ABSTRACT

Title of Dissertation: THE DIVERSITY OF SEX CHROMOSOMES WITHIN AFRICAN CICHLIDS

William Jacob Gammerdinger, Doctor of Philosophy, 2018

Dissertation directed by: Professor Thomas Kocher
Department of Biology

Genetic sex-determination is one of the most prevalent systems by which the sex of an organism can be established. The genes that determine sex reside on chromosomes that experience a unique pattern of evolutionarily processes, which often leads to the degradation of genes surrounding these sex-determination loci. The widespread degradation of sex chromosomes has been noted in the relatively old and heteromorphic sex chromosomes of therian mammals and birds. However, with the advent of next-generation sequencing, it is now possible to study the earliest stages of sex chromosome evolution in relatively homomorphic sex chromosomes.

African cichlid fishes are a powerful model system for studying the early stages of sex chromosome evolution because of the diversity and young age of their sex chromosomes. This dissertation develops methods for studying young sex chromosomes and employs these approaches to evaluate the sex chromosomes within tilapia and Lake Tanganyika cichlid fishes. Furthermore, this research demonstrates a

method for identifying the ancestral state for species sharing a common sex chromosome system and a process for evaluating the functional significance of these shared mutations. Lastly, this dissertation proposes a mechanism for the diversity of sex chromosome systems observed within African cichlid fishes.

This analysis not only characterizes the decay of several known young sex chromosomes, but also reveals multiple previously undiscovered sex chromosome systems within the African cichlid fishes. These novel sex chromosome systems likely represent only a fraction of the true variety of sex chromosome systems within this group, and therefore push forward the argument for characterizing the sex chromosome systems of more cichlid species in order to better understand the early stages of sex chromosome evolution.

THE DIVERSITY OF SEX CHROMOSOMES WITHIN AFRICAN CICHLIDS

by

William Jacob Gammerdinger

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Thomas Kocher, Chair
Professor Karen Carleton
Professor Eric Haag
Professor Carlos Machado
Professor Antony Jose

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Linkage Group (LG)
Single nucleotide polymorphism (SNP)

Chapter 1: Introduction

Sex, whether male or female, is a phenotype that has a substantial influence on an organism throughout its life cycle. There is a large diversity of mechanisms through which sex can be determined for an individual and it is important to understand the function and evolution of these diverse mechanisms. A clearer understanding of these diverse mechanisms is important because permutations of these mechanisms can enhance our understanding of network and chromosome evolution as well as promote economic interests in aquaculture.

Sex-determination systems

The sex of an organism can be determined predominantly by the environment[1,2], by genetic mechanisms[3,4] or through an interplay of both[5,6]. Temperature is often a key determinant of sex within species whose sex is determined by the environment[1,2]. However, other factors can be social and include the presence of one sex affecting the sex of other individuals[7]. Genetic mechanisms within animals often appear in one of four varieties: *i*) an XX-XO system, *ii*) a genic balance system, *iii*) a XX-XY system or *iv*) a ZZ-ZW system. The XX-XO system has been recognized in grasshoppers and occurs when females have two copies of the X chromosome, while males only have one[8]. A genic balance system has been witnessed in *Drosophila* species and occurs due to the ratio of X chromosomes to autosomes[9]. This ratio ultimately determines sex in *Drosophila* through the alternatively splicing of the gene *doublesex*[10]. An XX-XY system is common throughout many therian mammals and occurs when females have two X chromosomes and males only have one X and one Y chromosome. In the case of

many therian mammals, the Y chromosome, containing the gene *SRY*, determines sex[4,11]. In such systems, females are referred to as the homogametic sex, while males are referred to as the heterogametic sex. ZZ-ZW systems are commonly found as well and occur when males are the homogametic sex and females are the heterogametic sex. In birds, it appears as that the presence of two Z chromosomes, containing the gene *DMRT1*, is key for sex-determination[12]. However, it is possible that the W chromosomes can have a dominant mechanism for sex-determination, similar to the Y chromosome in mammals, as seen by *DM-W* in *Xenopus laevis*[13]. The research reported hereafter will focus specifically on the sex chromosomes and candidate genes for sex-determination which reside on these sex chromosomes.

Sex chromosomes and evolution

Some sex chromosomes, like those observed in therian mammals, are distinct from autosomes, because they contain a master sex-determination gene. Master sex-determination genes are part of the sex-determination network and for a given master sex-determination gene in a population it is the only gene from that network that differs between males and females. Sex chromosomes were once autosomes that have been co-opted due to the emergence of a sex-determination gene within them. Once this novel sex chromosome emerges, sexually antagonistic genes in linkage with the novel sex-determination gene are selectively favored[14,15]. The idea that sexually antagonistic alleles drive the emergence of a novel sex chromosome is called the sexually antagonistic model for sex chromosome evolution[16]. Linkage between sexually antagonistic alleles and sex-determination genes allows for loci selectively favored in one sex and counter-selected in the opposite sex to be resolved through

linkage disequilibrium. However, through recombination these haplotypes can be broken up. Therefore, sexually antagonistic alleles can drive a further reduction in recombination by favoring structural rearrangements of the genomic landscape[15]. Such rearrangements are selectively favored due to the repression of recombination between the sex-determination locus and the sexually antagonistic locus[15]. Evidence from the therian Y chromosome suggests that there have been several separate inversion events, which presumably have been integral to the resolution of sexually antagonistic alleles located on the Y chromosome[17].

Mutations rapidly begin accruing once recombination has been selectively repressed within the region containing the sex-determination gene and sexually antagonistic alleles. Through genetic hitchhiking, rare and selectively advantageous mutations can help drive the fixation of deleterious alleles[18] and through Muller's Ratchet, mutations continue to accumulate and are only purged by rare reversion mutations or gene conversion[19]. The amassing mutations convert functional genes into non-functional ones and degrade the gene content of the newly born sex chromosome[19,20]. This model is known as the mutational load model for sex chromosome evolution[21].

Recently, these two models for the emergence of novel sex chromosomes, through sexually antagonism and mutational load, have been merged together under the 'hot potato hypothesis'[22]. Modeling suggests that the combination of these two models underlying the 'hot potato hypothesis' can give rise to indefinite cycling of sex chromosomes.

As these functional genes become non-functional, individuals of the homogametic sex now have two functional copies, while their heterogametic counterparts only have one functional and one non-functional copy. This produces homogametic-biased expression patterns until dosage compensation mechanisms arise and provide equal gene expression for the dosage sensitive genes on the sex chromosomes with respect to autosomes. To resolve this issue, humans evolved X-inactivation, which packages one female X chromosome in heterochromatin and thus halts most transcription from this X chromosome[23–31]. Alternatively, species of *Drosophila* resolve this issue by increasing expression of genes from the X chromosome in males[32–38].

Sex-determination mechanisms and the scale of the sex-determination network within Teleosts

Recent years have brought a number of questions concerning the evolution of sex-determination. Sex-determination through *SRY* in therian mammals[4] has been known for decades, however, a plethora of new genes have only recently been discovered. These genes include *DMRT1* in birds[12], *DM-W* in *Xenopus laevis*[13], *Dmy* in medaka[39,40], *gsdf* in *Oryzias luzonesis*[41], *amhy* in Patagonian pejerrey[42], *amhr2* in fugu[43], *sdY* in rainbow trout[44] and *amh* in tilapia[45,46]. One apparent trend of this recent research is the wide diversity of sex-determination mechanisms within the teleost linage. Below, I will briefly outline sex-determination within this lineage.

Dmy is the crux of the XX-XY sex-determination system in medaka[40] and it was the first of the teleost-specific sex-determination genes to be discovered. A duplication of *DMRT1*, called *Dmy*, was the only functional gene within the sex-

determining region and reported male-specific expression during development.[39,40]. Two mutants of *Dmy* from wild populations of medaka produced all XY female offspring[39]. One mutant had a premature stop codon, leaving *Dmy* non-functional[39]. While the other mutant did not disrupt the coding region, it substantially downregulated the transcription of *Dmy*.

O. luzonesis is a recently derived relative of medaka, however it does not share the same sex-determination gene. *Gsdf* (gonadal soma derived growth factor) controls an XX-XY sex-determination in *O. luzonesis*[41]. *Gsdf* was identified through the construction of physical maps across the sex-determination region and no large deletions or insertions were found[41]. Predicted genes within this region were identified and examined by qRT-PCR. Only one gene within this region, *gsdf*, demonstrated male-biased expression prior to gonad differentiation[41]. The amino acid sequence of the X and Y copy of *gsdf* were identical[41]. A transgenic line was developed that contained the Y-copy of *gsdf* along with a short upstream and downstream region[41]. This transgenic line showed upregulation of *gsdf* and was sufficient to create XX male sex reversal[41].

Patagonian pejerrey, *Odontesthes hatcheri*, also demonstrate a XX-XY sex-determination system[47]. An AFLP study found a SNP marker to be tightly linked with sex near *amh* (anti-Müllerian hormone)[47]. A follow-up study discovered two unique *amh* transcripts, which corresponded to a duplication event that created two *amh* loci[42]. One allele was male specific, *amhy*, and sex reversal of XY males into females resulted when it was knocked down through the use of morpholinos during development[42].

The Japanese pufferfish, *Takifugu rubripes*, also has an XX-XY sex-determination system[48] and is controlled by *amhr2* (anti-Müllerian Hormone receptor type II)[43]. Genetic mapping through breakpoint analysis was able to define a 17.5kb region within which *amhr2* and another gene were located[43]. Due to a lack of linkage disequilibrium surrounding the region, association mapping within this region from wild populations was carried out. It yielded only one SNP that was heterozygous in all males and absent in all females[43]. This SNP causes a H384D missense mutation[43].

Rainbow trout, *Oncorhynchus mykiss*, has an XX-XY sex-determination system regulated by *sdY*. Identification of *sdY* utilized RNA-seq, which searched for genes that demonstrated a male-biased expression pattern and mapped to the Y-specific region[44]. *sdY* appears to be a diverged duplicate of *irf9* (interferon regulatory factor 9)[44]. *sdY* was found to be male-specific and expressed highly during gonadal differentiation[44]. *sdY* was also only expressed in the testes of adults[44]. Transgenic lines of XX individual carrying *sdY* and a 5kb upstream promoter region were able to yield males capable of producing testes and sperm[44]. F_1 individuals with zinc-finger nuclease knockouts of *sdY* developed ovaries[44].

A recent review hypothesized that the diversity of sex-determination mechanisms from medaka, *O. luzonesis*, Patagonian pejerrey and fugu were all alterations to a common TGF-β signaling network, which controls germ cell proliferation during early development[48]. Critical evidence for this hypothesis was gained from the *hotei* mutant of *amhr2* in medaka[49]. *hotei* is an autosomal recessive mutation and females homozygous for the *hotei* mutation showed grossly enlarged ovaries[49]. Amongst males who were homozygous for *hotei*, half showed a male

phenotype and the other half showed sex-reversal with enlarged ovaries[49]. The study suggested that *amhr2* is critical for the regulation of the proliferation of germ cells[49]. This coincides with an observation from *O. luzonensis* that the first indicator of gonadal differentiation is that females have more germ cells than males during early development[50]. Therefore, it is hypothesized that alterations which downregulate germ cell proliferation repress ovarian development of the gonads in favor of testes[48]. However, it is unclear how *sdY* in rainbow trout ties into this view[48].

The only known sex-determination locus that has been connected with a gene within cichlids is the linkage group (LG) 23 XY locus. A mapping study was able to narrow the sex-determination locus to ~1.5Mb region on LG23 and within this region was *anti-müllerian hormone* (*amh*), which also showed a significant difference in expression during the critical window for sex-determination[51]. Subsequent sequencing of *amh* showed that there was a duplication of *amh* and the second version was called *amhy*[46]. An additional study followed this up and created CRISPR/Cas-9 knockouts for *amhy* and they were able to create a male-to-female sex reversal[45].

While these sex determination mechanisms represent nodes that can be modulated in the sex-determination network, they certainly do not represent an exhaustive list of all of the genes that are capable of being master sex-determination genes within teleost. A study that evaluated the sex-determination network in mammals, which presumably has a similar topology to the teleost sex-determination network, modeled the core sex-determination network using 21 genes in a multi-step Boolean model. While there are a few key differences between the known

mammalian and teleost networks, such as the presence of *Sry* in the mammalian network and the presence of *Gsdf* in the teleost network, the core elements of these networks likely have a common ancestral state and seem to have many of the same genes. This study also reviewed the literature for manipulations of these genes and found that many created sex-reversals in the mammalian system. Therefore, this understandably simplified model of the sex-determination network therefore sets a minimum number of sex-determination mechanisms likely just shy of two-dozen.

Sex chromosomes and sex-determination genes within African cichlids

The remarkable diversity of cichlids within the lakes of East Africa consists of more than 2,000 species radiating within the past 10 million years[52]. Furthermore, within the past couple million years, over 500 species of cichlids have radiated within Lake Malawi[52]. We observe an incredible diversity of sex chromosomes within these cichlids and those sex chromosomes, and the sex-determination systems that reside on them, have been hypothesized to be a driving force behind the expansive radiation[52]. In addition to the diversity of sex-determination mechanisms between species, there are also accounts of multiple sex-determination loci within single species[53,54].

Within the haplochromine clade, Lake Malawi cichlids appears to segregate at least three prominent separate sex-determination systems[54]. One is a ZZ-ZW system on linkage group 5 (LG5), another is a XX-XY on linkage group 7 (LG7), and a last one is an XX-XY system on linkage group 20 (LG20)[54–56]. The LG5 ZZ-ZW sex-determination system is segregating in at least four species of haplochromines[54]. This sex-determination locus is linked to the orange blotch (OB) locus[54,55]. OB is believed to be a locus under strong sexual antagonism[55]. In females, OB presumably confers a

cryptic coloration from predators, but in males OB is hypothesized to lower fitness due to female reliance on male color pattern for recognition[55]. Recent evidence indicates that the OB locus has resolved this sexual antagonism through an inversion that spans the OB locus and the sex-determination gene (Conte et al., *in preparation*). The LG7 XX-XY locus has been recorded in eight species within Lake Malawi cichlids, but the sex-determination gene(s) have yet to be elucidated. *Astatotilapia burtoni* lives in the tributaries to Lake Tanganyika and different studies have reported an XYW system on LG13, an XX-XY system on LG18 and an XX-XY system on a fusion between LG5 and LG14[57,58].

Within the Tilapiine tribe, three sex-determination regions have been identified. They are located on linkage group 1 (LG1), linkage group 3 (LG3) and linkage group 23 (LG23)[51,59]. Depending on the population studied, Nile tilapia, *Oreochromis niloticus*, has sex-determination loci mapping to LG1 or LG23[51,59]. The LG1 and LG23 loci are both XX-XY sex-determination systems[51,59]. As previously mentioned, *amhy* has been demonstrated to be the sex-determination gene at the LG23 locus[45,46,51]. In blue tilapia, *O. aureus*, the ZZ-ZW LG3 sex-determination gene is epistatically dominant to the LG1 XX-XY sex-determination gene[53]. In the Mozambique tilapia, *O. mossambicus*, sex-determination has also been mapped to the LG1 XX-XY or LG3 ZZ-ZW regions, depending on the strain[59]. Lastly, the spotted tilapia*, Pelmatolapia mariae*, has only the LG3 ZZ-ZW sex-determination region. The master sex-determination genes located on LG1 and LG3 have yet to be identified.

Interestingly, some regions where sex has been mapped within cichlids do not contain genes from the core vertebrate sex-determination network. Since there appear to be sex-linked regions that lack genes from the core vertebrate network, it seems reasonable to conclude that the sex-determination network within cichlids is likely larger than just the core sex-determination network in vertebrates and could actually be larger by several handfuls of genes. As a result, cichlids represent an excellent model system to elucidate other critical genes in the sex-determination network in teleosts. A clearer understanding of these genes will inform our knowledge of the potential size of the vertebrate sex-determination network and also understand lineage-specific differences between the structure of the network in mammals and teleosts.

The literature is sparse with regards to mapping sex outside of the haplochromines and tilapia. Efforts to map sex have been minimal, or non-existent, in many of the cichlid tribes of Lake Tanganyika (Figure 1). Understanding where sex is mapping in these tribes may provide more insight into the ancestral character state that led to the diverse array of sex chromosomes within the Haplochromines. Additionally, studies of the sex chromosomes in these species may also uncover novel sex-determination systems and increase our estimate as to the size of the sex-determination network in cichlids and vertebrates, in general. Therefore, studies of these Lake Tanganyikan cichlids hold tremendous potential to inform our understanding of not only the size of the sex-determination network but also help us better understand the transitions between sex-determination systems.

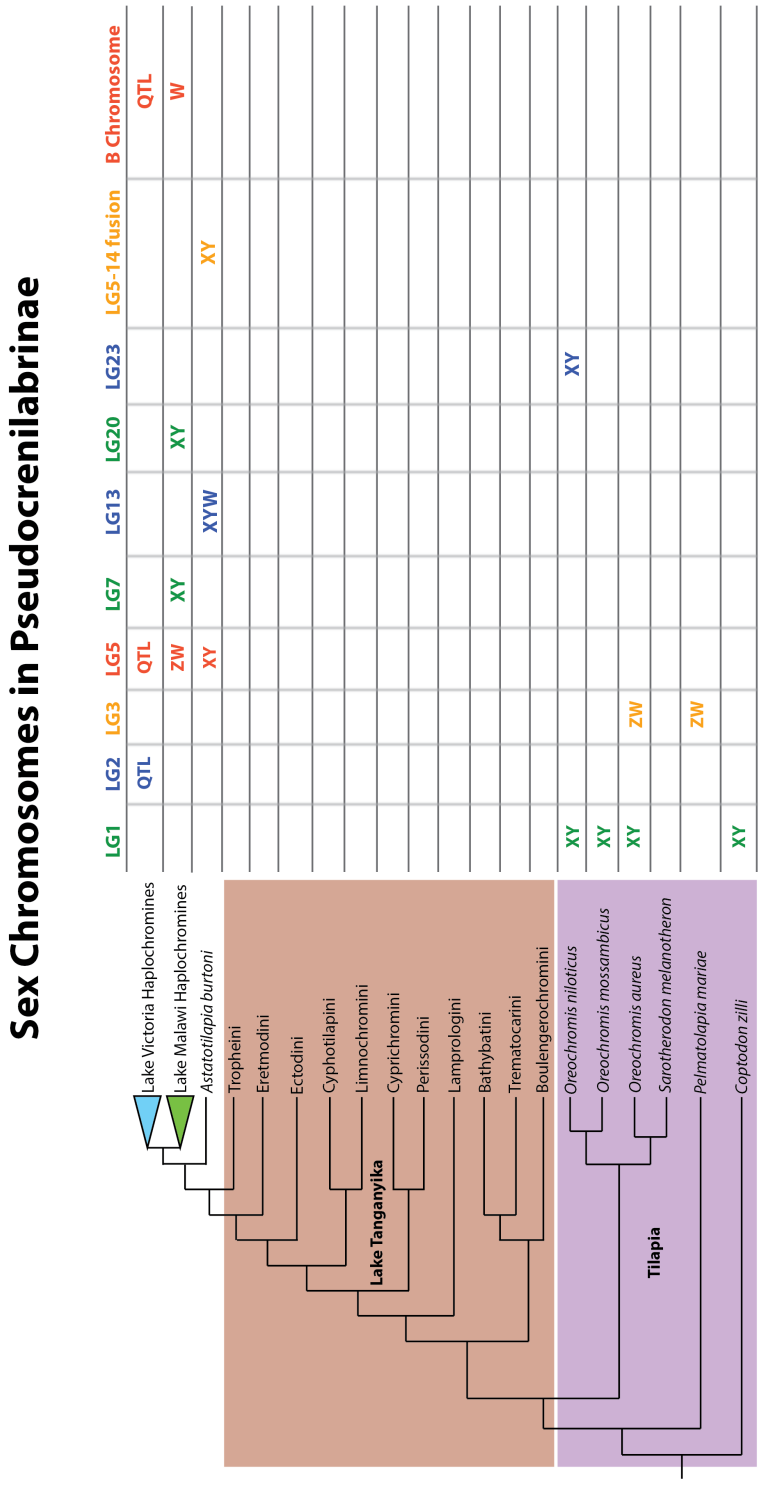


Figure 1. An overview of the diversity of known sex chromosome systems within Pseudocrenilabrinae[45,46,51,54–69].

Goals and broader impacts

The goal of this research is to identify and characterize sex chromosomes across the Pseudocrenilabrinae linage as well as identify possible sex-determination loci within sampled species of Pseudocrenilabrinae. The broader impacts of this research will shed insight into chromosome evolution and allow us to look for trends in the structure and decay of non-recombining chromosomes. Additionally, it will provide empirical data points to understand the very earliest stages of sex chromosome evolution. It will also provide greater insight into the size of sex-determination network by providing a natural mutant screen to look for possible alterations within the sex-determination network.

Chapter 2: Structure and Decay of a Proto-Y Region in Tilapia, *Oreochromis niloticus*

Previous published as: Gammerdinger, W. J., Conte, M. A., Acquah, E. A., Roberts, R. B., Kocher, T. D. "Structure and decay of a proto-Y region in Tilapia, *Oreochromis niloticus*." BMC Genomics 15.1 (2014): 975.

Abstract

Sex-determination genes drive the evolution of adjacent chromosomal regions. Sexually antagonistic selection favors the accumulation of inversions that reduce recombination in regions adjacent to the sex-determination gene. Once established, the clonal inheritance of sex-linked inversions leads to the accumulation of deleterious alleles, repetitive elements and a gradual decay of sex-linked genes. This in turn creates selective pressures for the evolution of mechanisms that compensate for the unequal dosage of gene expression. Here we use whole genome sequencing to characterize the structure of a young sex chromosome and quantify sex-specific gene expression in the developing gonad. We found an 8.8Mb block of strong differentiation between males and females that corresponds to the location of a previously mapped sex determiner on linkage group 1 of *Oreochromis niloticus*. Putatively disruptive mutations are found in many of the genes within this region. We also found a significant female-enrichment in the expression of genes within the block of differentiation compared to those outside the block of differentiation. Eight candidate sex-determination genes were identified within this region. This study demonstrates a block of differentiation on linkage group 1, suggestive of an 8.8Mb inversion encompassing the sex-determining locus. The enrichment of female-biased

gene expression inside the proposed inversion suggests incomplete dosage compensation. This study helps establish a model for studying the early-to-intermediate stages of sex chromosome evolution.

Background

The classic model of sex chromosome evolution begins with the emergence of a new sex- determining gene on an autosome[70]. The new sex-determiner may be linked with genes experiencing sexually antagonistic selection. Selection favors mechanisms, such as chromosomal inversions, that reduce recombination between the sex-determination locus and sexually antagonistic genes[15,71]. The human sex chromosomes have undergone at least four such inversions, which may have limited recombination between the sex-determination locus and nearby sexually antagonistic genes[17].

Inversions create a clonally inherited chromosomal segment with a relatively small effective population size when compared to the rest of the genome[72]. Sex chromosomes therefore become a haven for deleterious mutations and repetitive elements that are difficult to purge. These deleterious mutations accumulate via Muller's Ratchet, as well as by hitchhiking with advantageous mutations[18,20]. Degradation of functional genes on the Y- or W-chromosome leaves the homogametic sex carrying two functional copies of a particular gene, while the heterogametic sex carries only one functional copy. Therefore, mechanisms are needed to maintain appropriate expression of dosage-sensitive genes on emerging sex chromosomes[73,74,10]. In mammals, global dosage compensation is accomplished

through X-inactivation[23,75]. However, in many species, dosage compensation is partial and the expression of many genes is not compensated[73].

Some sex-determining genes are conserved for long periods of time. An example is *Sry*, a gene which has controlled sex-determination in therian mammals for approximately 180 million years[4,18,30,76]. Other sex-determination genes hold sway for much shorter periods of time. There have been at least five transitions in the mechanism controlling sex-determination in rice fish (genus *Oryzias*) during the last 20 million years[48]. Similarly rapid rates of sex chromosome evolution have been identified among sticklebacks (Family Gasterosteidae)[77].

The evolution of new sex-determination genes may have contributed to the rapid radiation of African cichlid fishes[52]. Among the closely related haplochromine cichlids of Lake Malawi, sex-determination regions have been localized to linkage groups 3 (ZW), 5 (ZW), 7 (XY) and 20 (XY)[54–56]. Among tilapia cichlids, sex-determination regions have been localized to linkage groups 1 (XY), 3 (ZW) and 23 (XY)[51,59]. Multiple sex-determination genes often segregate within a single species[53,54]. The blue tilapia, *Oreochromis aureus*, segregates both an XY system on linkage group 1 and a ZW system on linkage group 3[53]. Some strains of the Nile tilapia, *O. niloticus*, have an XY system on linkage group 1, while others segregate an XY system on linkage group 23[51,64,65].

