. Harrison, P. Bishop, and H. Braack. Smithsonian Institution Press, 260–263.

Villiers, F. A. de, Kock, M. de, and Measey, G. J. (2016). Controlling the African clawed frog *Xenopus laevis* to conserve the Cape platanna *Xenopus gilli* in South Africa. *Conservation Evironment* 13, 17.

Vogt, S., Villiers, F. A. de, Ihlow, F., D, R., and Measey, J. (2017). Competition and feeding ecology in two sympatric *Xenopus* species (Anura: Pipidae). *PeerJ* 5, e3130.

Wagenmakers, E.-J. and Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* 11(1), 192–196.

Wagner, C. A., Friedrich, B., Setiawan, I., Lang, F., and Bröer, S. (2000). The use of Xenopus laevis oocytes for the functional characterization of heterologously expressed membrane proteins. *Cellular Physiology and Biochemistry* 10(1-2), 1–12.

Wallis, M. C., Waters, P. D., Delbridge, M. L., Kirby, P. J., Pask, A. J., Grutzner, F., Rens, W., Ferguson-Smith, M. A., and Graves, J. A. (2007). Sex determination in platypus and echidna: autosomal location of *SOX3* confirms the absence of *SRY* from monotremes. *Chromosome Research* 15(8), 949–959.

Wang, Y., Jha, A. K., Chen, R., Doonan, J. H., and Yang, M. (2010). Polyploidy-associated genomic instability in *Arabidopsis thaliana*. *Genesis* 48(4), 254–263.

Watanabe, T., Takeda, A., Mise, K., Okuno, T., Suzuki, T., Minami, N., and Imai, H. (2005). Stage-specific expression of microRNAs during Xenopus development. *FEBS letters* 579(2), 318–324.

Weisman, A. I. and Coates, C. W. (1941). The frog test (*Xenopus laevis*), as a rapid diagnostic test for early pregnancy. *Endocrinology* 28(1), 141–142.

Weiss, G. and Haeseler, A. von (1998). Inference of population history using a likelihood approach. *Genetics* 149(3), 1539–1546.

Weiss, J., Meeks, J. J., Hurley, L., Raverot, G., Frassetto, A., and Jameson, J. L. (2003). *Sox3* is required for gonadal function, but not sex determination, in males and females. *Molular Cell Biology* 23(22), 8084–8091.

Weldon, C., Du Preez, L. H., Hyatt, A. D., Muller, R., and Speare, R. (2004). Origin of the amphibian chytrid fungus. *Emerging Infectious Diseases* 10(12), 2100.

Wells, D. E., Gutierrez, L., Xu, Z., Krylov, V., Macha, J., Blankenburg, K. P., Hitchens, M., Bellot, L. J., Spivey, M., Stemple, D. L., Kowis, A., Ye, Y., Pasternak, S., Owen, J., Tran, T., Slavikova, R., Tumova, L., Tlapakova, T., Seifertova, E., Scherer, S. E., and Sater, A. K. (June 2011). A genetic map of *Xenopus tropicalis*. *Developmental Biology* 354(1), 1–8.

Wilson, J. R., Gaertner, M., Griffiths, C. L., Kotzé, I., Le Maitre, D., Marr, S., Picker, M. D., Spear, D., Stafford, L., Richardson, D., et al. (2014). Biological invasions in the Cape Floristic Region: history, current patterns, impacts, and management challenges. *Fynbos: Ecology, Evolution, and Conservation of a Megadiverse Region*, 273.

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

Wong, A. K., Ruhe, A. L., Dumont, B. L., Robertson, K. R., Guerrero, G., Shull, S. M., Ziegle, J. S., Millon, L. V., Broman, K. W., Payseur, B. A., et al. (2010). A comprehensive linkage map of the dog genome. *Genetics* 184(2), 595–605.

Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences* 111(14), 5283–5288.

Wright, A. E., Darolti, I., Bloch, N. I., Oostra, V., Sandkam, B., Buechel, S. D., Kolm, N., Breden, F., Vicoso, B., and Mank, J. E. (2017). Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature Communications* 8, 14251.

Wright, A. E., Dean, R., Zimmer, F., and Mank, J. E. (2016). How to make a sex chromosome. *Nature Communications* 7.

Wu, R., Ma, C.-X., Painter, I., and Zeng, Z.-B. (2002). Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. *Theoretical Population Biology* 61(3), 349–363.

Wyman, M. J., Cutter, A. D., and Rowe, L. (2012). Gene duplication in the evolution of sexual dimorphism. *Evolution* 66(5), 1556–1566.

Yager, D. D. (1996). Sound production and acoustic communication in *Xenopus borealis*. In: *Symposia of the Zoological Society of London*. 68. London: The Society, 1960-1999.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in Bioscience* 13(5), 555–556.

Yang, Z. (Aug. 2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24(8), 1586–91.

Yang, Z. (2014). *Molecular Evolution A Statistical Approach*. Oxford, UK: Oxford University Press.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* 17(1), 32–43.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* 107(20), 9264–9269.

Yazdi, H. P. and Ellegren, H. (2014). Old but not (so) degenerated—slow evolution of largely homomorphic sex chromosomes in ratites. *Molecular Biology and Evolution*, msu101.

Yoshimoto, S., Ikeda, N., Izutsu, Y., Shiba, T., Takamatsu, N., and Ito, M. (2010). Opposite roles of DMRT1 and its W-linked paralogue, DM-W, in sexual dimorphism of Xenopus laevis: implications of a ZZ/ZW-type sex-determining system. *Development* 137(15), 2519–2526.

Yoshimoto, S., Okada, E., Umemoto, H., Tamura, K., Uno, Y., Nishida-Umehara, C., Matsuda, Y., Takamatsu, N., Shiba, T., and Ito, M. (2008). A W-linked DM-domain gene, *DM*-W, participates in primary ovary development in *Xenopus laevis*. *Proceedings of the National Academy of Sciences* 105(7), 2469–2474.

Zarkower, D. (2001). Establishing sexual dimorphism: conservation amidst diversity? *Nature Reviews Genetics* 2(3), 175–185.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16(7), 409–420.

Zhao, D., McBride, D., Nandi, S., McQueen, H., McGrew, M., Hocking, P., Lewis, P., Sang, H., and Clinton, M. (2010). Somatic sex identity is cell autonomous in the chicken. *Nature* 464(7286), 237–242.

Zhou, Q., Zhang, J., Bachtrog, D., An, N., Huang, Q., Jarvis, E. D., Gilbert, M. T. P., and Zhang, G. (2014). Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* 346(6215), 1246338.

Zickler, D. and Kleckner, N. (2016). A few of our favorite things: Pairing, the bouquet, crossover interference and evolution of meiosis. In: *Seminars in Cell & Developmental Biology*. Vol. 54. Elsevier, 135–148.