The goal of this study was to characterize the sex-determination locus on linkage group 1 in *O. niloticus*. We took a family-based strategy, separately pooling males and females from two crosses, and performing whole genome sequencing on the pooled DNAs. We cataloged the density and frequency of single nucleotide

polymorphisms (SNPs) and assessed their functional impact. We identified an 8.8Mb block of differentiation suggestive of a Y-linked inversion on linkage group 1. We found high densities of functionally significant SNPs within this differentiated block. Analysis of gonadal transcriptomes demonstrated an enrichment of female-biased gene models within the inversion, which suggests that dosage compensation is incomplete in this strain of *O. niloticus*.

Methods

Genomic DNA Pools

All animal procedures were conducted in accordance with University of Maryland IACUC Protocol #R-10-73. The fish sequenced are 3rd generation descendants of fish collected from a commercial tilapia farm in Amherst, Massachusetts USA. Individuals from two related lab-raised families were sacrificed and visually inspected for the presence testes or ovaries to determine the sex of each fish. Fish with ambiguous or immature gonads were excluded from the study. DNA was extracted from fin clips using a standard phenol/chloroform protocol. To confirm the family identity we genotyped each individual for two sex-linked microsatellite markers selected from the Broad anchored tilapia assembly on linkage group 1 (MS1045 at 14.32Mb and MS1141 at 15.53Mb). 33 males and 20 females from family BYL078 and 25 males and 13 females from family BYL084 were selected for pooling. DNA from each individual was then quantified by Picogreen fluorescence on a BioTek FLx800 spectrophotometer and appropriate dilutions were made to ensure equal representation of each individual in the pooled samples. The pooled male (or female) DNA from each family was sheared to 500bp using a Covaris

shearer and indexed separately during library construction. Paired-end libraries for each family/sex were constructed for Illumina sequencing using the Illumina TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA). The male (or female) libraries from each family were combined and each sex was sequenced in a separate lane on an Illumina HiSeq 2000. The male and female reads were deposited to NCBI with the accession numbers SRR1606298 and SRR1606304, respectively. Only reads passing the Illumina CASAVA filtering were retained.

Read qualities was checked with FASTQC[78]. Alignments to the *O. niloticus* anchored reference assembly[79] were performed with Bowtie 2[80] using the - - *very-sensitive* setting (Additional file 5). The mean alignment rate was 90.12% in males and 90.67% in females (Separate values for each family are given in Additional file 5). Read alignments were filtered for a minimum mapping quality (MAPQ) of 20 before further analysis. Insert sizes were analyzed using Picard tools CollectInsertSizeMetrics package[81]. The aligned mean insert size was 188.76bp (s.d. = 44.81bp) for males and 167.62bp (s.d. = 37.29bp) for females. Variants were called using GATK[82].

Genomic Analysis

Popoolation2[83] was used to calculate F_{ST} and Fisher's exact test on allele frequency differences between the male and female pools. Initial F_{ST} results from the individually adapter-indexed families were very similar, so all subsequent analyses were performed on the combined male or female pool, including unassigned reads from the male and female lane which could not be assigned to a particular family.

A custom Perl script, Sex_SNP_finder_now.pl (available at https://github.com/Gammerdinger/sex-SNP-finder), was used to identify SNPs at intermediate frequencies in the male pools and were fixed or nearly fixed in female pools at the same position. Intermediate SNPs were defined as SNPs with a frequency between 0.3 and 0.7 within the male pool. Fixed or nearly fixed sites required a frequency less than or equal to 0.1 or greater than or equal to 0.9 within the female pool. We used a non-overlapping window of 10kb to determine the density of these SNPs. The non-overlapping window did not include positions with coverage less than 10 reads in both sexes. The Sex_SNP_finder_now.pl script outputs a tab-delimited file with the number of SNPs per window along with an Integrative Genomics Viewer file[84] that lists all SNPs that were fixed in one designated pool and in intermediate frequency in the other.

We used SnpEff[85] to identify variants predicted to alter gene function. The SnpEff output was filtered to consider only the SNPs found using Sex_SNP_finder_now.pl. SnpSift[86] was used to extract out SNPs with similar effects and impacts.

Transcriptome Analysis

Gonads were dissected from individual larvae 28 days post-fertilization. The sex of each larvae was determined by genotyping microsatellite markers highly associated with sex. RNA from approximately 20 male or 20 female larvae was pooled and cDNA libraries were constructed using the Illumina TruSeq DNA Sample Prep Kit. Sequencing of these libraries yielded ~392 million reads for each male and female pool. Reads were aligned to the *O. niloticus* reference sequence with

TopHat2[87]. NCBI RefSeq annotations were used to guide the Cufflinks[88] assembly (-g) and Cuffdiff was used to was used to determine FPKM values for those gene models. The results were subsequently filtered to exclude gene models whose FPKM value was less than 0.05 in both males and females. Additionally, when comparisons between FPKM of the two sexes was carried out, if the FPKM value exceeded 0.05 in one sex and was zero in the other sex, it was considered an undefined bias favoring the sex with expression. Female-biased and male-biased gene models from inside and outside the proposed inversion were counted and statistical significance was looked for using χ^2 with Yates' correction on a 2x2 contingency table. These male and female reads from the RNA-Seq experiment were deposited to NCBI with the accession numbers SRR1606274 and SRR1606273, respectively.

Results and Discussion

Sequencing of male and female DNA pools

We obtained ~202 million reads from the pool of male DNA and ~219 million reads from the pool of female DNA. 90.12% of the male and 90.67% of the female reads were aligned to the *O. niloticus* reference genome. Genome-wide coverage was slightly lower in males (32.97, s.d. = 24.41 alignments per site), compared to females (36.68, s.d. = 31.39 alignments per site).

Large block of divergence on linkage group 1

The mean F_{ST} between the male- and female-pooled genomes at polymorphic sites over the entire genome was 0.0356 (s.d. = 0.030). A region between 10.1Mb and 18.9Mb on linkage group 1 showed a substantially higher value of F_{ST} = 0.0807

(s.d. = 0.061) (Figure 2a, 3a). This region corresponds to the previously mapped sex-determination region in this strain of *O. niloticus*[59,64–67]. Mean read coverage within the differentiated region was lower in males (34.65, s.d. = 10.56), compared to females (38.45, s.d. = 12.00), but this difference was consistent with the total number of reads obtained from each sex. We used Fisher's exact test to determine whether the allele frequency of SNPs was significantly different between males and females. We found a cluster of highly significant SNPs within the differentiated block on linkage group 1 (Figure 2b, 3b).

We also counted the number of positions per 10kb window that were fixed in female pools and had a SNP in intermediate frequency in male pools, as would be consistent with females having two X chromosomes and males having an X and a Y chromosome, using Sex_SNP_finder_now.pl. There were 40,514 of these SNPs found across the genome. 18,277 (2,076.932/Mb) lay inside the differentiated block and 22,237 (24.197/Mb) lay outside. Among the 300 non-overlapping 10kb windows with the highest frequency of these SNPs, 290 were found within the differentiated block on linkage group 1. The mean number of such SNPs per window was 21.81 (s.d. = 13.84) within the differentiated block and only 0.33 (s.d. = 1.29) outside of this region (Figure 2c, 3c). The elevated F_{ST}, along with the abundance of intermediate frequency SNPs in males that are fixed in females, suggests that this region has limited, if any, recombination between the X and Y alleles.

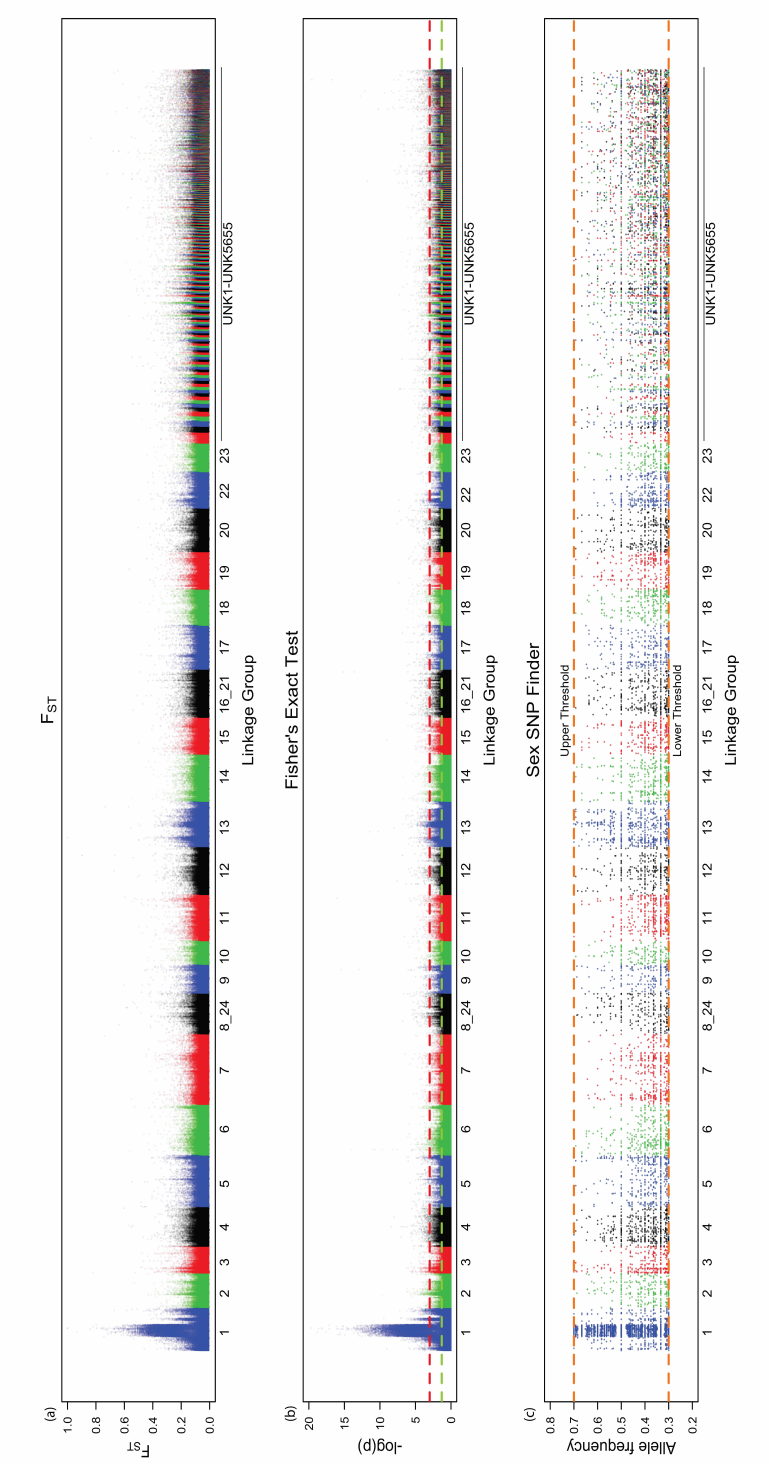


Figure 2. *Genome-wide scan for population differentiation.* *Genome-wide statistics for (a) F_{ST}, (b) Fisher's Exact Test and (c) intermediate frequency SNPs in males that are fixed or nearly fixed in females.*

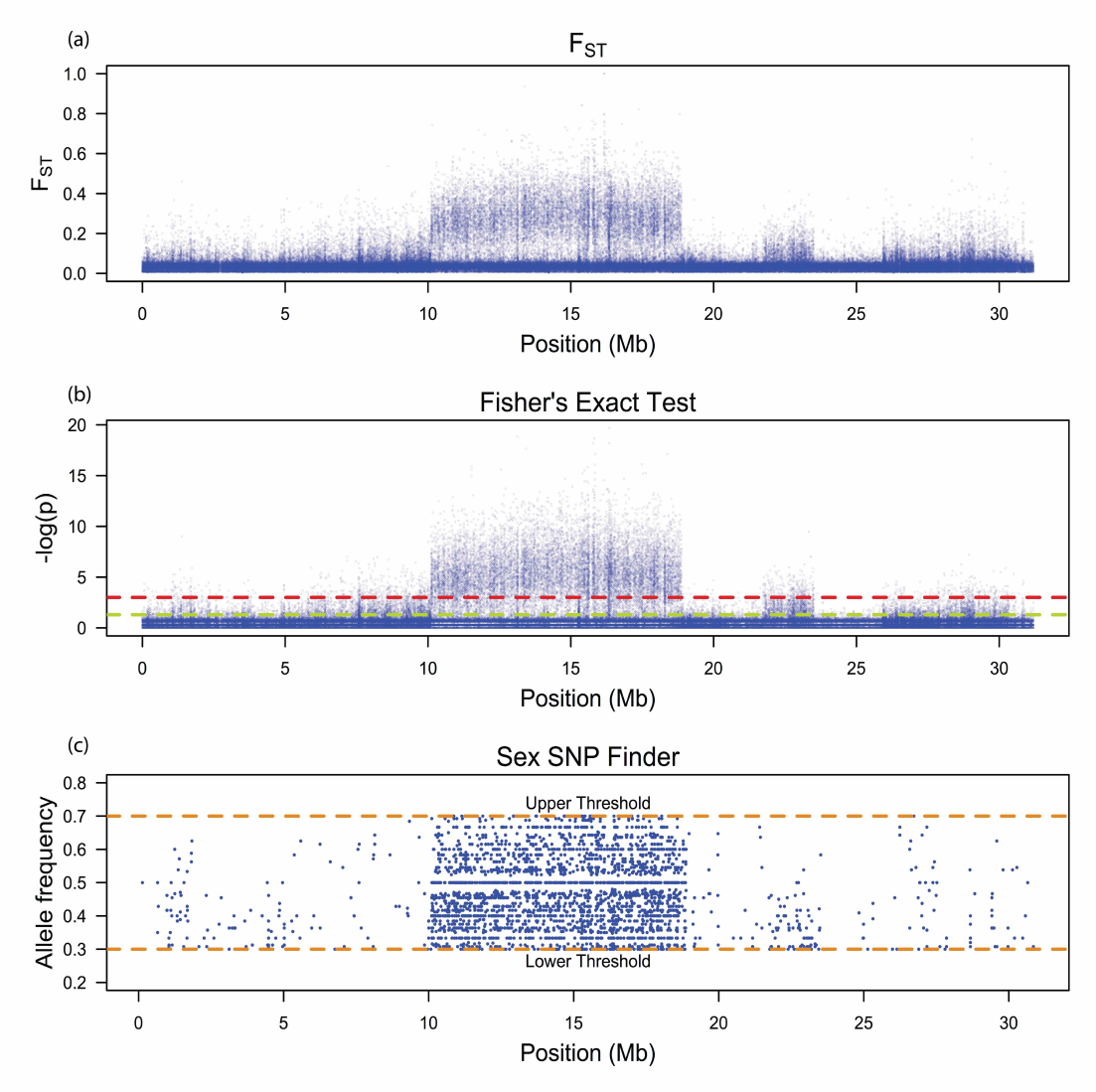


Figure 3. *Population differentiation on linkage group 1. Differentiation statistics for linkage group 1. (a) F_{ST}, (b) Fisher's Exact Test and (c) intermediate frequency SNPs in males that are fixed or nearly fixed in females.*

We considered the possibility that this block of differentiation is an artifact of the process by which we selected individuals for sequencing. We initially screened individuals by genotyping two sex-linked microsatellites in order to confirm family identity and sex. We required males to demonstrate a heterogametic pattern and females to demonstrate a homogametic pattern for both markers. Five male and five

female individuals were excluded by these criteria and may represent naturally sex-reversed individuals. The sharply defined edges of the block lie 4.22Mb upstream and 3.37Mb downstream of the microsatellites we genotyped (Figure 2 and 3), which would normally represent approximately 5cM of genetic distance in this species[89]. However, there is no evidence of an exponential decay of F_{ST} in the flanking regions as would be expected if there was recombination between the markers and the edges of the block. We considered the possibility that the high level of differentiation might be due to an 8.8Mb duplication on the Y. However, the depth of read coverage is relatively consistent across this entire linkage group. Additionally, cytogenetic studies have not revealed any evidence of heteromorphy in this chromosome pair as would be witnessed from a translocation[90]. The sum of the evidence suggests that this block of differentiation most likely reflects an 8.8Mb inversion on the Y-chromosome.

The relatively small size of the putative inversion, and its location in the middle of the chromosome, make it challenging to characterize using standard cytogenetic techniques. Ideally, we would characterize the breakpoints, but we were unable to identify anomalous Illumina mate pairs near the ends of the inversion in our short insert libraries. Longer reads or more widely spaced mate pairs will be needed to characterize the breakpoints of the proposed inversion.

Functionally significant SNPs

We examined the functional consequences of the SNPs that were fixed in female pools but at intermediate frequency in male pools at the same position using SnpEff and SnpSift[85,86]. Within the 8.8Mb differentiated block we found 13 stop codon changes (1.477/Mb), 3 start codon losses (0.341/Mb) and 2 splice site

alterations (0.227/Mb, Table 1). In the remaining 919Mb of the genome we found a total of 9 stop codon changes (0.010/Mb), no start codon losses, and 3 splice site alterations (0.003/Mb, Additional file 1). SNPs classified as non-synonymous coding changes by SnpEff totaled 168 (19.091/Mb) within the differentiated region and 147 (0.160/Mb) across the rest of the genome (Additional file 2).

Table 1. *Putative functional mutations in the proposed inversion*

Genes containing a stop codon, start codon or splice site alterations that were in intermediate frequency in males and fixed or nearly fixed in females within the proposed inversion on LG1.

Gene Name	SNP Location on LG1	Reference Codon	SNP Codon	Effect on the Y	Effect on the X	Frequency of SNP in Males	Frequency of SNP in Females
Ras-related protein R-Ras2 (LOC100693950)	10506882	CGA	**T**GA	Stop Gain	-	0.462	0
Signal peptide	10868192	TCA	T**G**A	-	Stop Gain	0.545	1
AMP deaminase 3 (LOC100694225)	11096201	TGA	T**C**A	Stop Lost	-	0.454	0
Zinc finger protein 821 (LOC100712266)	12466312	ATG	A**C**G	Start Lost	-	0.4	0
Zinc finger protein 821 (LOC100712266)	12466313	ATG	AT**A**	Start Lost	-	0.4	0
SAFB-like transcription modulator (LOC100711186)	12619332	CGA	**T**GA	-	Stop Gain	0.613	1
Hepatic lipase (Lipc)	12690753	TGA	**C**GA	Stop Lost	-	0.448	0
Ammonium transporter Rh type C 2 (LOC100706367)	13529856	ATG	AT**A**	-	Start Lost	0.575	1
AFG3-like protein 1 (LOC100702885)	13725056	TTA	T**G**A	Stop Gain	-	0.439	0

CUB and sushi domain-containing protein 1 (LOC100698036)	15189214	CAA	**T**AA	Stop Gain	-	0.391	0.027
CUB and sushi domain-containing protein 1 (LOC100698036)	15243948	TGC	TG**A**	Stop Gain	-	0.391	0
Neuromedin-K receptor (LOC100693904)	15788009	CAGG	CAG**C**	Splice Site Acceptor Lost	-	0.343	0
Protein FAM176A (LOC100700039)	16417062	CGA	**T**GA	Stop Gain	-	0.475	0
GC-rich sequence DNA-binding factor (LOC100700589)	16480688	CAGA	CAG **G**	Splice Site Acceptor Gain	-	0.512	0.018
BTB/POZ domain-containing protein KCTD3 (LOC100703295)	16809270	CGA	**T**GA	Stop Gain	-	0.52	0
Hypothetical protein (LOC100705710)	17489648	CAA	**T**AA	-	Stop Gain	0.452	0.97
Hypothetical protein (LOC100705710)	17507222	TAA	**C**AA	Stop Lost	-	0.469	0.04
Nuclear factor of activated T-cells	18194412	TGG	TG**A**	-	Stop Gain	0.5	0.982

The elevated density of high impact SNPs within the proposed inversion leads us to believe that deleterious alleles have begun to accumulate on this proto-Y. This is in accordance with the canonical model of heterogametic sex-chromosome evolution[14,15] and empirical observations of the eutherian mammal Y-chromosome, *Silene*, *Drosophila* and tongue sole[17,91–93].

Localization of the sex-determining gene

Previous studies have concluded that sex is multifactorial in *O. niloticus*[64,94] with a major sex-determination gene on LG1[59]. Our study confirms this previous

work, identifying an XY sex-determination locus in the middle of LG1 (Figure 4). The sex-determination gene was first mapped near microsatellite markers GM201 (13.96Mb) and UNH995 (18.02Mb, Figure 4a)[64]. Additional AFLP and FISH mapping found sex-associated markers at 13.79Mb, near 18Mb and at 19.43Mb (Figure 4b)[65,66]. Another study confirmed GM201 and UNH995 along with several other sex-associated markers spanning a region from 7.05Mb to 18.02Mb (Figure 4c)[59]. Lastly, a RAD-seq experiment found the highest associations at 14.95Mb (LOD score 18.5), but demonstrated a broad region spanning 10.92Mb to 16.44Mb with a LOD score above 15 (Figure 4d)[67].

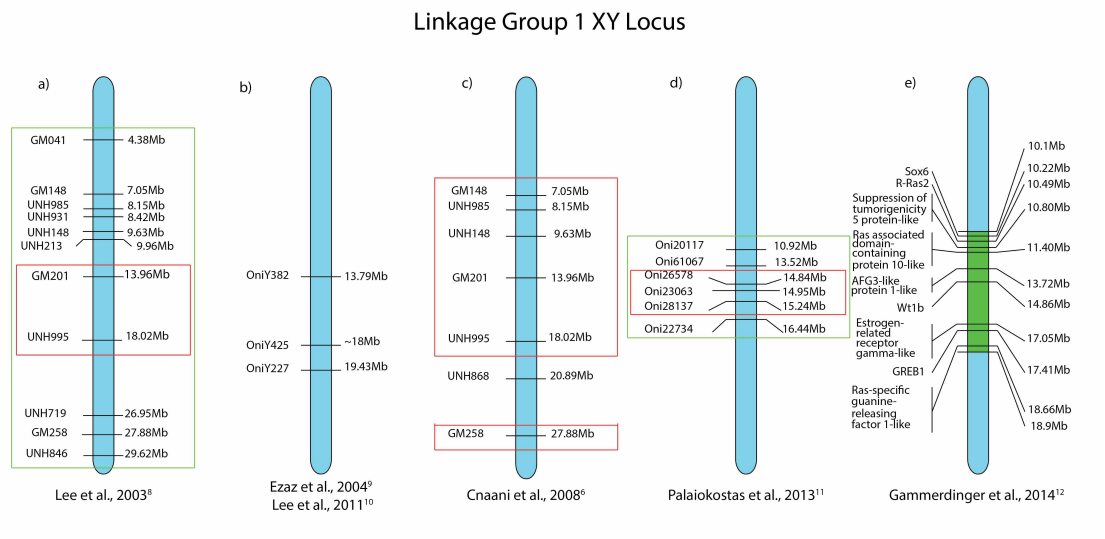


Figure 4. *Mapping of sex-determination locus on linkage group 1. Previous studies identifying sex-linked markers on LG1. (a) Lee et al., 2003 used a bulked segregant analysis. The green rectangle surrounds markers that were significantly sex-associated. The red rectangle encompasses the region with the highest significance. (b) Ezaz et al., 2004 identified three Y-specific AFLPs. OniY425 was assigned through BLAST to scaffold UNK43. It was placed on LG1 according to Lee et al., 2011, which used BAC contigs to place it within 100kb of UNH995. (c) Cnaani et al.,*

2008, also used a bulked segregant analysis. The markers within the red rectangles indicate markers that were significantly associated with sex. (d) Palaiokostas et al., 2013, identified sex-linked RAD-Seq markers. The green rectangle encompasses the markers with a LOD score greater than 15, while the red rectangle encloses the markers flanking the marker with the highest LOD score (Oni23063 with a LOD score of 18.5). (e) Proposed inversion in green with the eight candidate genes discussed in this paper.

The multifactorial nature of sex-determination in this species causes difficulties for genetic mapping studies. An XX individual may develop as a male due to other genetic factors, or environmental effects on differentiation. These individuals would appear to be recombinant in the sex interval. We previously claimed to exclude *Wilm's tumor protein homolog* (*Wt1b*) as the sex-determining gene on the basis of two recombinant individuals[95], but this conclusion is now in doubt. Conversely, the absence of recombination within the proposed inversion may preclude any further fine-mapping of the gene responsible for sex-determination.

Differences in gene expression

The block of differentiation on linkage group 1 comprises just more than 1% of the assembled genome and contains 257 RefSeq annotated genes. Cufflinks predicted 234 gene models within the block of differentiation and predicted 22,411 gene models across the entire transcriptome. Of the gene models that showed an FPKM of >0.05 in at least one sex, 7,977 gene models (37.4%) showed higher expression in males, while 13,375 (59.7%) gene models showed higher expression in females. Furthermore, within the inverted region, only 68 of these gene models

(29.6%) showed a male bias (Additional file 3), while 162 of these gene models (69.2%) showed a female bias (Additional file 4). The enrichment of female biased gene models within the proposed inversion is significant ($\chi^2 = 5.58$, $p < 0.05$). This data suggests that this sex chromosome is at an early-to-intermediate evolutionary phase where the degradation of a proto-Y, due to the putative inversion, has begun and a reduced expression of Y-linked genes in males is taking place. However, mechanisms for complete dosage compensation have yet to take hold.

Candidate sex determiners

Since the proposed inversion limits further attempts to fine-map the sex-determination gene, we evaluated candidate genes based upon putative functional polymorphisms, differential expression and prominence in pathways considered critical to sex-determination in other species. The complete list of candidate genes is presented in Table 2.

First, we analyzed all SNPs that SnpEff classified as high impact mutations (Table 1). One prominent candidate within the proposed inversion is *Ras-related protein R-Ras2* (10.49Mb-10.51Mb), which is part of the Ras-MEK-ERK pathway within the TGF-β signaling network[96]. Alterations to the TGF-β network have been suggested as the mechanism for sex-determination in several fish species[48]. *Ras2* has been implicated as particularly important in the proliferation of cells[97] and is expressed during early development in a hermaphroditic fish, *Kryptolebias marmoratus*[98]. *Ras-related protein R-ras2* has a stop codon gain in intermediate frequency in males that is absent in females. Disruption of *R-ras2* could lead to

decreased cell proliferation of primordial germ cells, resulting in increased likelihood of maleness[48,97].

Table 2. *Candidate genes in the proposed inversion. Sex-determination candidate genes within the proposed inversion with any codon changes and their FPKM values.*

Gene Name	Gene Location on LG1 (Mb)	SNP Location on LG1	Coding change	Pool frequency in males	Pool frequency in females	Male FPKM	Female FPKM
Transcription factor SOX-6 (LOC100694759)	10.22-10.30	10295869	T789K	0.3684	1	3.56951	4.24045
Ras-related protein R-Ras2 (LOC100693950)	10.48-10.51	10506882	R94STOP	0.4615	0	6.78248	5.16947
Suppression of tumorigenicity 5 protein (LOC100693420)	10.80-10.85	-	-	-	-	7.41571	2.09969
Ras association domain-containing protein 10 (LOC100693148)	11.40-11.41	-	-	-	-	0.252165	0.0688204
AFG3-like protein 1 (LOC100702885)	13.72-13.73	-	-	-	-	3.35696	0.36143
Wilms tumor protein homolog (LOC100701078)	14.86-14.88	14873730	A237V	0.4545	0	22.5644	13.2172
Estrogen-related receptor gamma (LOC100704106)	17.05-17.11	17093619	R172H	0.4333	0.06	0.370072	0.105597
Growth regulation by estrogen in breast cancer 1 (GREB1)	17.41-17.42	17424470	R1775C	0.5333	0.0571	0.961782	0.65826

Next, we evaluated SNPs that SnpEff categorized as missense mutations (Additional file 2). The first of these candidate genes is *Wilms tumor protein*

homolog, *Wt1b* (14.86Mb-14.88Mb), which has been implicated in gonadal development and acts directly upstream of *AMH*, the sex-determination gene in *Odontesthes hatcheri*[99]. *Wt1b* has also been demonstrated to bind to DNA and upregulate the sex-determination gene *Sry* in mammals. There is an A237V missense mutation in *Wt1b* that is absent in females and in intermediate frequency in males. Although our previous paper rejected *Wt1b* on the basis of two recombinant individuals[95], in light of the proposed inversion, we now believe that these individuals represented instances of natural sex reversal, not recombination.

A third candidate gene is *estrogen-related receptor gamma*, *ERR*γ (17.05-17.11Mb). It has a R172H missense mutation within a predicted DNA-binding domain[100]. *ERR*γ has been shown to be a transcriptional activator of *DAX-1*, and *DAX-1* has been implicated as having an antagonistic effect to *Sry* in mammals[101]. Therefore, a mutation in the DNA-binding domain of *ERR*γ could reduce *DAX-1* transcription and thus have a masculinizing effect.

Growth regulation by estrogen in breast cancer 1 (*GREB1*) is another candidate gene (17.41-17.42Mb) with a missense mutation. The R1775C mutation alters the side chain from a basic side chain to a polar side chain. *GREB1* has been shown to be predominantly expressed within ovaries of young mice[102]. Additionally, *GREB1* has been demonstrated to be a coactivator of estrogen receptor-α[103]. Therefore, the missense mutation in *GREB1* could downregulate the expression of estrogen receptor-α, resulting in a masculinizing effect on the developing embryo.

Another potential sex-determination gene is *transcription factor SOX-6* (10.22Mb-10.30Mb). There is a T789K missense mutation in intermediate frequency

in males that is fixed in females and changes a polar side chain into a basic one. *SOX-6* protein is localized to the same nuclear speckles as *Sry* and it has been suggested that it might play a role in sex-specific splicing in mammals[104].

We also evaluated gene models showing differential expression between males and females (Additional file 3 and 4). One candidate *AFG3(ATPase Family Gene 3)-like protein 1* (13.72Mb-13.73Mb) has over a nine-fold male-biased expression. It is also on the list of SNPs with high impact coding alterations with a stop codon gain. However, a clear tie to sex-determination has yet to be elucidated.

Suppression of tumorigenicity 5 protein (10.80Mb-10.85Mb) and *Ras association domain-containing protein 10* (11.40-11.41Mb) were also identified for having over a three-fold male-biased expression pattern. Ras association domain family proteins have been implicated as tumor suppressors[105–107]. Therefore, upregulation of these genes could suppress primordial germ cell proliferation leading to maleness.

Lastly, it is possible that there could be Y-specific genes that were not captured in our study, because the reference genome that the reads were align to is a homozygous clonal XX individual. A complete list of candidate genes within the proposed inversion is in Additional file 6.

Conclusions

Inversions have been well-documented in sex-chromosome evolution and are one possible mechanism for resolving sexually antagonistic selection near the novel sex-determiner through a reduction in recombination[17]. This study revealed an 8.8Mb block of differentiation between males and females. The variety of evidence

presented here is most consistent with the presence of an inversion. The decay of genes and overall level of differentiation indicate that this region has substantially reduced recombination. We have also documented an accumulation of SNPs causing functional alterations within this region, as would be expected for a genomic region suffering both the deleterious effects of Muller's Ratchet and accumulation of deleterious alleles hitchhiking to fixation with advantageous alleles. The transcriptome data indicates that genes inside the proposed inversion show significant enrichment for female-biased expression. These data suggest that *O. niloticus* has not yet evolved complete dosage compensation. Future functional studies are needed to identify the master sex-determination gene(s) within this region. Further research on cichlid sex determination will help unravel the underlying sex-determination network that underlies the rapid turnover of sex-determination mechanisms within teleosts.

Competing Interests

The authors declare that they have no competing interests.

Author's contributions

WJG helped conceive the study and was involved in specimen collection, genotyping of specimens, quantifying DNA, computational analysis and drafting of the manuscript. MAC helped conceive the study, substantially contributed to the computational analysis and aided in drafting of the manuscript. EAA substantially aided in specimen collection and genotyping of specimens. RBR constructed the cDNA libraries and aided in drafting of the manuscript. TDK helped conceive the study, constructed the pooled DNA libraries, guided the analysis and helped draft the manuscript.

Acknowledgements

We wish to thank the National Science Foundation for providing computational space and support through Indiana University Pervasive Technology Institute on the National Center for Genome Analysis Support (NCGAS) Mason Computing Cluster. We also wish to thank Karen Carleton for her insightful comments and suggestions on the manuscript. This work was supported in part by the National Science Foundation under Grant Number DEB-1143920.

Additional Files

Additional files reference for this chapter can be found at:

https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-975

Chapter 3: Comparative Analysis of a Sex Chromosome from the Blackchin Tilapia, *Sarotherodon melanotheron*

Previously published as: Gammerdinger, W. J., Conte, M. A., Baroiller, J. F., D'Cotta, H., Kocher, T.D. "Comparative analysis of a sex chromosome from the blackchin tilapia, *Sarotherodon melanotheron*." *BMC Genomics* 17.1 (2016): 808.

Abstract

Inversions and other structural polymorphisms often reduce the rate of recombination between sex chromosomes, making it impossible to fine map sex-determination loci using traditional genetic mapping techniques. Here we compare distantly related species of tilapia that each segregate an XY system of sex-determination on linkage group 1. We use whole genome sequencing to identify shared sex-patterned polymorphisms, which are candidates for the ancestral sex-determination mutation. We found that *Sarotherodon melanotheron* segregates an XY system on LG1 in the same region identified in *Oreochromis niloticus*. Both species have higher densities of sex-patterned SNPs, as well as elevated number of ancestral copy number variants in this region when compared to the rest of the genome, but the pattern of differentiation along LG1 differs between species. The number of sex-patterned SNPs shared by the two species is small, but larger than expected by chance, suggesting that a novel Y-chromosome arose just before the divergence of the two species. We identified a shared sex-patterned SNP that alters a *Gata4* binding site near *Wilms tumor protein* that might be responsible for sex-determination. Shared sex-patterned SNPs, insertions and deletions suggest an ancestral sex-determination

system that is common to both *S. melanotheron* and *O. niloticus*. Functional analyses are needed to evaluate shared SNPs near candidate genes that might play a role in sex-determination of these species. Interspecific variation in the sex chromosomes of tilapia species provides an excellent model system for understanding the evolution of vertebrate sex chromosomes.

Background

Natural selection alters the local genomic environment around newly evolved sex-determination loci in two ways. First, selection favors the accumulation of sexually antagonistic alleles and their association with the appropriate X- or Y-haplotype. Second, selection also favors a reduction in the rate of recombination between the proto-X and proto-Y to preserve the associations between the sexually antagonistic alleles and the sex-determination locus[18,108]. The process is cyclic, in that selection will favor the recruitment of additional sexually antagonistic loci within a growing region of reduced recombination[71]. Once a region of reduced recombination is established, deleterious mutations will begin to accumulate via Muller's Ratchet[19]. Through these mechanisms, the region of differentiation between the X- and Y-chromosome grows, creating a series of evolutionary strata from the oldest and most decayed regions to the newest and least decayed[17]. Structural rearrangements, such as inversions, are one mechanism that reduces recombination and contributes to the creation of evolutionary strata on sex chromosomes[17,18,108].

The initial sex-determination loci are buried within the oldest evolutionary strata because the first selectively favored chromosomal rearrangement encompasses both the sex-determination locus and at least one sexually antagonistic locus.

Traditional genetic mapping techniques cannot be used to fine map the sex-determination locus because of the lack of recombination inside these structural polymorphisms. An alternative approach is to identify shared ancestral polymorphisms inside the deepest and most decayed stratum among species that share a common ancestral sex-determination mechanism.

Cichlid fishes are a model system for studying evolutionary processes[52]. Sex-determination has evolved rapidly among African cichlids. Sex-determination loci have been identified on linkage groups (LG) 1 (XY), 3 (WZ) and 23 (XY) among species of tilapia, and on linkage groups 5 (WZ), 7 (XY) and possibly others in haplochromine cichlids[54,56,59]. The sex determination locus on LG23 appears to be a duplication of *amh*, but the sex-determination genes on the other chromosomes have not yet been identified [45,46].

Previous studies have mapped the sex-determination locus on linkage group 1 in the Nile tilapia (*Oreochromis niloticus*) to a region of a few centimorgans.[59,64,66,67]. We recently identified a high density of sex-patterned SNPs on LG1 in the region from 10.1 to 18.9Mb[109]. This 8.8Mb region is significantly enriched for missense mutations and likely corresponds to an inversion that restricts recombination between the X- and Y-chromosomes.

The blackchin tilapia, *Sarotherodon melanotheron*, is a sister group to the genus *Oreochromis*[110] and has received relatively little attention with respect to sex-determination. One study reported that the chromosomes of *S. melanotheron* were homomorphic with a chromosome count consistent with that of other species of *Oreochromis*[111]. In this study we identify the sex-determination region in *S.*

melanotheron, characterize the shared regions of differentiation on LG1 between *O. niloticus* and *S. melanotheron* and catalog shared SNPs that may be responsible for sex-determination in these species.

Methods

Materials

Sarotherodon melanotheron were originally collected from Lake Guiers, Senegal. The individuals sampled are the 4th generation progeny of the wild-caught fish. Fins were sampled from a total of 22 male and 22 female fish and preserved in a salt-DMSO preservation solution. The sex ratio of the family was 50:50.

Sequencing

DNA was extracted from each fin-clip by phenol-chloroform extraction using phase-lock gel tubes (5Prime, Gaithersburg, Maryland). DNA concentrations were measured by fluorescence spectrometry and normalized during library preparation. DNA from 21 males and 22 females were then separately pooled (one likely male was excluded due to ambiguous gonads). Sequencing libraries were constructed using the TruSeq DNA PCR-Free LT Kit (Illumina, San Diego, California). A full lane of Illumina 183bp paired-end sequencing was performed for each sex.

For comparative analyses, we reanalyzed whole genome sequencing data for male and female *Oreochromis niloticus* from our previous study[109]. These data consist of 100bp paired-end reads from separate male and female DNA pools.

Read mapping

Reads that did not pass Illumina CASAVA 1.8 filtering were removed. Read qualities for both lanes were analyzed using FastQC[78]. The reads were then mapped to the *O. niloticus* reference assembly with Bowtie2 using the *--very-sensitive* setting and a minimum insert length of 200bp[79,80]. Variants were called using GATK in the *UnifiedGenotyper* mode[82]. Read qualities with a PHRED score of at least 20 were required for variant calling.

Identification of sex-patterned SNPs

A custom script was used to identify sites showing a sex-patterned signature. Sex_SNP_Finder_GA.pl is an expansion of the Sex_SNP_finder_now.pl script released previously[109,112]. Sex-patterned SNPs are sites that are fixed or nearly fixed in the homogametic sex and in a frequency between 0.3 and 0.7 in the heterogametic sex. We required a minimum read depth of ten and a minimum allele count of two. We assessed the overall density of sex-patterned SNPs in 10kb non-overlapping windows. We also counted regions enriched for sex-patterned SNPs by identifying non-overlapping 10kb windows containing at least 10 sex-patterned SNPs.

In addition to finding sex-patterned SNPs, the updated script calculates F_{ST}, d_{xy}, d_a, Nei's D and C_P for each nucleotide position across the genome. F_{ST} is calculated in accordance with the method used in PoPoolation2, except that we calculated values at each site instead of by windows[83]. Due to numeric difficulties handling the upper bound of Nei's D, the calculation is modified for alternatively fixed positions to make the calculation based upon the maximum coverage level specified by the user. C_P is a metric of population differentiation that ranges from 0 (no population differentiation)

to 1 (complete population differentiation), where x_i and y_i are the frequencies of each allele in populations x and y[113].

$$C_p = \sum_{i=1}^{n} \frac{1}{2} | x_i - y_i |$$

Functional annotation

The functional significance of sex-patterned SNPs was evaluated with SnpEff and SnpSift using the gene annotations obtained from NCBI (RefSeq release 70)[85,86]. We excluded mRNA models that did not have full-length protein coding sequences. Non-synonymous substitutions were subsequently evaluated with PROVEAN to predict functional impacts on protein structure[114]. Three missense mutations, R203C in XM_003448054.2 (LG6), E1235V in XM_003438386.2 (LG13) and G215A in XM_005466093.1 (LG22), were excluded from the PROVEAN analysis due to low quality gene annotations as defined by PROVEAN. Missense mutations harboring PROVEAN scores less than the recommended threshold of -2.5 were considered deleterious.

Identification of shared sex-specific SNPs

Sex-patterned SNPs in *S. melanotheron* were compared to the sex-patterned SNPs identified in *O. niloticus*, which also segregates a sex-determination system on LG1. The null hypothesis for the expected number of shared sex-patterned SNPs was calculated by multiplying the frequency of sex-patterned SNPs within the previously identified region from 10.1Mb to 18.9Mb on LG1 in *O. niloticus,* by the frequency of sex-patterned SNPs within the same region in *S. melanotheron* by the size of the region.

Transcription factor binding site analysis

SNPs that were sex-patterned and shared between *O. niloticus* and *S. melanotheron* were inspected to ensure that they shared common X- and Y-alleles. Flanking regions for the SNPs passing this criteria were extracted using Bedtools and compared to the JASPAR CORE Vertebrata 2016 database to identify possible transcription factor binding sites[115,116]. A JASPAR relative score threshold of 0.80 was used to assess the significance of putative binding sites.

Identification of copy number variants

Copy number variation within the *S. melanotheron* and *O. niloticus* datasets was assessed with VarScan 2, using a minimum window size of 100bp, a maximum window size of 1kb and *amp* and *del* thresholds equal to 0.2[117]. A custom script (Varscan_multiple.pl) was developed to find conserved copy number variants between *n*-number of VarScan 2 comparisions[118]. We then utilized a non-overlapping window approach, which excluded assembly gaps from the windows, to quantify the density of these conserved copy number variants within 10kb windows across the genome.

Statistical Assessment

Mann-Whitney U-tests were used to determine if a region was enriched for sex-patterned SNPs or conserved copy number variants. All testing used 10kb non-overlapping windows as samples. A Bonferroni correction was applied to control for 12 comparisons resulting in a significance threshold of α=0.004167.

Results

Sequencing and mapping of reads

We obtained a total of ~160 million and ~162 million 183bp paired-end reads from the female and male pools of *S. melanotheron*, respectively. The alignment rates to the *O. niloticus* reference sequence were 85.50% and 85.69%, respectively. The mean coverage in the female pool was 35.12 reads, while the mean coverage in the male pool was 35.75 reads.

We obtained a total of ~219 million and ~202 million 100bp paired-end reads from the female and male pools of *O. niloticus*, respectively. The alignment rates to the *O. niloticus* reference were 90.49% and 89.96%, respectively. The mean coverage in the female pool was 27.17 reads, while coverage in the male pool was 26.84 reads.

F_{ST} Differentiation in *S. melanotheron*

Examination of the whole genome F_{ST} plot comparing male and female *S. melanotheron* identifies a strong signal on LG1 (Figure 5). A closer examination of LG1 reveals that this region overlaps with the previously identified XY sex-determination region in *O. niloticus* (Figure 6). The region of divergence on LG1 is broader in *S. melanotheron* than in *O. niloticus* and spans from approximately 10.1Mb to 28Mb. The boundaries of the differentiated region in *S. melanotheron* are gradual, unlike the sharp boundaries found in *O. niloticus* on LG1 from 10.1Mb to 18.9Mb and from 21.7Mb to 23.6Mb.

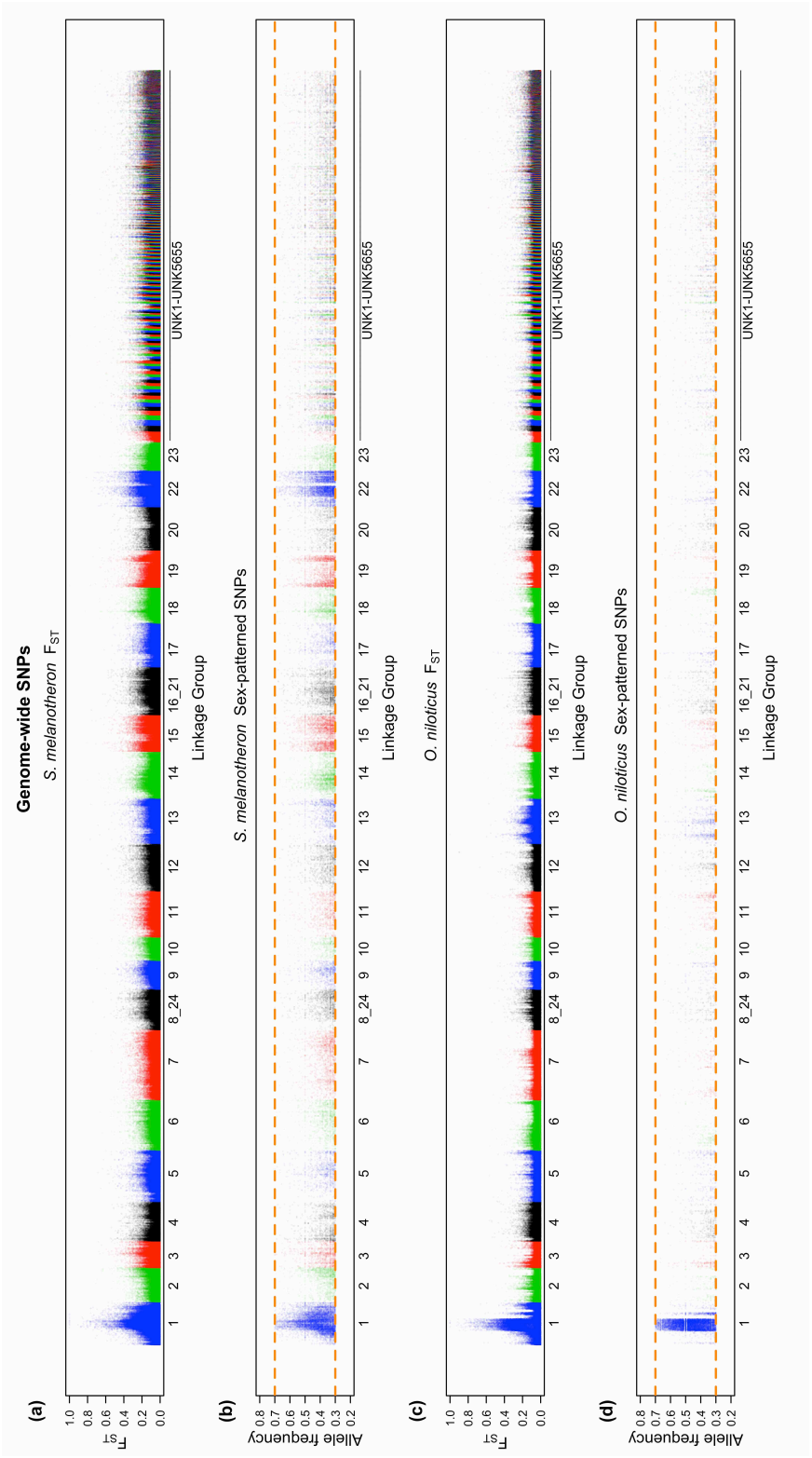


Figure 5. *Whole genome survey of (a) F_{ST} in S. melanotheron, (b) sex-patterned SNPs in S. melanotheron, (c) F_{ST} in O. niloticus and (d) sex-patterned SNPs in O. niloticus.*

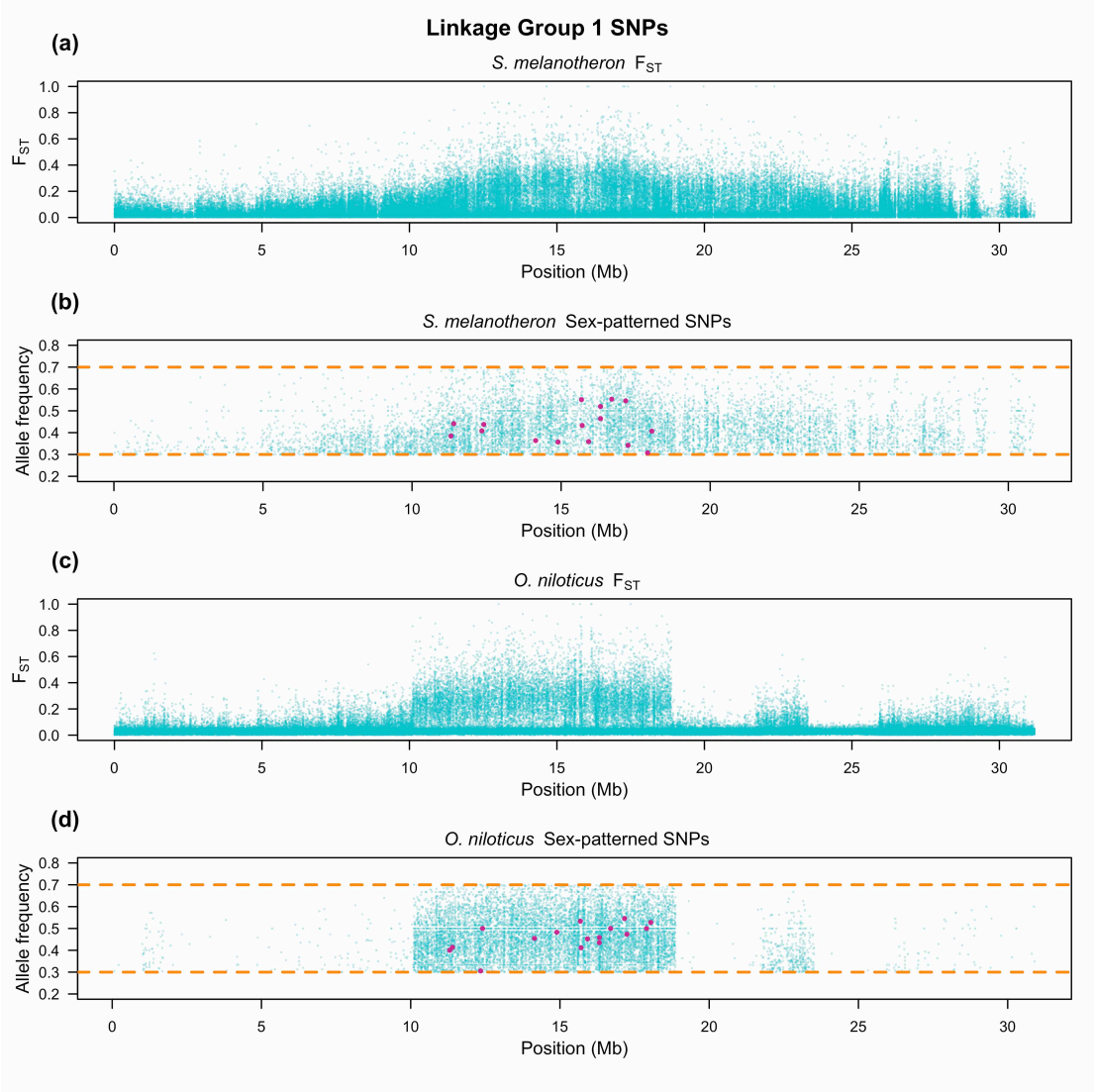


Figure 6. *Linkage group 1 survey of (a) F_{ST} in S. melanotheron, (b) sex-patterned SNPs in S. melanotheron, (c) F_{ST} in O. niloticus and (d) sex-patterned SNPs in O. niloticus. The red points represent the shared sex-patterned SNPs between O. niloticus and S. melanotheron.*

S. melanotheron also shows indications of divergence between the sexes on LG22. There is no previous evidence to suggest that this region is associated with sex in any cichlid species and the signal is not as strong as on LG1 (Figure 7 and Table 3). Importantly, *S. melanotheron* shows no strong sex-patterned signal from LG3, a

region known to harbor a sex-determination locus in *Pelmatotilapia mariae* and some members of *Oreochromis*[59].

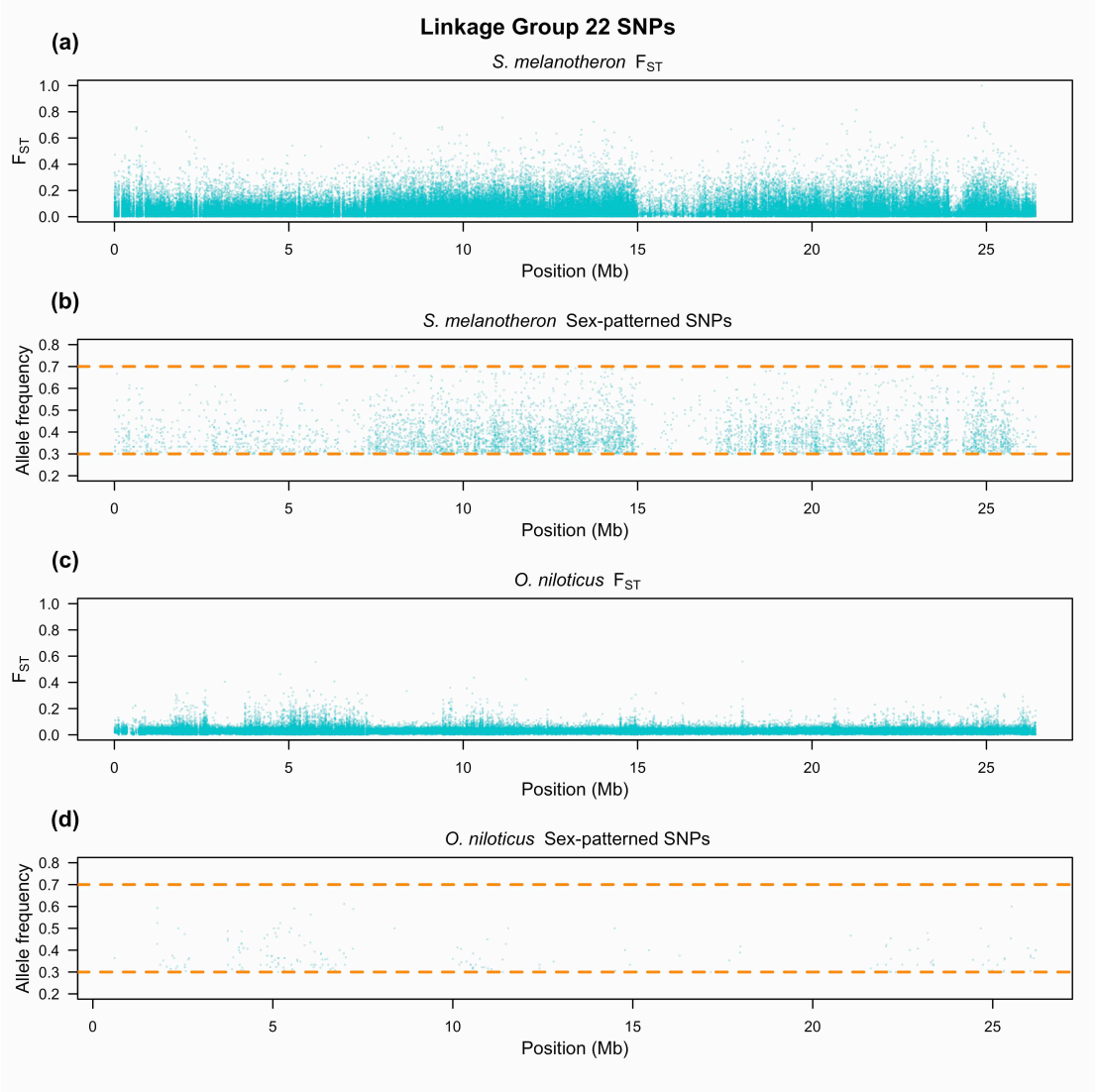


Figure 7. *Linkage group 22 survey of (a) F_{ST} in S. melanotheron, (b) sex-patterned SNPs in S. melanotheron, (c) F_{ST} in O. niloticus and (d) sex-patterned SNPs in O. niloticus.*

Table 3. *Average F_{ST} across all polymorphic sites along with total counts and densities of sex-patterned SNPs across the regions of differentiation in S. melanotheron and O. niloticus.*

	Average F_{ST} across polymorphic sites	*Sex-patterned SNPs (per Mb)*	*Enriched 10kb windows (per Mb)*	*Missense SNPs (per Mb)*	*PROVEAN SNPs <-2.5 (per Mb)*	*Synonymous SNPs (per Mb)*
S. melanotheron						
LG1: 10.1Mb-18.9Mb	0.111	5,342 (607.05)	193 (21.93)	61 (6.93)	6 (0.68)	121 (13.75)
LG1: 18.9Mb-28Mb	0.081	2,702 (296.92)	52 (5.71)	38 (4.18)	5 (0.55)	40 (4.40)
LG22	0.057	5,709 (216.16)	80 (3.03)	42 (1.59)	9 (0.34)	108 (4.09)
Rest of the Genome	0.034	36,335 (41.13)	123 (0.14)	669 (0.76)	103 (0.12)	962 (1.09)
Total		***50,088***	***448***	***810***	***123***	***1,231***
O. niloticus						
LG1: 10.1Mb-18.9Mb	0.039	10,792 (1,226.36)	517 (58.75)	159 (18.07)	25 (2.84)	256 (29.09)
LG1: 21.7Mb-23.6Mb	0.024	399 (210.00)	3 (0.33)	5 (2.63)	1 (0.53)	6 (3.16)
Rest of the Genome	0.021	8,602 (9.38)	19 (0.02)	167 (0.18)	27 (0.03)	180 (0.20)
Total		***19,793***	***539***	***331***	***53***	***442***

Sex-patterned SNPs

We found a total of 50,088 SNPs in *S. melanotheron* and 19,793 SNPs in *O. niloticus* fitting the sex-patterned criteria. There were 448 and 539 non-overlapping 10kb windows with at least 10 sex-patterned SNPs in *S. melanotheron* and *O. niloticus*, respectively (Table 3). The highest densities of sex-patterned SNPs occurred between 10.1Mb and 18.9Mb in both *O. niloticus* and *S. melanotheron*. Sex-

patterned SNPs found on LG1 between 10.1Mb and 18.9Mb are at significantly higher densities in this region than any other region in either species ($p < 0.0001$, for all five comparisons to other regions noted in Table 3).

In *S. melanotheron,* the region on LG1 from 18.9Mb to 28Mb, as well as LG22, both showed a significantly higher density of sex-patterned SNPs when compared to the rest of the genome ($p < 0.0001$, for both comparisons). The region on LG1 from 18.9Mb to 28Mb also demonstrated a significantly higher density of sex-patterned SNPs than LG22 ($p < 0.0001$). In *O. niloticus*, the region on LG1 from 21.7Mb to 23.6Mb showed a significantly higher density of sex-patterned SNPs than the rest of the genome ($p < 0.0001$).

Functional impacts

We evaluated the functional impacts of the sex-patterned SNPs and found a total of 810 missense and 1231 synonymous mutations in *S. melanotheron*. Table 3 shows the average F_{ST} across all polymorphic SNPs within each region as well as the genomic distribution of each subset of the sex-patterned SNPs: total sex-patterned SNPs, 10kb windows enriched for sex-patterned SNPs, missense SNPs, missense SNPs with a PROVEAN score less than -2.5 and synonymous mutations. The density of each class of sex-patterned SNPs in *S. melanotheron* was consistently from highest to lowest: LG1 between 10.1Mb and 18.9Mb, LG1 between 18.9Mb and 28Mb, LG22, then across the rest of the genome. There were seven stop codon gains or losses, but only one mapped to a region of divergence on LG1 (Additional file 1).

In *O. niloticus* we identified a total of 331 missense and 442 synonymous mutations. The density of each class of sex-patterned SNPs in *O. niloticus* was

consistently from highest to lowest: LG1 between 10.1Mb and 18.9Mb, LG1 between 21.7Mb and 23.6Mb, then across the rest of the genome. There were six stop codon gains or losses, but none mapped to the regions of divergence (Additional file 2). These counts differ slightly from Gammerdinger et al., 2014 due to a difference in the gene annotation versions used. There were no conserved stop codon gains or losses between *S. melanotheron* and *O. niloticus*.

Copy Number Variants

We identified a significantly higher density of conserved duplications and deletions between *O. niloticus* and *S. melanotheron* on both LG1 from 10.1Mb to 18.9Mb ($p < 0.0001$) and LG1 from 21.7Mb to 23.6Mb ($p < 0.0002$) when each is compared to the rest of the genome. However, we did not detect any significant difference in the density of duplications and deletions conserved between the species in these divergent regions on LG1.

Shared SNPs

There are 42 sex-patterned SNPs conserved between *S. melanotheron* and *O. niloticus*, compared to the null expectation of 6.55 conserved SNPs. We examined all 42 positions to confirm that the X- and Y-alleles were consistent in the two species. There were 13 SNPs where the X- and Y-alleles are switched between species (*e.g.* when the X-allele in *O. niloticus* is the Y-allele in *S. melanotheron*). Eight sites showed a pattern where the Y-alleles are different but the X-alleles were the same in each lineage. This narrowed the list to 21 SNPs where the X- and Y-alleles were shared between the two species. We filtered these 21 SNPs by genomic position. Sixteen fell within the region on LG1 between 10.1 and 18.9Mb, one lay on LG14

and four fell on unanchored scaffolds (Additional file 3 and 4). Of these shared SNPs, none were missense mutations and only one was a synonymous mutation. The rest were located in non-coding regions. We identified shared sex-patterned SNPs within 20kb of two previously identified candidate genes for sex-determination on LG1[109]. One was 19kb downstream of *Wilms' tumor protein* and the other was 5kb downstream of *Ras association domain containing protein 10*.

We used JASPAR to examine a short region flanking each of the 21 shared sex-patterned SNPs for transcription factor binding sites (Additional file 5). A sex-patterned SNP at position 14,895,959 on LG1, near *Wilms' tumor protein*, resulted in a loss of a *Gata4* binding site. The sex-patterned SNP at 11,400,015 on LG1, near *Ras association domain containing protein 10* did not alter binding sites for any transcription factor binding sites known to be involved in the sex-determination network (Additional file 5).

Discussion

Evidence for a shared LG1 sex-determination mechanism between 10.1Mb and 18.9Mb

The initial line of evidence that *S. melanotheron* shares an ancestral sex-determination mechanism with *O. niloticus* is the overlap of the region harboring sex-determination on LG1 (Figure 5 and 6). Convergence for sex-determination within the same 8.8Mb region within a 927Mb genome, while possible, seems unlikely. This region also contains the highest density of sex-patterned SNPs across the genome in both species. The high density of sex-patterned SNPs explains the elevated sex-related F_{ST} witnessed within this region. Furthermore, the copy number variant

analysis demonstrates that the region on LG1 from 10.1Mb to 18.9Mb is not significantly more divergent than the region from 21.7Mb to 23.6Mb, but it does have significantly more conserved insertions and deletions compared to the rest of the genome. Lastly, the number of shared sex-patterned SNPs is several times more than expected by chance (6.55 shared sex-patterned SNPs). Collectively, the evidence strongly suggests the idea that *O. niloticus* and *S. melanotheron* share a common ancestral sex-determination mechanism on LG1.

Evidence for a shared sexually antagonistic locus on LG1 between 21.7Mb and 23.6Mb

The region from 21.7Mb to 23.6Mb on LG1 has a moderately strong signal of differentiation between males and females in both lineages. The region from 18.9Mb to 28Mb in *S. melanotheron* includes this region and is the second most densely sex-patterned SNP region in the genome. The region from 21.7Mb to 23.6Mb does not have significantly more conserved insertions and deletions than the region from 10.1Mb to 18.9Mb, but it does have significantly more conserved insertions and deletions when compared to the rest of the genome. This suggests that the block of differentiation between 21.7Mb and 23.6Mb has a shared deep history with the sex-determination region.

Within this region, we found no shared SNPs with a consistent allelic pattern between *O. niloticus* and *S. melanotheron*, but there was a sex-patterned locus in both species where the X- and Y-alleles were switched. This suggests that there could have been an ancestral polymorphism that was alternatively sorted during speciation and the emergence of the sex chromosome. The evidence suggests that the diverged

region on LG1 from 21.7Mb to 23.6Mb was present during the initial divergence of sex-determination ancestral to both *O. niloticus* and *S. melanotheron* on LG1 from 10.1Mb to 18.9Mb.

This linked, differentiated region from 21.7Mb to 23.6Mb may harbor a sexually antagonistic locus. However, the sequence between these regions of divergence, from 18.9Mb to 21.7Mb, shows differentiation in *S. melanotheron*, but not in *O. niloticus*. Previous theoretical work has suggested that sexually antagonistic loci can be in linkage disequilibrium with the sex-determination locus over large distances while not showing strong signs of differentiation between the two loci[119]. It appears that *O. niloticus* has potentially maintained linkage disequilibrium between these two pockets of differentiation, while *S. melanotheron* has accumulated mutations across this entire region. This theoretical prediction also postulates that the sexually antagonistic region should have a lower level of divergence than the sex-determination region. This proposition is consistent with the data collected in *O. niloticus*.

From our data, it is unclear if the intervening region from 18.9Mb to 21.7Mb diverged initially and then recombination was restored in the *O. niloticus* lineage or if these were initially two separate regions of divergence held together by sexually antagonistic selection and recombination was subsequently reduced across the entire region in the *S. melanotheron* lineage. Figure 8 illustrates the latter sequence of events, but the former sequence of events is also plausible. Sequences of more species harboring the LG1 sex-determination system might distinguish these two hypotheses.

Alternatively, this region could represent a misassembly in the *O. niloticus* reference genome or a Y-specific structural rearrangement.

Evidence for speciation soon after sex chromosome emergence

Of the 42 loci that demonstrate a sex-patterned profile in both data sets, 21 loci had the X- and Y-alleles conserved, while 13 loci had X- and Y-alleles switched. The 21 conserved X- and Y-alleles indicate that this sex chromosome had a shared ancestry between *O. niloticus* and *S. melanotheron*. The 13 loci where the X- and Y-alleles were switched, indicate that this period of shared ancestry was limited, and that some sites that underwent alternative lineage sorting as the proto-X and proto-Y were diverging. This indicates that speciation occurred soon after sex chromosome divergence.

Evidence for structural rearrangements and their timing

The pronounced boundaries of divergence, high densities of sex-patterned SNPs and predicted deleterious mutations above the background level in *O. niloticus*, when taken together with the knowledge that chromosomes of this species are homomorphic in cytogenetic analyses, provide strong evidence for an inversion between the X- and Y-chromosomes in *O. niloticus*[90]. The absence of similar sharp boundaries in *S. melanotheron*, along with lower densities of sex-patterned SNPs and predicted deleterious mutations, suggest that this inversion is likely not present in the *S. melanotheron* lineage. We suggest the most parsimonious scenario is that this inversion arose following the divergence of *O. niloticus* and *S. melanotheron*.

The differentiation between males and females on LG22 of *S. melanotheron* suggests an association between this linkage group and sex-determination. While the

levels of differentiation are lower than those witnessed on LG1 between 10.1Mb and 28Mb, LG22 has a high density of sex-patterned deleterious mutations and the overall level of differentiation is significantly higher than the genomic background. Currently, this patterned is consistent with four hypotheses: 1) The signal could be the result of a reciprocal translocation of similar sized fragments between LG1 and LG22 after the divergence of *O. niloticus* and *S. melanotheron*. A previous karyotyping study has concluded that the chromosomes are homomorphic and that *S. melanotheron* have the same chromosome count as *O. niloticus*, but this study might have failed to detect such a translocation[111]. 2) LG22 is in linkage disequilibrium with the LG1 XY system. This seems unlikely because of the strong selection needed to overcome linkage disequilibrium between two independently assorting chromosomes. 3) A region on LG22 epistatically contributes to sex-determination within LG1 in this family. Multiple sex-determination systems have been observed in some cichlid species, however no sex-determination loci been previously reported from LG22 in cichlids[53,54,59]. 4) The signal could be an artifact of the sampling structure. This also seems unlikely as it would suppose that two copies of LG22 were segregating in the family and that males disproportionately received more of one copy and females received almost none of that copy by chance meiotic events. We suggest that a reciprocal translocation from LG22 to LG1 following the divergence of *S. melanotheron* and *O. niloticus* is the most likely scenario for explaining the observed patterned of sex-patterned differentiation on LG22. Future cytogenetic studies may provide data to test this hypothesis.

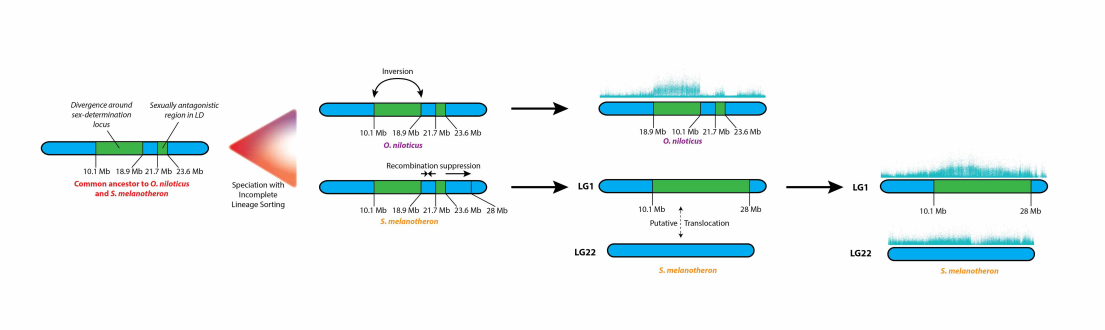


Figure 8. *Proposed model for the evolution of the linkage group 1 sex-determination system in tilapia.*

We summarize our findings by proposing a model for the evolution of the LG1 sex chromosome in tilapia (Figure 8). The sex-determination mechanism resided on LG1 in the common ancestor of *O. niloticus* and *S. melanotheron*. Early stages of sex differentiation encompassed a region on LG1 from 10.1Mb to 18.9Mb. The region from either 18.9Mb or 21.7Mb to 23.6Mb was maintained in linkage disequilibrium with sex-determination locus by sexually antagonistic selection[119]. Due to a lack of shared SNPs between 18.9Mb and 21.7Mb, we are unable to determine whether this region was originally not part of the diverging region and recombination was reduced only in the *S. melanotheron* lineage (Figure 8) or this region was originally part of a block of divergence and recombination was restored to it in the *O. niloticus* lineage. *Oreochromis niloticus* and *S. melanotheron* diverged shortly after the emergence of the novel sex-determination locus. We propose that in the *O. niloticus* lineage, the region from 10.1Mb to 18.9Mb experienced a chromosomal inversion and began to rapidly accumulate mutations. In the *S. melanotheron* lineage, the region did not experience this inversion and the region of sex differentiation expanded to encompass a region from 10.1Mb to 28Mb. In *S.*

melanotheron, LG1 may have further experienced a translocation of a region from LG22.

Candidate sex-determination mutations

We identified 21 shared SNPs with consistent sex-patterned profiles in *S. melanotheron* and *O. niloticus*, which represent candidate sex-determination mutations for this lineage. Two of these are located near previously identified sex-determination candidate genes, so we evaluated how these shared sex-patterned SNPs might alter transcription factor binding sites. The mutation near *Ras association domain containing protein 10* did not affect any transcription factor binding sites known to play a role in sex-determination. The mutation near *Wilms tumor protein* altered a *Gata4* transcription factor binding site. *Gata4* is a transcription factor that has been associated with the sex-determination pathway[120]. *Gata4* has been shown to bind in the promoter of the *Wilms tumor protein* in mammals and teleosts and *Wilms tumor protein* is a strong activator of *Amhr*, a known sex-determination gene[43,121]. A Y-specific loss of a *Gata4* binding site would decrease transcription of *Wilms tumor protein* in males, which in turn would not activate *Amhr* and result in masculinization. This variant currently represents the strongest SNP candidate for sex-determination on LG1 and further functional analysis is necessary to assess its role in sex-determination.

Our analysis is limited by the inability of the short-read Illumina data to clearly resolve the large number of deletions and insertions that are likely present on the Y-chromosome. Future studies should consider using longer read sequencing technologies to assemble X- and Y-specific sequences for both species. It may then

be possible to more clearly identify conserved insertion and deletion events on the proto-Y chromosome in both species.

Conclusion

This study utilized a comparative approach in order to identify the ancestral state of the LG1 sex-determination locus. We determined that *S. melanotheron* shares an ancestral XY sex-determination mechanism with *O. niloticus*. In both species, the highest differentiation between the X- and Y-chromosomes is found in the region from 10.1Mb to 18.9Mb on LG1. We identified 21 sex-patterned SNPs shared between the two species. One of these SNPs alters a *Gata4* transcription factor binding site near *Wilms tumor protein,* which might alter the function of the sex-determination pathway. Future studies should assess the role of this candidate SNP in sex-determination. Investigation of additional species segregating this sex-determination system on LG1 will provide a clearer understanding of evolutionary processes during the early stages of sex chromosome divergence.

Ethics Approval

All animal procedures were conducted in accordance with University of Maryland IACUC Protocol #R-10-73. All the experimental procedures took place in CIRAD facilities in Montpellier under the Laboratory agreement for animal experimentation number A-34-172-24 and the author's personal authorization for animal experimentation N° 35-15, both delivered by the French Government.

Availability of Data and Materials

The *Oreochromis niloticus* male and female reads are available from NCBI under SRR1606298 and SRR1606304, respectively. The *Sarotherodon melanotheron*

male and female reads are available NCBI under SRR3473459 and SRR3473461, respectively.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by the National Science Foundation under Grant Number DEB-1143920 and through a U.S. Department of Education Graduate Assistance in Areas of National Need fellowship offered through the University of Maryland Biology Department under Award Number P200A150160.

Authors' contributions

WJG helped conceive the study and was involved in DNA extraction, DNA quantification, computational analyses and drafting of the manuscript. MAC was involved in computational analyses and editing of the manuscript. HD raised the specimens, collected the fin clips and editing of the manuscript. J-FB raised the specimens and collected the fin clips. TDK helped conceive the study, constructed the pooled DNA libraries, guided the analysis and helped draft the manuscript. All authors have read and approved the manuscript.

Acknowledgements

The authors acknowledge the University of Maryland supercomputing resources (http://www.it.umd.edu/hpcc) made available for conducting the research reported in this paper. The authors would like to thank Aston Belew for his programming advice.

Additional Files

Additional files for this chapter can be found at:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3163-7

Chapter 4: Novel sex chromosomes in three cichlid fishes from Lake Tanganyika

Previously published as: Gammerdinger, W. J., Conte, M. A., Sandkam, B. A., Ziegelbecker, A., Koblmüller, S., Kocher, T. D. "Novel sex chromosomes in three cichlid fishes from Lake Tanganyika." *Journal of Heredity* (2018).

Abstract

African cichlids are well known for their adaptive radiations, but it is now apparent that they also harbor an extraordinary diversity of sex chromosome systems. In this study, we sequenced pools of males and females from species in three different genera of cichlids from Lake Tanganyika. We then searched for regions that were differentiated following the patterns expected for sex chromosomes. We report two novel sex chromosomes systems, an XY system on LG19 in *Tropheus* sp. 'black' and a ZW system on LG7 in *Hemibates stenosoma*. We also identify a ZW system on LG5 in *Cyprichromis leptosoma* that may be convergent with a system previously described in Lake Malawi cichlids. Our data also identify candidate SNPs for the blue/yellow tail color polymorphism observed among male *Cyprichromis leptosoma.*

Background

Theories about the evolution of sex chromosomes have been strongly influenced by the early discovery of old and highly heterochromatic systems, including the XY system found in eutherian mammals and the ZW system found in carinate birds. Each of these systems arose more than 100 million years ago, giving

ample time for the genomic landscape around the sex locus to be sculpted by a variety of evolutionary processes[122]. These processes include sexually antagonistic selection, as well as the suppression of recombination between the sex-determination locus and nearby sexually antagonistic genes[18]. Beneficial combinations of sexually antagonistic alleles linked to the sex-determination gene are selectively favored and may help drive a novel sex-determination locus to fixation. At the same time selection favors mechanisms, such as inversions, that more tightly link these favorable alleles with the sex locus[18,123]. This reduction in recombination leads to an accumulation of deleterious mutations and repetitive elements and therefore the gene content of these novel sex chromosomes begins to decay via Muller's ratchet[19].

In striking contrast to the ancient sex chromosomes of mammals and birds, the sex chromosomes of many fishes are young and experience rapid turnover[108,124]. Species within the same family, or even within the same genus, have often evolved different sex-determination loci. In ricefish (genus *Oryzias)*, at least three different systems of sex determination have evolved in the last 15 million years[41]. Despite this diversity of sex chromosomes, many of the sex-determination loci identified in fishes also play a critical role in mammalian sexual differentiation, suggesting a deep conservation of the gene network underlying sex-determination[39–43,45,125].

African cichlids are best known for their spectacular adaptive radiations, but they have also evolved a diversity of sex chromosomes[52,54–58,62,63]. Three sex chromosome systems, on three different linkage groups (LG1 XY, LG3 ZW and LG23 XY), have been characterized within the tilapia clade of African cichlids[45,53,59,64,65,67–69,109,126]. The causative locus for the LG23 system has been

shown to contain a duplication of *Amh*[45]. Additionally, many species of tilapia segregate multiple sex-determination genes. For example, *Oreochromis aureus* segregates both the LG1 and LG3 sex chromosome systems within families, while sex has been mapped to either LG1 or LG23 in different strains of *O. niloticus*[45,51,53,64,65,67,109,126]. The haplochromine cichlids of Lake Malawi also display a plethora of sex-determination mechanisms, including a ZW system on LG5 and XY systems on LG7 and LG20[54–57]. In addition, some species carry a female-limited B chromosome that may function as a dominant female sex-determination gene[62]. Because the sex chromosome systems of tilapia and the cichlid species of Lake Malawi do not overlap, it is not yet possible to identify the ancestral state. It is therefore difficult to reconstruct the evolutionary history of sex chromosome transitions within cichlids.

Lake Tanganyika is inhabited by several lineages of cichlids that may provide insight into how sex-determination evolved in haplochromine cichlids after they diverged from their last common ancestor with tilapia. Previous research on Lake Tanganyika cichlids is limited to studies of the haplochromine *Astatotilapia burtoni*[57,58], which is not particularly common in the lake proper, but it is one of the dominant species in the tributaries of the lake. Three systems within *A. burtoni* have been reported: an XYW system on LG13, an XY system on LG18, as well as an XY locus located on a fusion between LG5 and LG14[57,58].

The Lake Tanganyika cichlid species flock comprises more than a dozen tribes - distinct lineages with unique ecologies and life histories[127–130]. This report explores the diversity of sex chromosomes in species representing three tribes:

Tropheus sp. 'black' (Tropheini), *Hemibates stenosoma* (Bathybatini) and *Cyprichromis leptosoma* (Cyprichromini). Recent work has shown that the tribe Tropheini is nested within the Haplochromini and is closely related to a monophyletic assemblage that includes the adaptive radiations in the other East African Great Lakes along with several riverine cichlid lineages, including *A. burtoni*[131–133]. The Haplochromini, together with a number of other tribes, including the Cyprichromini, originated in the so called primary lacustrine radiation in Lake Tanganyika, whereas the Bathybatini are an ancient tribe whose origin predates the primary Lake Tanganyika radiation[134–136]. The phylogenetic relationship of the species studied is shown in Figure 9[130]. All three study species are maternal mouth brooders. *Cyprichromis leptosoma* and *Hemibates stenosoma* exhibit a pronounced level of sexual dichromatism, whereas *Tropheus* sp. 'black' are sexually monochromatic, similar to other *Tropheus* but contrary to most other Tropheini or Haplochromini[137].

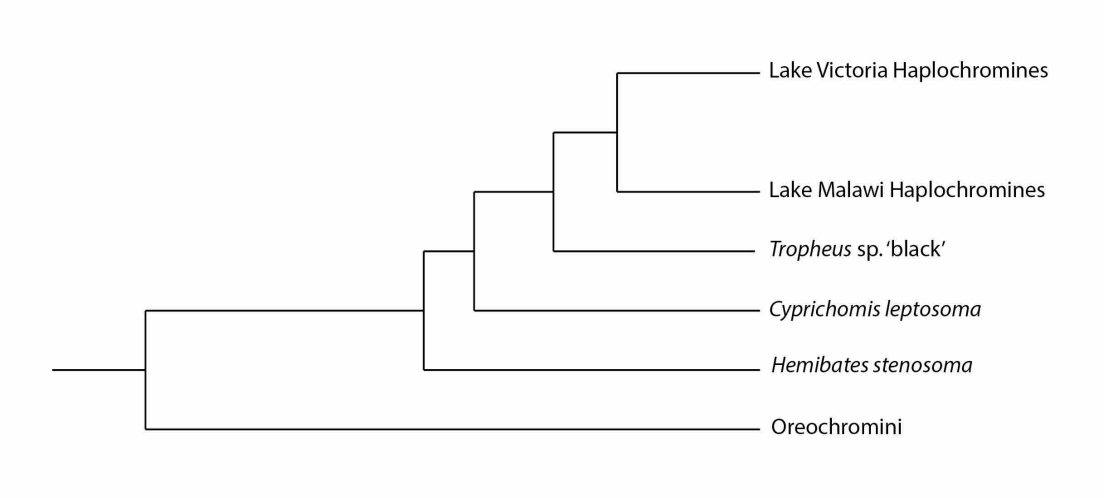


Figure 9. *A phylogenetic tree of the species studied. Adapted from Meyer et al., 2015.*

Cyprichromis leptosoma has a distinct fin color dimorphism in males. A previous RAD mapping study had mapped this trait to a 161kb region of LG16[138].

Yellow males are homozygous, while blue males are heterozygous or homozygous for an alternate allele across this region suggesting a single locus trait with two alleles[138]. We now provide a characterization of variants within this region.

Materials and Methods

Materials

All of the tissues used were sampled from wild-caught fish collected from Lake Tanganyika. We sampled 30 males and 24 females *Tropheus* sp. 'black' from Ikola, Tanzania, 25 males and 25 females *Hemibates stenosoma* from Mpulungu, Zambia and 26 males and 30 females *Cyprichromis leptosoma* near Kalambo Lodge in Zambia. Of the 26 male *C. leptosoma* samples, 17 were from individuals with the blue color morph and 9 were from individuals with the yellow color morph.

Sequencing

DNA was purified from fin clips by phenol-chloroform extraction using phase-lock gel tubes (5Prime, Gaithersburg, Maryland). Fluorescence spectroscopy was utilized to quantify DNA concentrations for each individual in order to ensure equal representation of each individual in the pooled libraries. Males and females from each species were pooled separately and the male *C. leptosoma* pool was further divided into separate pools based upon their color morph. The libraries were constructed using the TruSeq DNA PCR-Free LT Kit (Illumina, San Diego, California). Illumina 100bp paired-end sequencing was carried out with one lane per species. The *C. leptosoma* libraries received an additional lane of sequencing.

Read mapping

Reads were filtered using the CASAVA 1.8 filter. Reads that did not pass this criterion were removed. Read qualities were visually assessed using FastQC[78]. Reads were then mapped to an *Oreochromis niloticus* reference genome (O_niloticus_UMD1) with BWA version 0.7.12 using the default parameters along with read group labels[69,139]. The alignments were sorted, marked for duplicates and indexed using Picard version 1.119[81].

Variant identification

Alignments were converted into an mpileup file using Samtools version 0.1.18 and subsequently into a sync file using Popoolation2[83,140]. Base calls with a PHRED score less than 20 were filtered out of the data set. Each species was then separately analyzed with Sex_SNP_finder_GA.pl (https://github.com/Gammerdinger/sex-SNP-finder) for both XY- and ZW-patterns in order to find sex-patterned SNPs that could be used to determine the likely sex-determination system as well as to measure F_{ST} across the genome[126]. The density of sex-patterned single nucleotide polymorphisms (SNPs) was measured in 10kb non-overlapping windows using a minimum coverage threshold of 10. The sex system (XY or ZW) was determined by identifying the Sex_SNP_finder_GA.pl output with the greatest number of sex-patterned SNPs in regions of F_{ST} differentiation. We analyzed the distribution of sex-patterned SNPs in non-overlapping 10kb windows for each species and assigned a threshold near the middle of the bimodal distribution, which we used to call regions of high differentiation.

In parallel to the aforementioned methods, variants were called from the pool alignments with GATK version 3.7 using *HaplotypeCaller* with the maximum allowed ploidy for a sample set to ten[82]. Males and females of a species were combined using GATK's *GenotypeGVCF* command to assemble a VCF file of all variants in each species[82]. We then filtered the output VCFs from GATK's *GenotypeGVCF* using the list of sex-patterned SNPs in each species. These sex-patterned SNPs were then filtered using the BEDtools *intersect* command to identify whether they were located within the regions of high differentiation [115].

The color locus in *C. leptosoma* also was evaluated using Sex_SNP_finder_GA.pl, however the parameters were set to search for a dominant-recessive pattern. Since it has been previously reported that the blue color morph is dominant to the yellow color morph, we looked for sites in the genome at which the yellow morph was fixed, or nearly fixed, and the blue morph had an alternative allele with a frequency between 0.3 and 1.

Functional annotation

Sex-patterned SNPs were evaluated for functional significance using SnpEff version 4.3 and subsequently filtered by their functional impacts using SnpSift[85,86]. SNPs with non-synonymous impacts were further evaluated using PROVEAN version 1.1.5 in order to predict whether or not these missense SNPs were deleterious[114]. Scores below the recommended PROVEAN threshold of -2.5 were considered deleterious. Low quality gene annotations caused us to remove from the analysis 34 sex-patterned missense mutations in *Tropheus* sp. 'black' (in proteins XP_005461945.2, XP_005477739.2, XP_003457270.2, XP_005461986.2,

XP_005477064.2 and XP_019204497.1), 21 in *H. stenosoma* (in proteins XP_013129257.2, XP_019213473.1, XP_013119637.2, XP_013130691.2, XP_019217006.1 and XP_005450963.2) and 40 in *C. leptosoma* (in proteins XP_019214201.1, XP_003441585.2, XP_013124716.2, XP_013124709.2, XP_019214472.1, XP_003441539.3 and XP_013128682.2).

Copy number variant identification

Mean coverage of read alignments in each male or female pool was quantified with Picard's CollectWgsMetrics[81]. We then ran Varscan 2 version 2.3.7 to identify copy number variants using a windowed approach, taking into account the differences in read coverage in each pool. We used a minimum coverage of 8, a minimum window size of 100bp and a maximum window size of 1kb[117]. The *amp* and *del* thresholds were both set to 0.2.

Defining the color locus region in *Cyprichromis leptosoma*

Primer sequences from a previous mapping study were run through BLAST to find the corresponding region in the most recent *O. niloticus* assembly using the *–task blastn-short* setting in BLAST[138,141]. The 16 primer sequences correspond to 6 microsatellite markers (Sg07, Sg08, Sg09, Sg11, Sg13, Sg16) and 2 RAD markers (RAD206877 and RAD208776)[138]. The analysis of the color locus spanned the entirety of the previously mapped region of significance between Sg07 and Sg16.

Results

Sequencing and read mapping

We obtained ~100 million and ~107 million 100bp paired-end reads from *Tropheus* sp. 'black' males and females, respectively. The mean coverage for the

male reads was 12.03X and the mean coverage for female reads was 12.88X when aligned to *O. niloticus*.

Sequencing of *H. stenosoma* yielded ~102 million and 79 million reads from males and females, respectively. The mean male coverage in *H. stenosoma* was 12.46X and the mean female coverage was 8.68X when aligned to *O. niloticus*.

The sequencing for yellow *C. leptosoma* males yielded ~116 million 100bp paired-end reads, while blue males yielded ~71 million reads. The *C. leptosoma* females yielded ~167 million 100bp paired-end reads. The estimated coverage in the yellow males was 13.39X, in the blue males was 8.13X and in the females was 19.25X, when aligned to *O. niloticus*.

Differentiation in *Tropheus* sp. 'black'

The distribution of non-overlapping windows containing sex-patterned SNPs in *Tropheus* sp. 'black' showed a strongly bimodal signal in the XY dataset (Supplementary file 1). Non-overlapping windows with a density of XY-patterned SNPs of 25 or greater were designated as regions of high differentiation in this species. By this criterion, ~19Mb of the *Tropheus* sp. 'black' genome is in highly differentiated regions (Figure 10). The vast majority of XY-patterned SNPs fall within an ~18.9Mb region of increased F_{ST} on LG19 (Figure 11). No other linkage groups had non-overlapping, 10kb windows with more than 25 sex-patterned SNPs. Eight other contigs of the genome assembly that were not anchored to the linkage map had only one sex-patterned window each. From the Varscan analysis we can see that the overall percentage of copy number variants, duplications and deletions

combined, was higher within the region of high differentiation compared to the rest of the genome (Table 4).

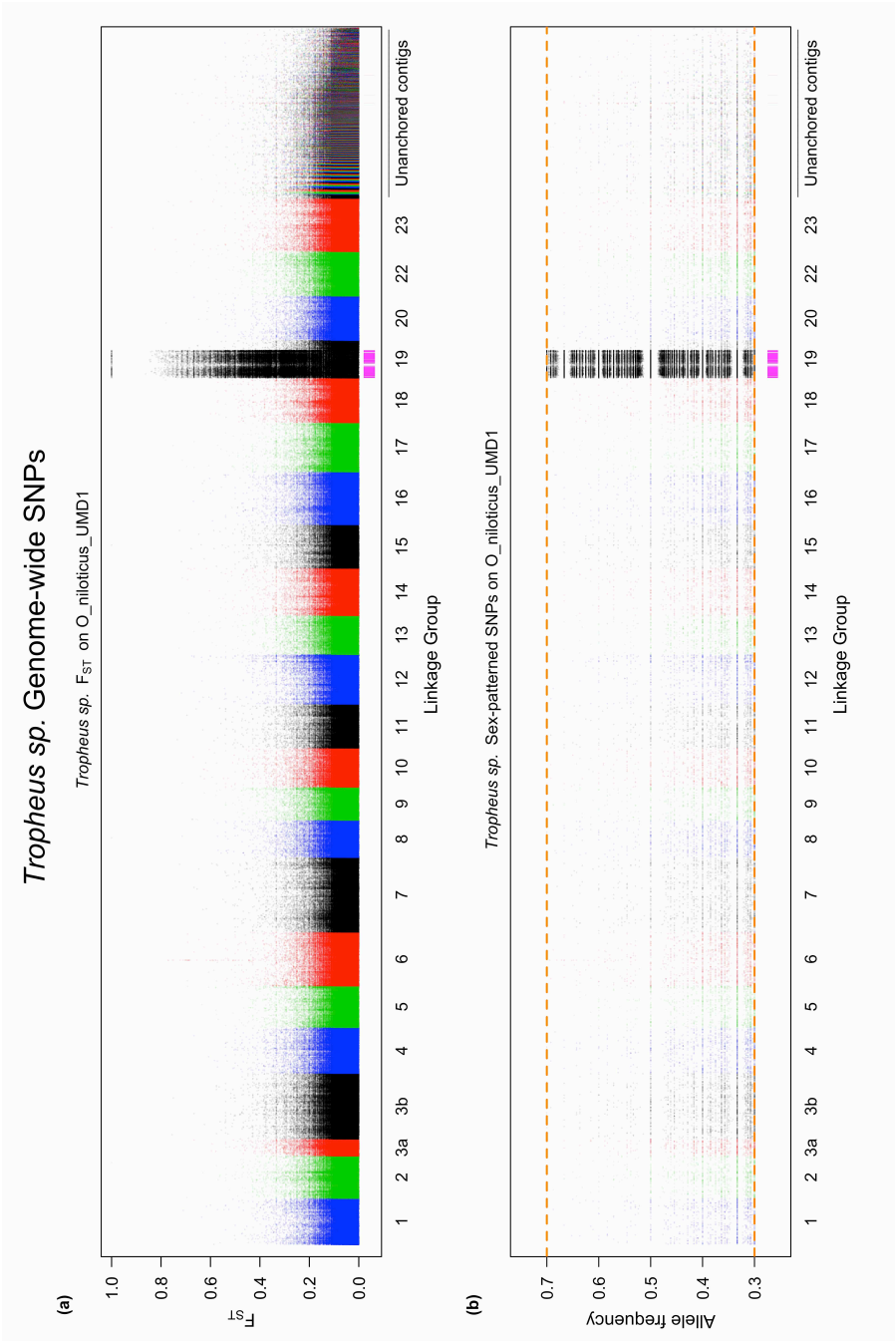


Figure 10. *A genome-wide scan for Tropheus sp. 'black' evaluating (a) F_{ST} and (b) sex-patterned SNPs in the XY direction. The rectangles underneath each figure identify the regions of high differentiation.*

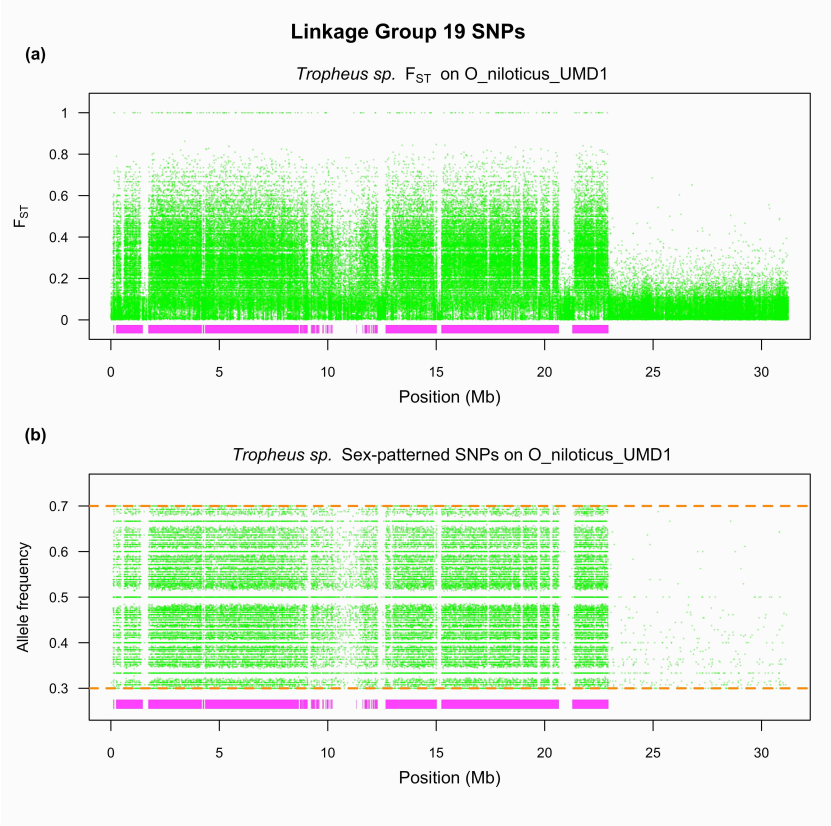


Figure 11. *A scan of LG19 for Tropheus sp.'black' evaluating (a) F_{ST} and (b) sex-patterned SNPs in the XY direction. The rectangles underneath each figure identify the regions of high differentiation.*

Table 4. Results from the VarScan 2 analysis evaluating coverage differences between male and female pools inside and outside the regions of differentiation in *Tropheus* sp. 'black', *H. stenosoma* and *C. leptosoma*.

***Tropheus* sp. 'black' Copy Number Variants**			
	Duplication	Deletion	Equal Coverage
Inside Regions of High Differentiation	3,089,106bp (21.8%)	2,772,768bp (19.6%)	8,306,306bp (58.6%)
Outside Regions of High Differentiation	92,003,122bp (16.1%)	113,351,692bp (19.9%)	365,178,238bp (64.0%)
***H. stenosoma* Copy Number Variants**			
Inside Regions of High Differentiation	5,862,875bp (19.9%)	6,838,210bp (23.2%)	16,830,707bp (57.0%)
Outside Regions of High Differentiation	104,174,455bp (19.6%)	109,260,371bp (20.6%)	316,942,356bp (59.8%)

C. leptosoma Copy Number Variants			
Inside Regions of High Differentiation	3,016,849bp (12.7%)	4,977,415bp (21.0%)	15,742,022bp (66.3%)
Outside Regions of High Differentiation	66,895,323bp (10.6%)	129,432,080bp (20.5%)	435,280,746bp (68.9%)

We detected 2,062 genomic mutations responsible for creating 4,740 missense transcript mutations. PROVEAN predicted that 1,415 of these 4,740 missense transcript mutations were deleterious. An additional 35 genomic mutations were responsible for 59 "high-impact" transcript mutations as defined by SnpEff. A full list of these "high-impact" and predicted deleterious mutations can be found for all of the species in Supplementary file 2.

Differentiation in *Hemibates stenosoma*

Our analysis of the distribution of non-overlapping windows spanning sex-patterned SNPs in *H. stenosoma* revealed a bimodal signal in the ZW dataset (Supplementary file 3). We considered 10kb windows with a sex-patterned SNP density of 14 or greater as regions of high differentiation. This analysis shows that ~42.8Mb of the *H. stenosoma* genome lies within these regions of differentiation (Figure 12). ~37.5Mb of this differentiation resides on LG7 (Figure 13), while ~1.6Mb is on LG4 and 1.1Mb is on contig245. 13 additional linkage groups had at least one of these highly differentiated windows, but none had more than six windows. 61 unanchored contigs had one differentiated window, while another unanchored contig had two differentiated windows. The high density of ZW-patterned SNPs within these regions on LG7 correspond with an elevation in F_{ST} (Figure 13). There is an increase in the overall percentage of copy number variants,

duplications and deletions combined, from the region of high differentiation when compared to the rest of the genome (Table 4).

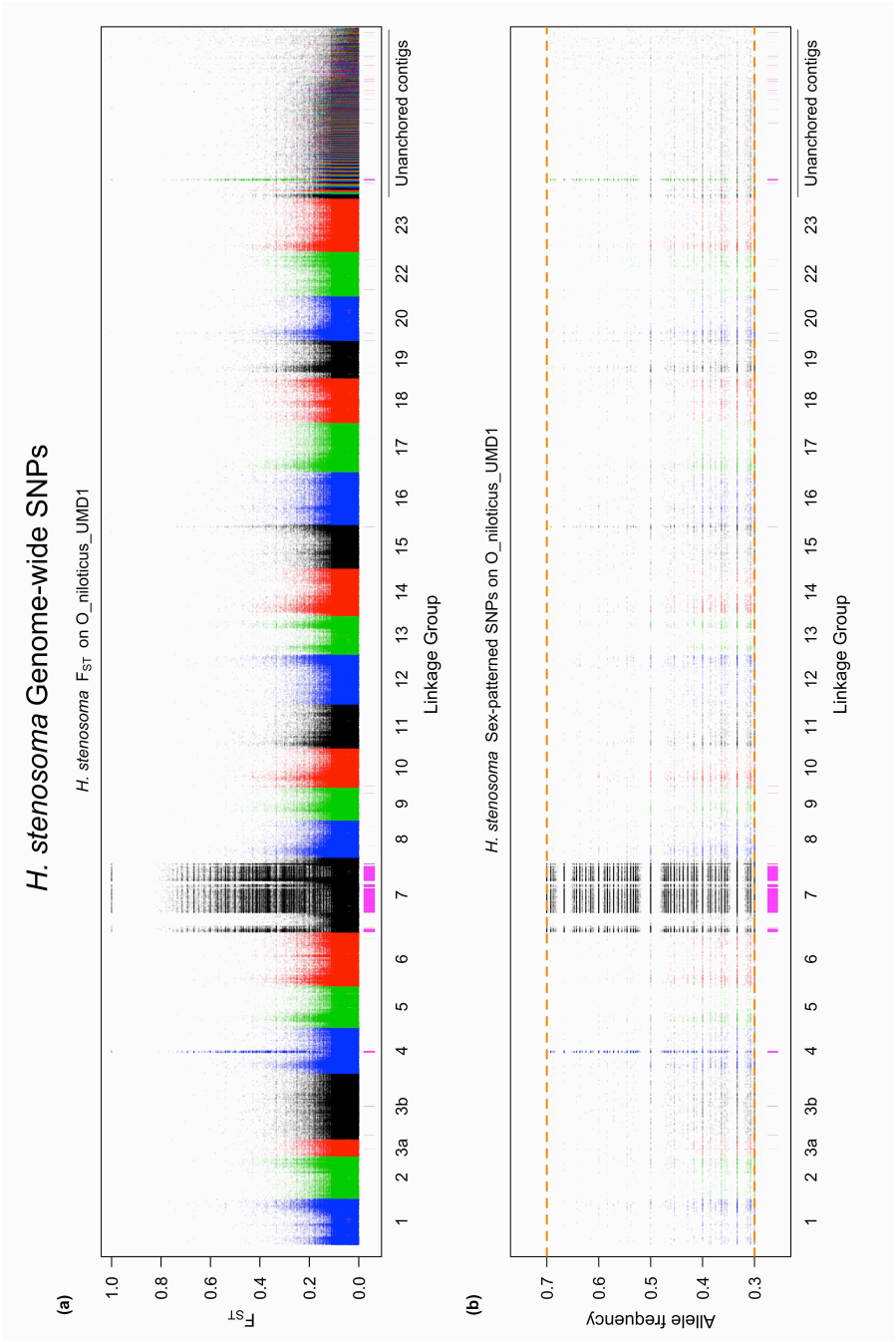


Figure 12. *A genome-wide scan for H. stenosoma evaluating (a) F_{ST} and (b) sex-patterned SNPs in the WZ direction. The rectangles underneath each figure identify the regions of high differentiation.*

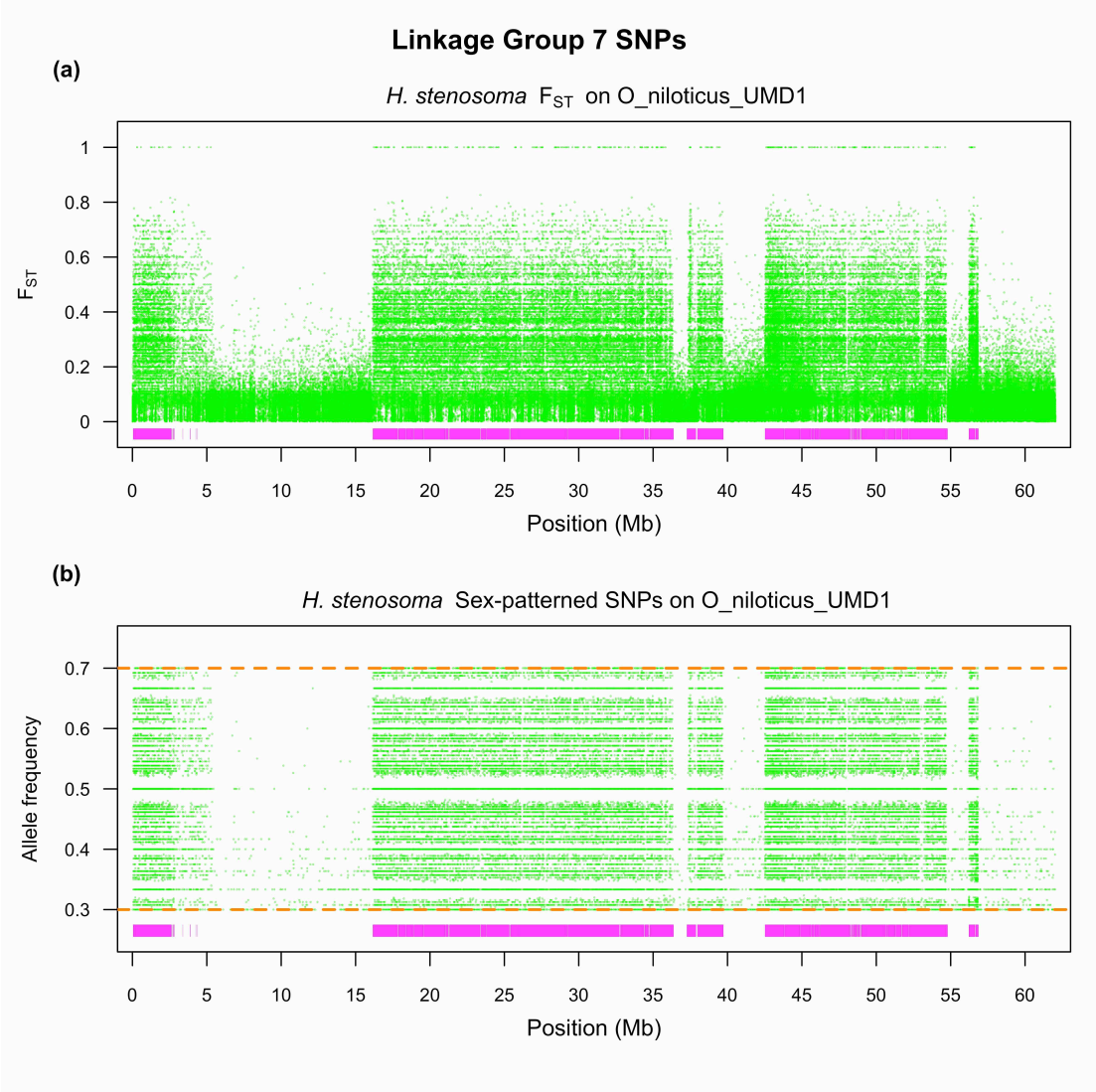


Figure 13. *A scan of LG7 for H. stenosoma evaluating (a) F_{ST} and (b) sex-patterned SNPs in the WZ direction. The rectangles underneath each figure identify the regions of high differentiation.*

We found 1,901 genomic mutations causing 4,664 missense transcript mutations in their associated proteins. PROVEAN analysis determined that 1,557 of these 4,664 missense transcript mutations were predicted to be deleterious. Additionally, we found 38 genomic mutations that created 69 “high-impact” transcript mutations.

Differentiation in *Cyprichromis leptosoma*

When we analyzed the distribution of non-overlapping windows containing sex-patterned SNPs in *C. leptosoma*, we determined that the distribution of non-overlapping ZW-patterned windows was bimodal and any windows with a density of 22 sex-patterned SNPs or greater were assigned to the category of differentiated (Supplementary file 4). Using this benchmark, *C. leptosoma* had ~29.6Mb of differentiation between males and females. Two regions showed particularly elevated F_{ST} (Figure 14). The first region of differentiation is located on LG5 (Figure 15) and totals ~26.4Mb. A few additional regions of high differentiation from LG13 (Figure 16) amounted to ~1.1Mb. While 28 unanchored contigs provided at least 1 differentiated window, the strongest signal from the unanchored contigs came from contigs313 and contig400, which had ~573kb and 379kb of differentiation, respectively. Seven additional linkage groups had at least one sex-patterned window meeting our criteria, however six of these only had one window and another was responsible for only ~87kb of differentiation. The differentiation as measured by F_{ST} is attributable to the high density of ZW-patterned SNPs within these regions of differentiation. Additionally, an analysis of the Varscan output shows an increase in the total number of copy number variants, duplications and deletions combined, within the region of high differentiation as compared to the rest of the genome (Table 4).

Within the regions of high differentiation we detected 1,861 genomic mutations that were responsible for 4,519 missense transcript mutations. PROVEAN predicted that 1,200 of these 4,519 missense transcript mutations had a predicted

deleterious effect on their associated proteins. Our analysis also reveals 32 genomic mutations that are scored as having a "high-impact" on gene function in 74 transcripts models.

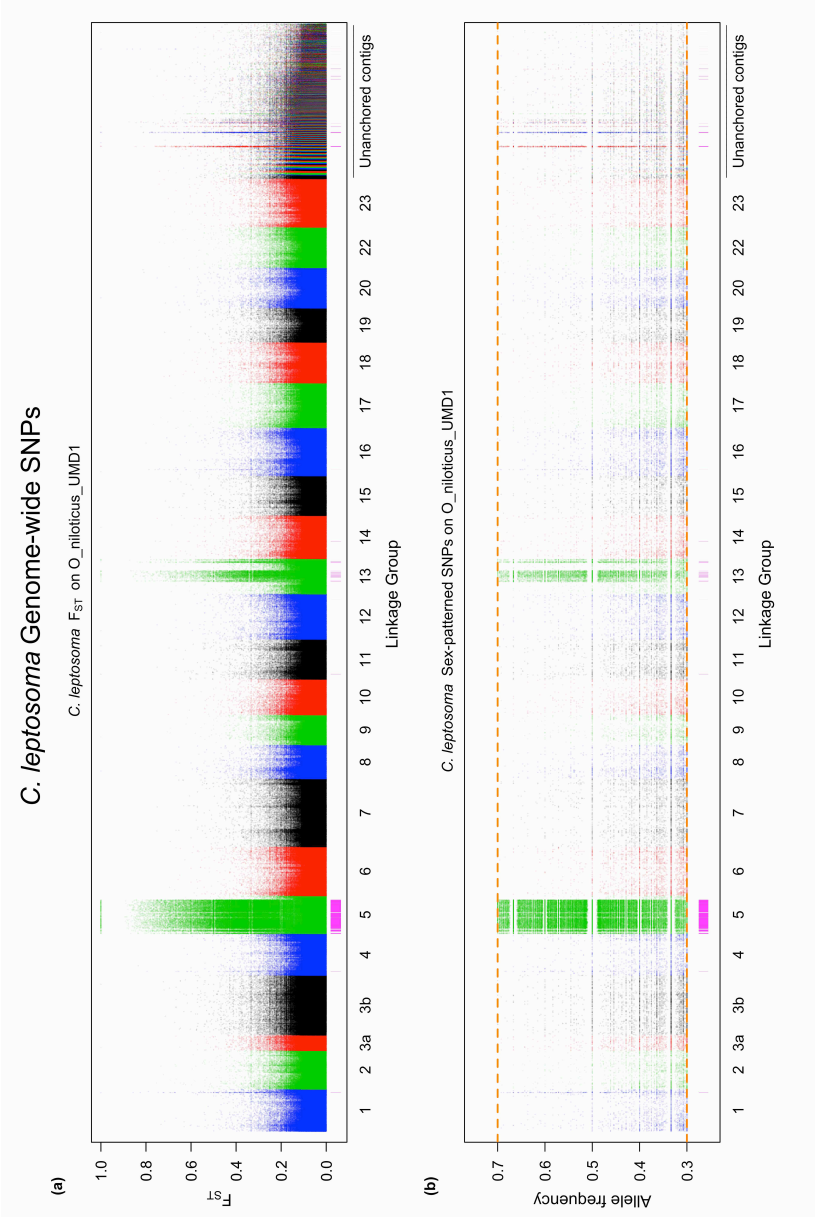


Figure 14. *A genome-wide scan for C. leptosoma evaluating (a) F_{ST} and (b) sex-patterned SNPs in the WZ direction. The rectangles underneath each figure identify the regions of high differentiation.*

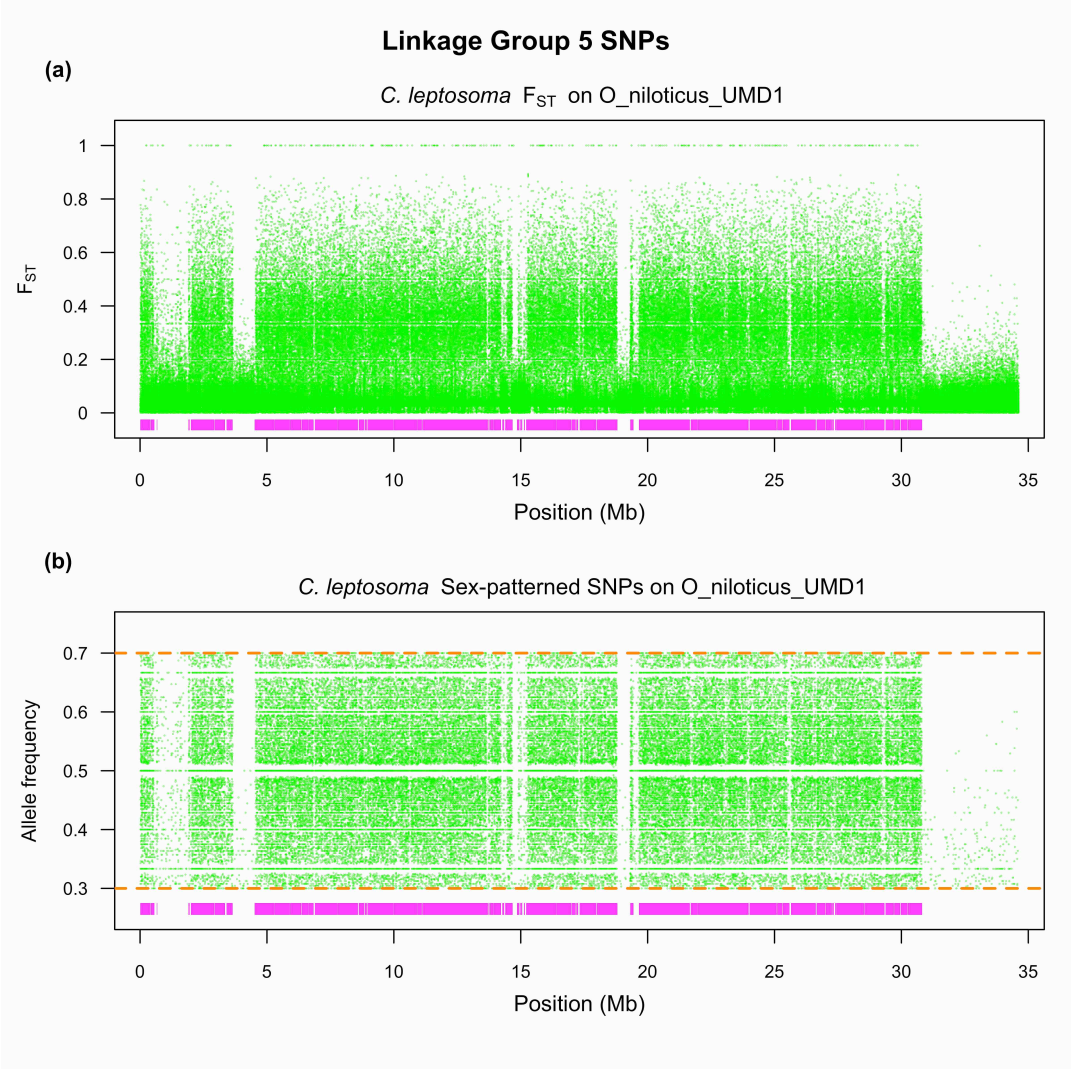


Figure 15. *A scan of LG5 for C. leptosoma evaluating (a) F_{ST} and (b) sex-patterned SNPs in the WZ direction. The rectangles underneath each figure identify the regions of high differentiation.*

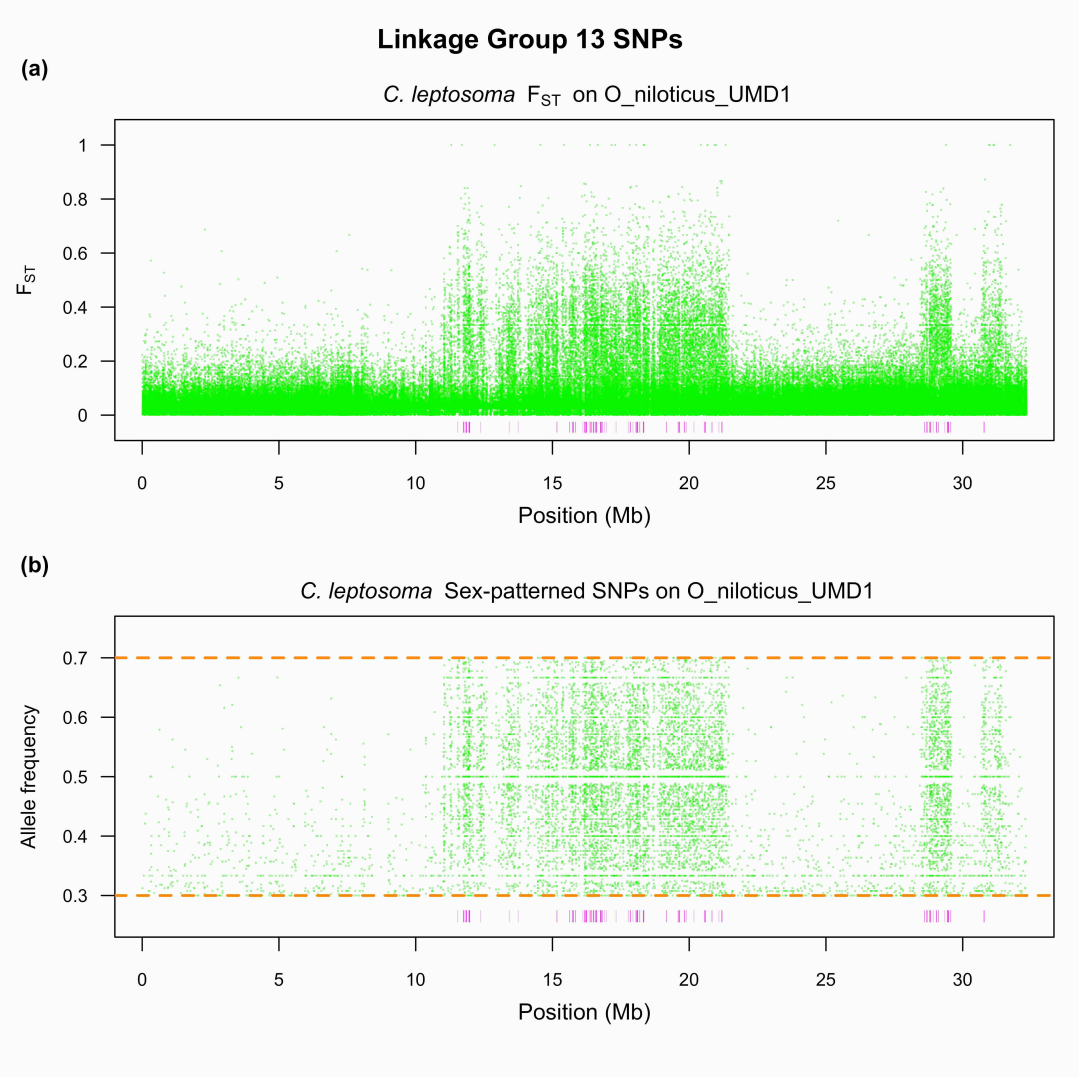


Figure 16. *A scan of LG13 for C. leptosoma evaluating (a) F_{ST} and (b) sex-patterned SNPs in the WZ direction. The rectangles underneath each figure identify the regions of high differentiation.*

Male color polymorphism in *C. leptosoma*

14 of the 16 primer sequences were mapped to linkage group 16 between 8.57Mb and 8.81Mb and retained synteny between the previous study and the more recent genome assembly[138]. Within this region we identified 218 SNPs with a pattern consistent with blue dominance. Of these SNPs, 11 genomic SNPs were responsible

for 13 missense transcript mutations in their associated gene models, but none were scored as having a "high-impact" as defined by SnpEff. Seven of these thirteen missense transcript mutations were scored as deleterious (Supplementary file 2).

Discussion

LG19 in *Tropheus* sp.'black'

The high number of XY sex-patterned SNPs and increased copy number variants provide strong evidence that LG19 is a sex chromosome in *Tropheus* sp. 'black'. This XY system on LG19 represents a previously unreported sex-determination system in African cichlids.

Within the regions of differentiation on LG19, there were several mutations scored by SnpEff as having a "high-impact" on gene function that are interesting for understanding the emergence and evolution of sex chromosomes. The first mutation is a splice junction variant on LG19 at 15,270,722 for *cathepsin L1*. *Cathepsin L1* is highly expressed in Sertoli cells of rats. Disruption of *cathepsin L1* leads to increased apoptosis of adult germ cells, but it is unclear if this is also true of primordial germ cells[142,143]. A common theme observed in sex differentiation is that males have fewer primordial germ cells in development and this appears to lead to the differentiation of testes[144]. Therefore, alterations of genes that affect germ cell numbers might create novel sex-determination genes[48]. Another interesting gene is *apolipoprotein B-100*, which has accumulated three nonsense mutations on the Y-chromosome. The product of this gene is secreted from granulosa cells in humans and improves the fertility of women undergoing in vitro fertilization[145]. While this gene has not been shown to have a deleterious effect in males, this gene could represent a gene that experienced

sexually antagonistic selection, as females with higher expression increase their fertility at the cost of an unknown detriment to males. Therefore, the gain of a stop codon in this gene on the Y-chromosome could alleviate sexual conflict that may have arisen from this gene. Additionally, four missense mutations in *fibroblast growth factor receptor-like 1* (*FGFR1*) located at 18,130,650, 18,131,447, 18,140,975 and 18,141,011 on LG19 were all predicted to be deleterious. *FGFR1* has been implicated as a candidate gene in humans for causing ambiguous genitalia. The paralog of *FGFR1*, *FGFR2*, is a critical receptor for *FGF9*, a well-studied gene in mammalian sex-determination networks[146,147]. Lastly, within *TGF-β3* there are several Y-patterned variants, including a predicted non-deleterious missense mutation (L268P), along with four mutations in the 5'-UTR (two of which create premature start codons) and three in the 3'-UTR. The *TGF-β* pathway has been strongly implicated in altering the vertebrate sex-determination network[48,148].

LG7 in *Hemibates stenosoma*

The ZW signal on LG7 in *H. stenosoma* is quite pronounced, while the small sex-patterned signal from LG4 likely represents a structural rearrangement between *H. stenosoma* and the reference genome. There is a high density of ZW-patterned SNPs and copy number variants inside the regions of differentiation. The identification of a ZW system on LG7 is particularly interesting because of previous reports of an XY system on LG7 in some Lake Malawi species[54]. This is yet another indication that the sex chromosomes of African cichlids are turning over rapidly.

None of the "high-impact" mutations found in the ZW-patterned SNPs showed a clear connection to the sex-determination network. There were two W-

patterned, missense mutations of note that were predicted to be deleterious. One is a deleterious missense mutation in *Wnt-2* on LG4 at position 17,579,780. *Wnt-2* signaling with *beta-catenin* promotes granulosa cell proliferation and *Wnt-2* is expressed in rainbow trout ovaries during gonadal differentiation[149,150]. However, it is not clear how this mutation would lead to female differentiation. Another predicted deleterious, missense mutation was within *Bardet-Biedl syndrome protein 1* on LG8 at position 9,112,074. Bardet-Biedl syndrome is an autosomal recessive disorder in humans with symptoms including obesity, mental retardation, hypogonadism and reproductive abnormalities[151]. Neither of these mutations is on LG7 of the reference genome and so these may suggest structural rearrangements since the divergence between tilapia and *H. stenosoma*. Also, *brain aromatase* has a W-patterned SNP in the 3'-UTR. *Ovarian aromatase* is well known within cichlids for being a key gene involved with sex differentiation in several species, but *brain aromatase* does not appear to be differentially expressed in the gonads of Nile tilapia[152–156]. The genes for *SMAD2*, *SMAD3, SMAD4*, *SMAD6* and *SMAD7* are all within the regions of high differentiation and have been shown to interact with the *TGF-β* signaling pathway, but lack any deleterious missense or "high-impact" mutations[96,157].

LG5 in *Cyprichromis leptosoma*

The strongest signal for sex-determination in *C. leptosoma* is found on LG5 in the form of a ZW system. This region overlaps the previously reported sex-determination region on LG5 in orange-blotched Lake Malawi cichlids and thus may represent a shared ancestral sex-determination system or be an example of convergence[54]. In *C. leptosoma* there is also a ZW signal coming from LG13 (Figure

14). Upon inspecting LG13, it is clear that there is a block of ZW-patterned SNPs, but the density of these SNPs is lower than on LG5. This fainter LG13 signal could therefore represent a more recent translocation of a region of LG13 to LG5. Alternatively, it may correspond to all or a part of the XYW system previously reported on LG13 in *A. burtoni* or another sex chromosome system[58].

None of the genes with "high-impact" mutations in the *C. leptosoma* sex comparison had any known ties to the sex-determination network. However, there was a deleterious missense mutation in *Wnt7a* protein on the Z-chromosome. *Wnt7a* has been previously tied to abnormal development of the female reproductive tract in mice, with homozygous mutants showing significant abnormalities[158]. Thus, this mutation may push development of ZZ individuals towards males. The most interesting candidates for sex-determination in this species are *WNT4* and *FGFR2*. *WNT4* is in the differentiated region on LG5, and has one W-patterned variant in the 5'-UTR and two W-patterned variants in the 3'-UTR. *WNT4* has been shown to be one part of a key genetic switch in the mammalian sex-determination network[146,159]. *FGFR2* is in the differentiated region on LG13 and has a predicted non-deleterious W-patterned missense (Proline to Alanine) mutation. *FGFR2* is the main receptor for *FGF9*. Mutations in *FGFR2* in humans and mice demonstrate a complete sex reversal[160,161].

Limited options?

Some have argued that the evolutionary history of sex chromosomes in vertebrates can be best explained by shared ancestry of sex chromosomes or a limited core set of genes capable of becoming sex-determination genes[162]. However, the work

presented here challenges both of these paradigms. We studied three species from Lake Tanganyika and discovered at least two novel sex chromosome systems. Research on African cichlids has now identified at least 9 autosomes that have become sex chromosomes in the last 15 million years: LG1 (XY), LG3 (ZW), LG5 (ZW), LG7 (distinct XY and ZW systems), LG13 (XYZ), LG18 (XY), LG19 (XY), LG20 (XY), LG23 (XY) and possibly a female-determining B chromosome[51,53–59,62,64,65,67,68,109,126]. These data show that there are many genes scattered across the genome that can become top-level sex-determination genes in fishes, and suggests that many additional sex-determination genes remain to be discovered among the African cichlids. It could be argued that the discovery of a novel ZW system on LG7 represents a situation similar to the *DMRT1* gene, which has been co-opted to become both a ZW system (*DM-W* in *Xenopus*), and an XY system (*DMRT1Y* in medaka)[13,40]. The possibly convergent or ancestrally shared LG5 ZW might also be construed as support of the notion of limited options in the sex-determination network. However, as more sex chromosome systems are found within African cichlids, some are likely to be found on the same chromosomes by chance rather than because of any innate predisposition to becoming sex chromosomes. The extraordinary diversity of sex chromosome systems witnessed in African cichlids seems to defy current paradigms for the emergence of novel sex chromosome systems. Sexual antagonism has been proposed as a mechanism aiding the emergence of novel sex chromosomes. Cichlids show varying levels of sexual dimorphism with respect to parental investment. It follows that cichlids are a useful model for studying the role of sexual antagonism in the emergence of new sex chromsomes[52,163].

Cyprichromis color polymorphism

While our analysis did not reveal any "high-impact" mutations within the region of interest for color in *C. leptosoma*, there was an intriguing predicted deleterious missense mutation. This missense mutation was in *exportin-4*, at 8,747,219, which is responsible for the nuclear import of *Sox* proteins in mammals[164]. The *Sox* proteins known to utilize *exportin-4* are *SRY* and *Sox2*, but it is possible that other *Sox* proteins also utilize this protein[164]. If *exportin-4* regulates *Sox10* nuclear import, mutations in *exportin-4* might disrupt the specification of pigment cells from the neural crest to produce the observed color phenotype [165,166].

Caveats

Our catalog of sex-patterned SNPs is subject to three kinds of error. First, at low genomic coverage, the total number of sex-patterned SNPs is likely underestimated. Second, some positions may be classified as sex-patterned due to sampling error in the relatively small number of individuals of each sex that we studied. Third, our relatively low coverage combined with the inherent variance in Illumina coverage means that we may have overestimated the number of copy number variants in our samples. However, the regions of high differentiation between our samples consistently harbored more copy number variants compared to the rest of the genome. This is consistent with other observations that sex chromosomes tend accumulate copy number variants[18,108].

We did not analyze mutations in the vast number of potential transcription factor binding sites, which could modulate expression of genes critical to sex-determination and color dimorphism. A more thorough genomic and transcriptomic analysis

consisting of several related species sharing a common sex chromosome system is needed to identify causal variants for a particular phenotype. Lastly, this analysis was carried out on the *O. niloticus* reference genome and there are likely structural rearrangements between this reference and each species analyzed.

Conclusions

This study reports the presence of three sex chromosome systems in African cichlids from Lake Tanganyika. Two of these systems, LG7 ZW and LG19 XY, appear to be previously unknown and the third may be convergent on a system identified previously in Lake Malawi. The majority of cichlid lineages from Lake Tanganyika have yet to be analyzed for sex chromosomes, but we suggest that further sampling of Lake Tanganyika cichlids would identify many additional sex chromosomes. The lack of shared sex systems among lineages makes it difficult to identify the ancestral state for sex chromosomes in Lake Tanganyika. Analyses of additional species are needed to understand ancestral states and the patterns of evolutionary transition between sex-determination systems, as well as providing a framework for understanding why African cichlids have such rapid turnover in sex chromosome systems.

Funding

This work was supported by the National Science Foundation under Grant Number DEB-1143920; U.S. Department of Education Graduate Assistance in Areas of National Need fellowship offered through the University of Maryland Biology Department under Award Number P200A150160; and the American Genetics Association through an Ecological, Evolutionary and Conservation Genomics Award.

Acknowledgements

The authors acknowledge the University of Maryland supercomputing resources (http://www.it.umd.edu/hpcc) made available for conducting the research reported in this paper. Furthermore, we would like to thank C. Sturmbauer, K.M. Sefc, C. Börger and T. Schenekar for logistics and help in fieldwork. Samples were processed in accordance with University of Maryland IACUC protocol #R-10-73.

Data Availability

All reads have been deposited in the NCBI SRA under Bioproject accession number PRJNA400462.

Authors' contributions

WJG helped conceive the study, extracted and quantified the DNA, constructed the libraries for sequencing, carried out the bioinformatic approaches and wrote the manuscript. MAC aided in the bioinformatic approaches. BAS helped quantify the DNA. SK and AZ collected the fin clip samples, sexed the fish and provided background on the study species. TDK helped conceive the study, aided in library construction and revised the manuscript.

Supplementary Files

Supplementary files for this chapter can be found at: https://academic.oup.com/jhered/advance-article/doi/10.1093/jhered/esy003/4850530

Chapter 5: Characterization of sex chromosomes in three species of tilapia (Teleostei: Cichlidae)

Submitted to Hydrobiologia as: Gammerdinger, W.J., Conte, M. A., Sandkam, B. A., Penman, D. J., Kocher, T. D. "Characterization of sex chromosomes in three species of tilapia (Teleostei: Cichlidae)"

Abstract

The African cichlid radiation has created thousands of new cichlid species with a wide diversity of trophic morphologies, behaviors, sensory systems and pigment patterns. In addition, recent research has uncovered a surprising number of young sex chromosome systems within African cichlids. Here we refine methods to describe the differentiation of young sex chromosomes from whole genome comparisons. We identified a novel XY sex chromosome system on linkage group 14 in *Oreochromis mossambicus*, confirmed a linkage group 1 XY in *Coptodon zillii* and also defined the limits of our methodology by examining the ZW system on linkage group 3 in *Pelmatolapia mariae*. These data further demonstrate that cichlids are an excellent model system for understanding the earliest stages of sex chromosome evolution.

Background

Traditionally, sex chromosomes were thought to be stable features of animal karyotypes [167]. For example, the sex chromosomes of eutherian mammals arose ~181 million years ago and remain similar in structure among species [76]. However, mounting evidence from other vertebrate groups suggests that sex chromosomes may

turnover much more rapidly in some lineages [41,53,54,59,168]. Transitions from one sex chromosome system to another begin when a mutation creates a novel sex-determination allele on an autosome. Selection on linked, sexually antagonistic alleles may help drive this novel sex-determination allele to fixation. Inversions, and other mechanisms that reduce recombination between the novel sex-determination allele and nearby sexually antagonistic variation, are selectively favored [15,108]. However, the loss of recombination also triggers Muller's ratchet, which contributes to the decay of many genes linked to the new sex-determination allele [19,21]. This decay may gradually reduce the fitness of the new sex chromosome, allowing new mutations in the sex-determination network to begin the cycle anew, a process that has been dubbed the 'hot-potato model' [22]. To better understand this process, there is a need to characterize the patterns of decay during the earliest stages of sex chromosome divergence.

The adaptive radiation of African cichlids has created diversity in trophic morphology, behavior, pigmentation and sensory systems [52]. Recently, it has been discovered that they also harbor a diverse collection of sex chromosomes. Among haplochromine cichlid species, distinct XY systems have been discovered on linkage group (LG) 7, LG18, LG20 and on a fusion of LG5 and LG14, while a ZW system has been discovered on LG5, and an XYW system found on LG13 [54–58,63]. Some species segregate multiple sex chromosome systems simultaneously [54]. Additional studies of three tribes of Lake Tanganyika cichlids have identified a ZW system on LG7, an XY system on LG19 and a possibly convergent ZW system on LG5 [169].

Tilapia, a sister group to the haplochromines and Lake Tanganyikan cichlids, segregate several additional sex chromosome systems, including XY systems on LG1 and LG23 and a ZW system on LG3. The causative sex-determination gene on LG23 is a duplication of *anti-Müllerian hormone*, which is known to play an important role in the vertebrate sex-determination network [45,46]. In both the haplochromines and tilapia, multiple sex-determination loci sometimes segregate within a given genus or even within a given species [53,54,59]. For example, some populations of the blue tilapia, *Oreochromis aureus*, simultaneously segregate both the LG1 XY system and LG3 ZW system [53].

In this study, we provide a more detailed characterization of sex chromosomes within the tilapia clade. We analyzed the sex chromosome systems of three species: *Oreochromis mossambicus*, *Coptodon zillii* and *Pelmatolapia mariae*. Previous research suggested that *O. mossambicus* has either a LG1 XY system or a LG3 ZW system, while *C. zillii* only has a LG1 XY system and *P. mariae* only has a LG3 ZW system [59]. However, this previous study used a small number of microsatellite markers to characterize these systems and thus could not characterize sequence divergence between the sex chromosomes in each species.

Materials and Methods

Materials

The *O. mossambicus* and *P. mariae* samples were lab-reared from fish purchased through the aquarium trade and the *O. mossambicus* stock likely came originally from South Africa. The *C. zillii* samples were derived from individuals originally collected in Lake Manzala, Egypt. Finclips were collected from each

individual and their gonads were visually inspected to determine their sex. For *O. mossambicus* we sampled two families (Family 1: 8M, 12 F; Family 2: 15M, 15F). We also sampled two families of *P. mariae* (Family 1: 25M, 21F; Family 2: 14M, 18F). The single family of *C. zillii* consisted of 9 males and 13 females.

Sequencing

DNA was purified from fin clips by phenol-chloroform extraction using phase-lock gel tubes (5Prime, Gaithersburg, Maryland). DNA concentrations for each sample were quantified using fluorescence spectroscopy and DNA pools for each sex were constructed with equal representation of each individual. Libraries were constructed using the TruSeq DNA PCR-Free LT Kit (Illumina, San Diego, California). For *O. mossambicus* and *P. mariae*, we pooled the two families of each species separately. Males and females from *C. zillii* and *O. mossambicus* each shared a lane of 100bp paired-end Illumina sequencing. Males and females from *P. mariae* each received a lane of 100bp paired-end Illumina sequencing.

Read mapping

Raw reads were de-multiplexed and filtered using the CASAVA 1.8 filter and low-quality reads that did not meet this criterion were discarded. The read qualities for each pool were visually inspected using FastQC version 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to the *Oreochromis niloticus* (O_niloticus_UMD_1) reference genome using BWA version 0.7.12 with default parameters and utilizing the read groups option [69,139]. The alignments were sorted, marked for duplicates, indexed using Picard version 1.119 (http://broadinstitute.github.io/picard/). Samtools was used to filter-out reads with an

alignment quality of zero [140]. Then, the alignments from families of the same species were merged and indexed using Picard version 1.119. Coverage estimates were obtained across the genome and for each linkage group using the *INTERVALS* option within the *CollectWgsMetrics* program for both sexes of each species with Picard version 2.17.4 (http://broadinstitute.github.io/picard/).

Variant calling

Variants were called with GATK version 3.7 software package [82]. We used the *HaplotypeCaller* command of GATK for each sex within each species and set the sample ploidy *(--sample_ploidy*) to its maximum of 10 due to the pooled nature of our datasets. We followed this up with GATK's *GenotypeGVCF* to combine the sexes for each of our species into a single output VCF file.

We also converted our alignment files into mpileup files using Samtools version 0.1.18 and subsequently into sync files with a PHRED score cutoff of 20 using Popoolation2 [83,140]. Next, each species was run through Sex_SNP_finder_GA.pl (https://github.com/Gammerdinger/sex-SNP-finder) in order to identify XY and ZW sex-patterned SNPs, and to calculate F_{ST} using a minimum coverage threshold of 10 and a maximum coverage threshold of 50. We also quantified the density of these sex-patterned SNPs in non-overlapping 10kb windows and looked for an elongation of the high SNP density tail in the distributions of both the XY and ZW patterned SNPs. If one tail was more elongated than the other, we ran a Mann-Whitney U-test to compare the distributions using a significance threshold of 0.05. If there was a significant difference, then we fit a negative binomial distribution to the distribution without the elongated tail, using the *fitdistr* function from the MASS library in R

[170,171], to create a null expectation for the distribution of non-overlapping, sex-patterned 10kb windows in a given species. Next, we calculated the top 0.1% of this fitted negative binomial to define regions of high differentiation using the *qnbinom* function in R [171]. We filtered the VCF file of each species by the list of sex-patterned SNPs generated by Sex_SNP_finder_GA.pl and used Bedtools version 2.26.0 *intersect* command to extract the sex-patterned SNPs within regions of high differentiation [109,115,126].

Functional annotation

The functional significance of each sex-patterned SNP was evaluated using SnpEff version 4.3 [85]. Sex-patterned SNPs classified as "high-impact" or missense mutations were extracted using SnpSift [86]. Missense mutations were subsequently analyzed with PROVEAN version 1.1.5 to predict the functional impact of each missense mutation on their respective protein [114]. We used the recommended PROVEAN score threshold of -2.5. Substitutions lower than this threshold were considered to be deleterious.

Results

Read mapping

For *P. mariae*, we obtained an 80.25% alignment rate for 22.96X coverage in the merged male pool and an 82.47% alignment rate for 19.41X coverage in the merged female pool. The sequencing from *C. zillii* returned an alignment rate of 77.59% for a coverage of 8.84X in males and an alignment rate of 78.74% for a coverage of 17.52X in females. From *O. mossambicus*, we retrieved a 79.28%

alignment rate for 12.89X coverage in the merged male pool and a 79.35% alignment rate for 14.91X coverage in the merged female pool.

Differentiation in *Oreochromis mossambicus*

In order to differentiate if the system is an XY or ZW system and also define the regions of high differentiation, we needed to compare the XY distribution to the ZW distribution of the non-overlapping, 10kb windows with sex-patterned SNPs. The tail of the XY distribution of non-overlapping, 10kb windows with sex-patterned SNPs is much larger than the tail of the ZW distribution, and the XY distribution is shifted significantly higher than the ZW distribution (Appendix A; $p < 0.05$). The top 0.1% of the negative binomial distribution fit to the ZW distribution corresponded to 13 or more sex-patterned SNPs per non-overlapping, 10kb window. The corresponding regions of high XY differentiation cover ~5.1Mb of the genome. Approximately 2.6Mb are found on LG14 and ~890kb on LG3b (Figure 17). No other linkage group or unanchored contig contained more than 200kb of highly differentiated regions. These regions of high differentiation overlap with the regions that show a strong signal of F_{ST} differentiation The region of differentiation on LG14 is tightly clustered almost entirely within the first 10Mb (Figure 18). We identified 69 genomic missense mutations responsible for 112 transcript missense mutations within the regions of high differentiation. PROVEAN predicted 27 of these 112 transcript missense mutations to be deleterious. Also, SnpEff predicted three mutations to be "high-impact" mutations. A complete list of all predicted deleterious missense and "high-impact" SNPs found in the regions of high differentiation in *O. mossambicus* can be found in Appendix B.

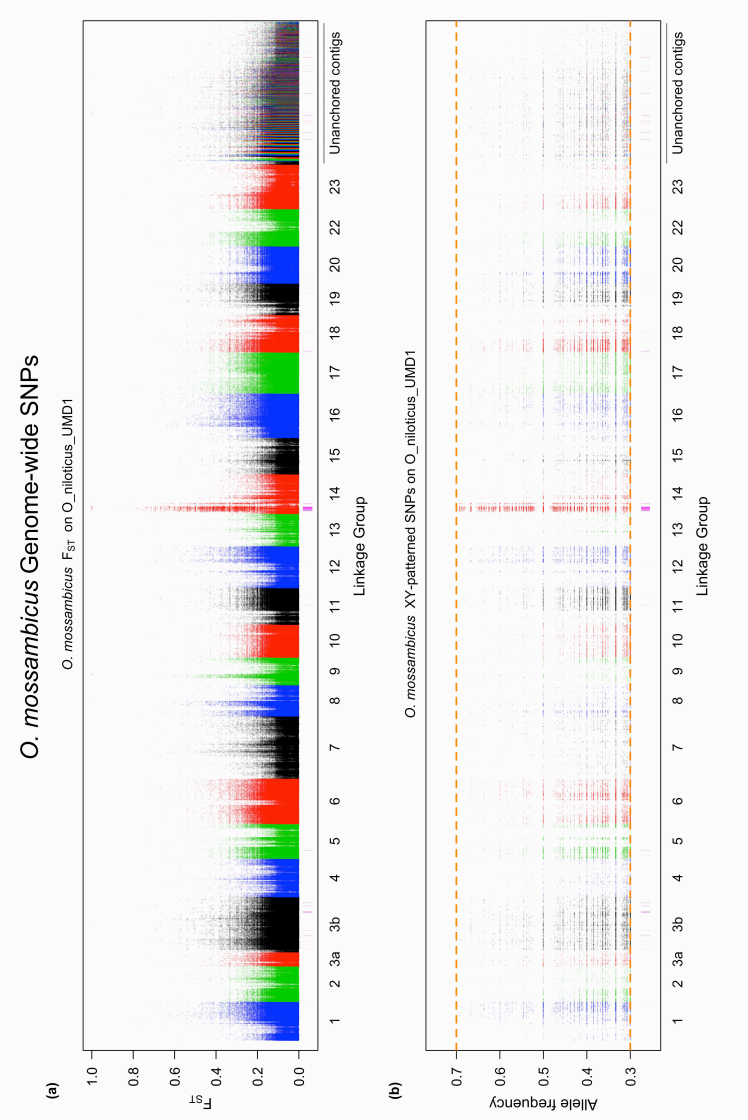


Figure 17. *Whole genome scan of O. mossambicus for a) F_{ST} and b) the allele frequency of the Y SNP in the male pool. The magenta rectangles underneath each figure represent regions of high differentiation.*

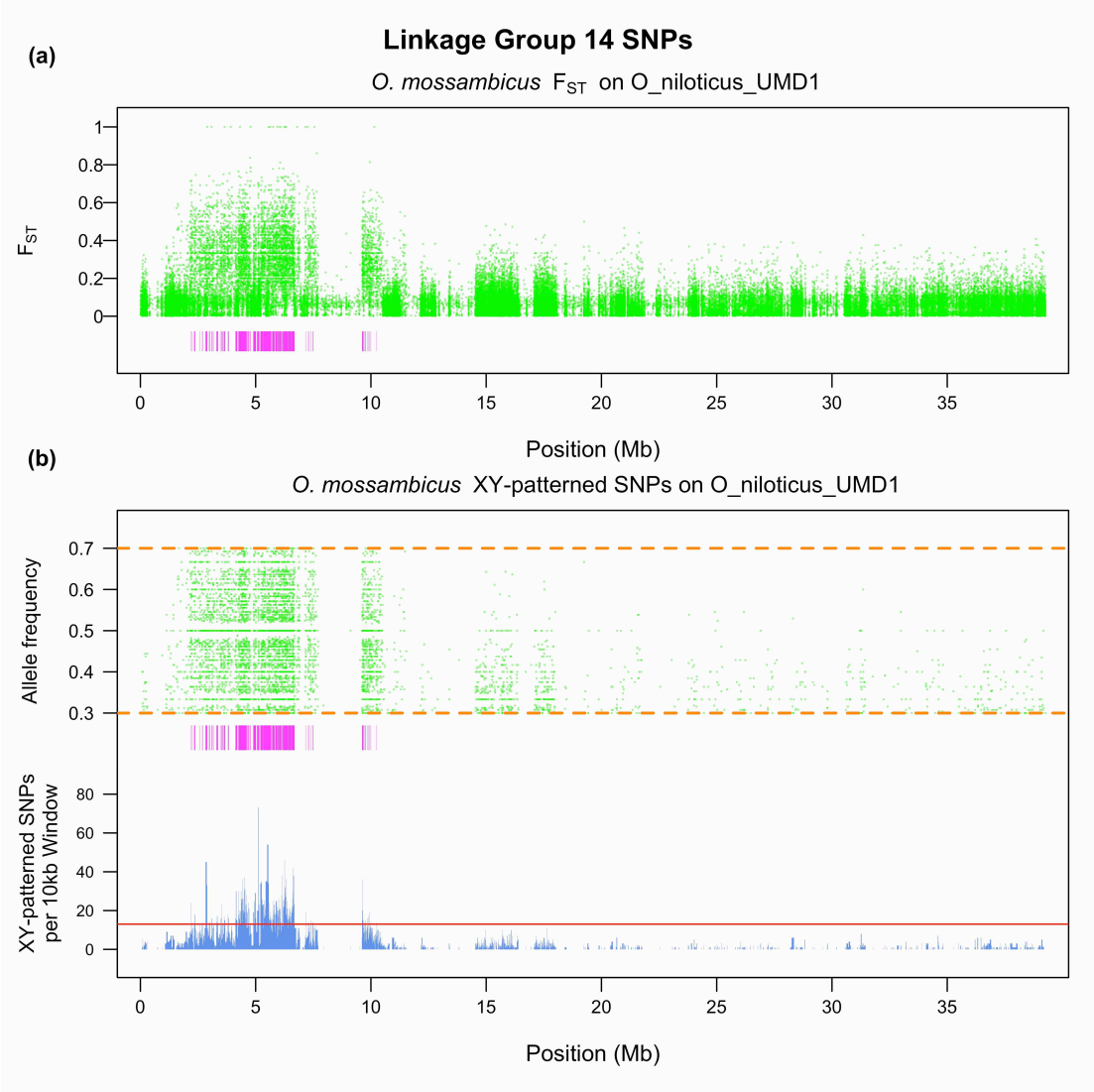


Figure 18. *LG14 scan of O. mossambicus for a) F_{ST} and b) the allele frequency of the Y SNP in the male pool. The magenta rectangles underneath each figure represent regions of high differentiation. The blue bars underneath panel (b) represent the number of XY-patterned SNPs in each non-overlapping, 10kb window and the red line represents the threshold for assigning a region as a region of high differentiation.*

Differentiation in *Coptodon zillii*

The tail of the distribution of non-overlapping, 10kb windows containing XY sex-patterned SNPs is much larger than the tail of the ZW distribution in *C. zillii* ($p < 0.05$) (Appendix C). The threshold derived from the top 0.1% of the negative binomial distribution fit to the ZW dataset was 14 or more sex-patterned SNPs per non-overlapping, 10kb window. Using this threshold, ~13.8 Mb of the genome falls within a region of high XY differentiation. Approximately 7.9Mb of this corresponds to LG1, with an additional ~1.1Mb on LG22 and ~728kb on LG3b (Figure 19). While there were many other small regions of differentiation scattered across the genome, no other linkage group or unanchored contig contained more than 350kb of highly differentiated regions. These regions of high differentiation explain the elevated F_{ST} on LG1 relative to the rest of the genome, and this differentiation corresponds to one end of the anchored linkage group (Figure 20). There are 266 genomic sex-patterned mutations creating 499 transcript missense mutations within the regions of high differentiation. PROVEAN predicts that 92 of the 499 missense mutations are deleterious. Lastly, a single mutation caused a "high-impact" mutation within this region of high differentiation. This mutation, along with a complete list of the predicted deleterious missense mutations in the regions of high differentiation in *C. zillii,* can be found in Appendix B.

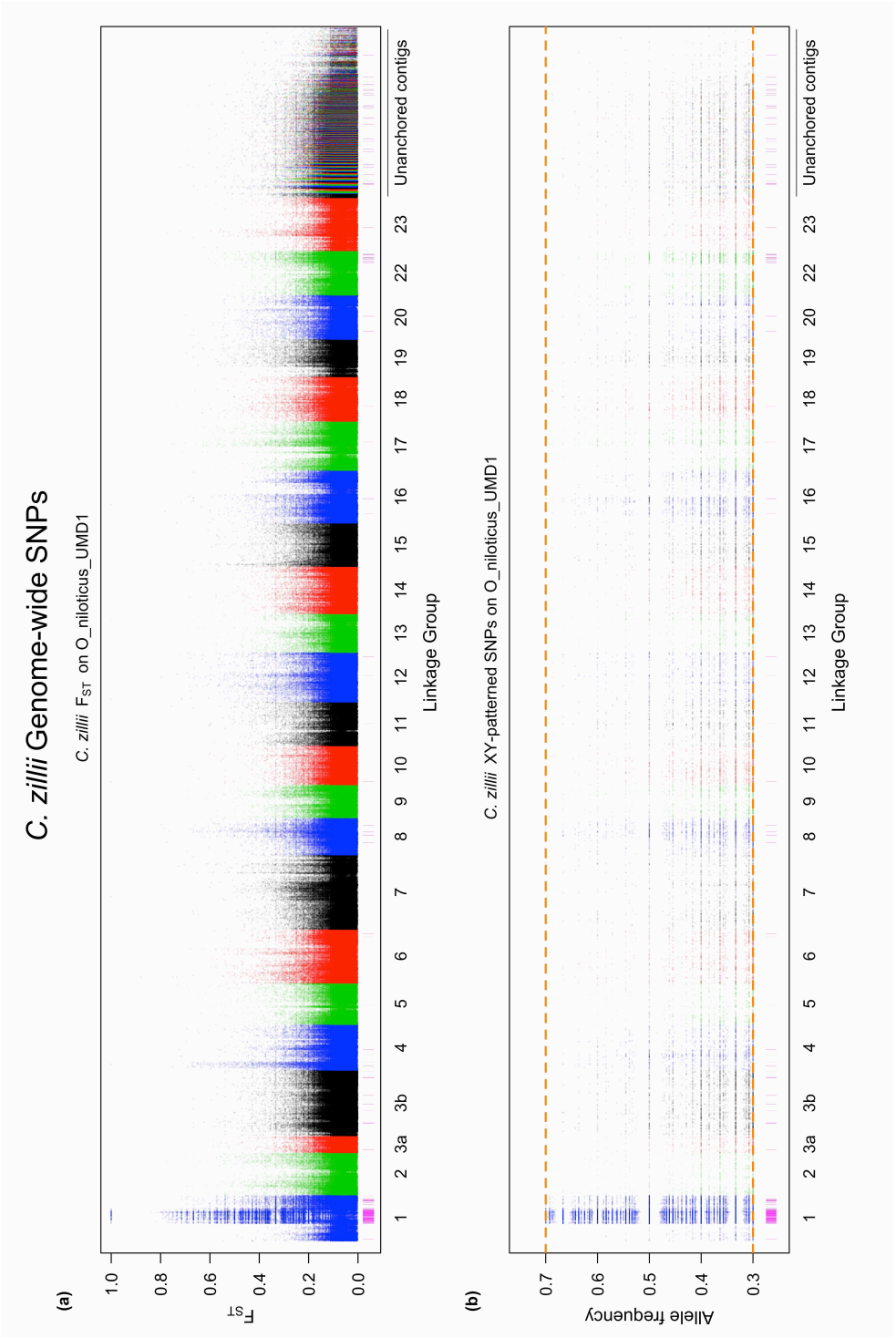


Figure 19. *Whole genome scan of C. zillii for a) F_{ST} and b) the allele frequency of the Y SNP in the male pool. The magenta rectangles underneath each figure represent regions of high differentiation.*

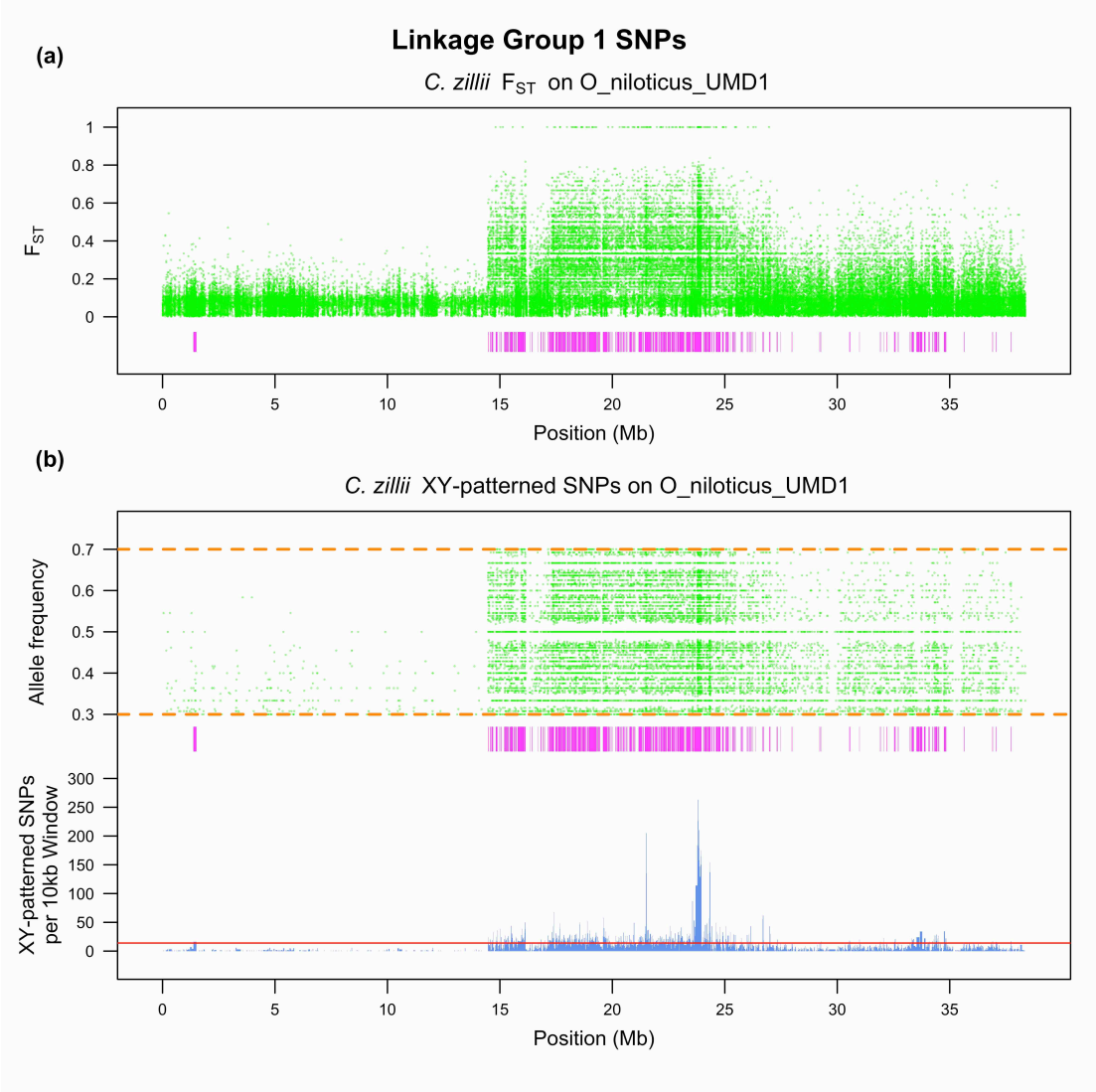


Figure 20. *LG1 scan of C. zillii for a) F_{ST} and b) the allele frequency of the Y SNP in the male pool. The magenta rectangles underneath each figure represent regions of high differentiation. The blue bars underneath panel (b) represent the number of XY-patterned SNPs in each non-overlapping, 10kb window and the red line represents the threshold for assigning a region as a region of high differentiation.*

Differentiation in *Pelmatolapia mariae*

Analysis of *P. mariae* revealed a high level of both XY and ZW signal. We detected 82,937 sex-patterned SNPs in the XY direction and 80,945 in the ZW

direction. Neither tail appeared more elongated than the other (Appendix D). Figure 21 and 22 show that LG3b has a high level of F_{ST} differentiation. There are 18,974 sex-patterned XY SNPs and 20,718 sex-patterned ZW SNPs on LG3b. We did not attempt to identify regions of high differentiation in this species for two reasons: (1) neither tail appeared to be elongated relative to the other and (2) there was a large excess of sex-patterned XY SNPs relative to what would have been expected given the results of a previous microsatellite study.

Additionally, our analysis of coverage from each linkage group revealed that LG3a and LG3b consistently contained the lowest mean coverage, and the highest standard deviation in coverage, in both sexes for all of the species analyzed (Appendix E). An example of this alignment problem is visualized in Appendix F, which compares a region on LG3b to a similarly sized region on LG6 and illustrates how reads align much better to LG6 than LG3b.

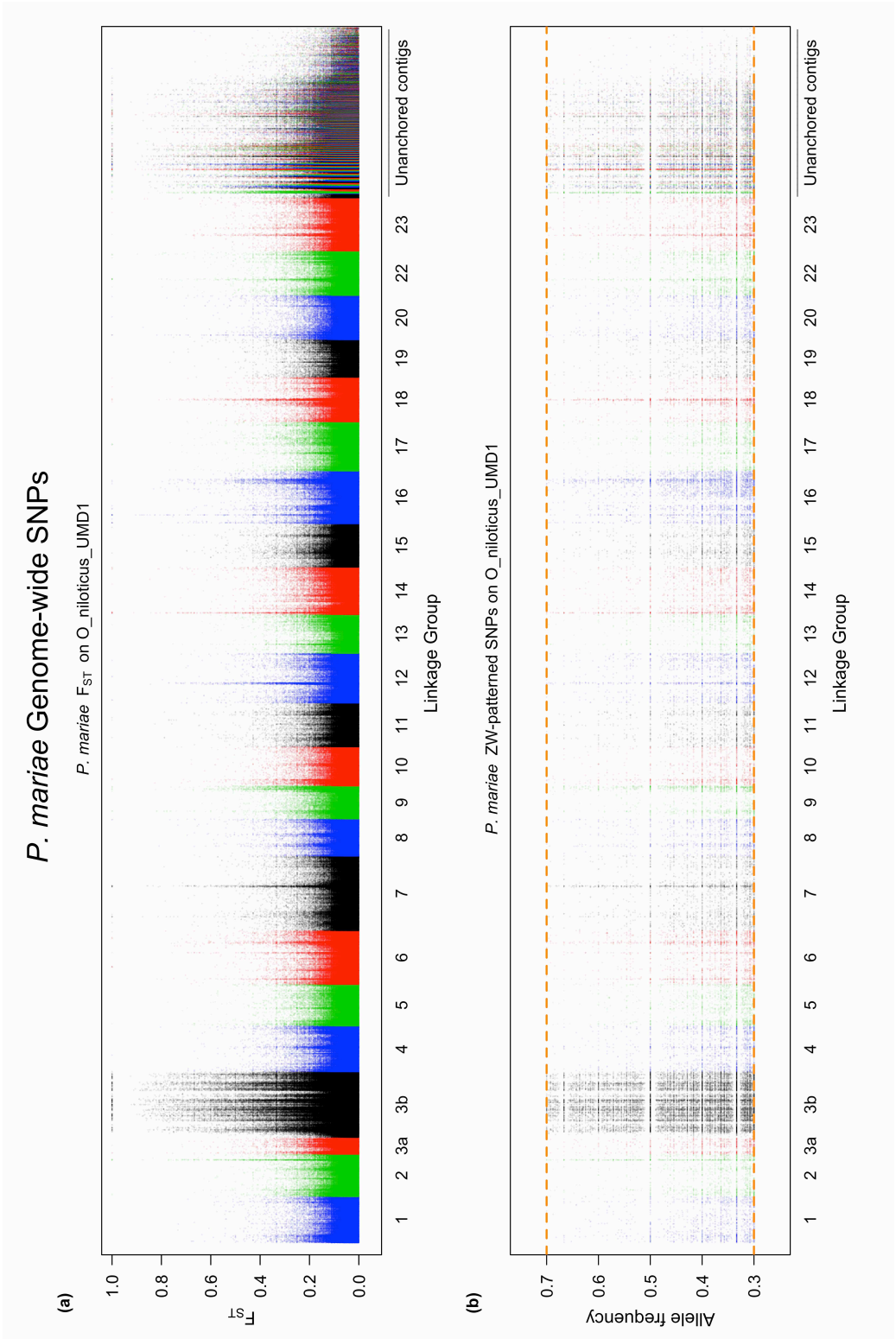


Figure 21. *Whole genome scan of P. mariae for a) F_{ST} and b) the allele frequency of the W SNP in the female pool.*

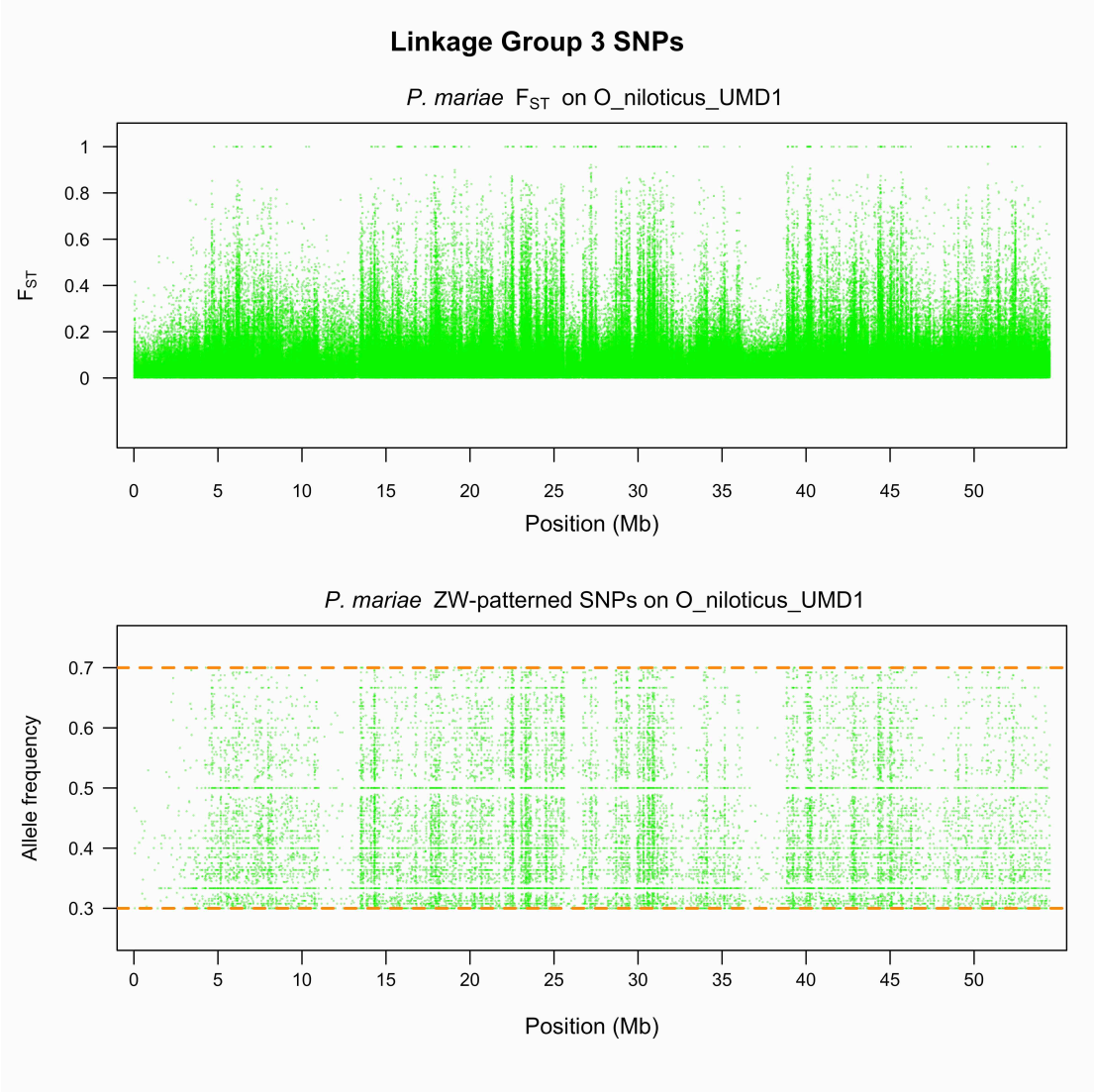


Figure 22. *LG3b scan of P. mariae for a) F_{ST} and b) the allele frequency of the W SNP in the female pool.*

Discussion

In this study, we used whole genome sequencing of pools of males and females to identify regions harboring sex-determination genes in species of tilapia. Our results provide additional support to the narrative that sex chromosomes

transitions occur at a rapid rate in cichlid fishes. However, they also illustrate the limits of our methodology and we propose solutions for moving forward.

LG14 in *Oreochromis mossambicus*

The high density of XY sex-patterned SNPs on LG14 likely indicates that this region of the genome harbors a novel sex-determination locus. Additionally, the lack of a bimodal distribution of the non-overlapping sex-patterned 10kb windows, as was observed in a recent study of Lake Tanganyikan cichlids, may provide evidence that this novel sex chromosome system is younger than its Lake Tanganyikan counterparts [169]. This is further corroborated by its small size, which suggests it has not had sufficient time to accumulate structural rearrangements that expand the regions of high differentiation, a characteristic of older sex chromosome systems. Furthermore, there are relatively few genes that have accumulated predicted deleterious missense and "high-impact" mutations as would be expected in an established sex chromosome. Taken together, this evidence suggests this is a young sex chromosome.

Since LG14 has not been previously described as a sex chromosome, it could be that this system is epistatically recessive to the previously reported LG1 XY and LG3 ZW systems. Studies of *O. aureus* demonstrated a LG3 W that was epistastically dominant to a LG1 Y, and individuals who were ZZ at LG3 and XX at LG1 were a mix between males and females, suggesting at least one additional factor for controlling sex in tilapia [53]. It could be that LG14 represents this additional factor, and which may have been revealed in our strain if the LG3 W and LG1 Y were lost after several generations of lab-rearing. Alternatively, there may be geographic variation in sex-determination for this species.

A candidate gene for sex determination within this region on LG14 is *AHNAK*, which has a "high-impact" SNP altering a splice acceptor site on the X-chromosome. *AHNAK* is involved in is a multitude of cellular processes including the regulation of calcium channels, cell signaling and membrane repair. But it also interacts with the *TGF-β* pathway as a tumor suppressor gene [172,173]. The *TGF-β* pathway has been previously implicated to play a critical role in sex-determination by modulating the germ cell count in the bipotential gonad [48]. It has been observed that bipotential gonads with fewer germ cells develop into testes, while bipotential gonads with more germ cells develop into ovaries [48]. Because *AHNAK* acts as a tumor suppressor for the *TGF-β* pathway, a SNP that disrupts this function on the X-chromosome could promote more germ cell growth in XX versus XY individuals. If this is true, the XX individuals with two mutant copies of *AHNAK* would develop into females and XY individuals with only one mutant copy of *AHNAK* would develop into males. However, the pleiotropic effects of such a mutation might impose large fitness costs on individuals.

LG1 in *Coptodon zillii*

The high density of XY sex-patterned SNPs on LG1 is consistent with a previous report of an XY system on LG1 in *C. zilli* [59]. Similar to *O. mossambicus*, we did not detect a bimodal distribution of 10kb windows with sex-patterned SNPs, suggesting this is also a young sex chromosome. However, as in *O. mossambicus*, we did see an extended tail in the XY distribution, suggesting that the LG1 system could be younger than the systems of the Lake Tanganyikan cichlids. However, phylogenetic evidence would suggest that the divergence between *C. zillii* and the

composite group of *O. niloticus* and *S. melanotheron*, which share a common LG1 XY system, predates the emergence of the Lake Tanganyikan cichlids [130]. Therefore, this pattern could be more plausibly explained by a mechanism through which the LG1 system is accumulating new mutations on the sex chromosomes at a slower rate than the sex chromosomes in Lake Tanganyikan cichlids. It could be that the mechanisms that have reduced recombination in the tilapia LG1 system have emerged more recently than the mechanisms that reduced recombination in the Lake Tanganyikan cichlid sex chromosomes. As a result, the Lake Tanganyikan sex chromosomes systems would have accumulated sex-patterned SNPs at a much more rapid rate and thus appear older than the LG1 XY system in tilapia. Alternatively, this LG1 system could represent convergence on a LG1 XY system.

The *C. zillii* system overlaps with the region previously shown to be associated with sex in *O. niloticus* and *S. melanotheron* [69,126]. The candidate genes for controlling sex from this region are the same as we discussed in previous work: *Ras-related protein R-Ras2*, *Suppression of tumorigenicity 5 protein*, *Ras association domain-containing protein 10*, *AFG3-like protein 1*, *Wilms tumor protein homolog*, *Estrogen-related receptor gamma* and *Growth regulation by estrogen in breast cancer 1* [109,126].

Interestingly, there is a less divergent block of differentiation centered around ~30Mb in *O. niloticus* that doesn't appear to be prominent in *C. zillii* [69]. Similarly, *C. zillii* seems to have a less divergent block of differentiation around ~34Mb suggesting that these regions represent different evolutionary strata in each lineage. Additionally, *S. melanotheron* shows widespread differentiation across this region. Theory suggests

that a region harboring sexually antagonistic alleles can remain in linkage disequilibrium with a sex-determination locus, and that the intervening region can show less differentiation than the differentiation at either locus [119]. However, given the distance between the sex-determination region and the sexually antagonistic locus in *C. zillii*, there would need to be very high levels of sexual antagonism to keep these regions in linkage disequilibrium. Therefore, it is more likely that the region around ~34Mb represents a structural rearrangement that has engulfed a new region of LG1 and created a less differentiated stratum in *C. zillii*.

LG3 in *Pelmatolapia mariae*

Previous research suggested that *P. mariae* has a ZW sex chromosome system on LG3. Our F_{ST} data (Figure 21) also clearly indicate the presence of a sex chromosome system on LG3. However, our data show a large number of XY patterned SNPs which make it difficult to confidently call regions of high differentiation. This 'noise' is likely to be the result of several factors. First, LG3b harbors many repetitive blocks that make alignment of short reads difficult. While our read alignments were relatively even on other linkage groups, we had relatively poor alignments in both sexes on LG3b with reads piling up in some locations and absent in others (Supplementary files 5 and 6). It could be that because the repetitive reads originating from non-homologous regions are piling up in this alignment, it creates an artifact that resembles an XY signal. Second, it could be that the Z-chromosome in *P. mariae* harbors a high level of polymorphism and thus creates spurious XY signal. Third, and likely most importantly, it could be that our LG3 sequence is simply too diverged from the reference genome to effectively align short reads. We were able to

align these short reads to the rest of the genome as seen in Supplementary files 5 and 6, but the reference sequence for this linkage group appears to be particularly problematic. It seems that our approach for finding sex-patterned SNPs has limitations that prevent us from determining whether it represents an XY or ZW signal; however, we are confident that our samples have a sex determination system on LG3b because we are getting a high level of F_{ST} signal on LG3b. The approach we have developed in this and previous studies seems effective at defining sex chromosomes in their earliest stages, but breaks down when sequences diverge strongly from the reference sequence. Longer sequencing read technologies and a better reference sequence will likely facilitate future studies of the LG3 ZW system on this sex chromosome pair.

Conclusions

We report and quantify the decay of a novel XY sex chromosome system on LG14 in *O. mossambicus*. The discovery of this new sex chromosome system illustrates the incredible diversity in cichlid sex chromosomes and argues for more studies to quantify their rich variety. Additionally, we confirm a previous study that reported a XY system on LG1 in *C. zillii* and we quantified the level of decay on this chromosome. We were unable to distinguish the presumed ZW system from an XY system in *P. mariae,* likely due to both technical and biological issues. We suggest refinements to our approach, such as longer reads and a more closely related reference assembly of LG3, which should allow for the characterization of this sex chromosome in future studies.

Funding

This work was supported by the National Science Foundation under Grant Number DEB-1143920; U.S. Department of Education Graduate Assistance in Areas of National Need fellowship offered through the University of Maryland Biology Department under Award Number P200A150160; and the American Genetics Association through an Ecological, Evolutionary and Conservation Genomics Award.

Acknowledgements

The authors acknowledge the University of Maryland supercomputing resources (http://www.it.umd.edu/hpcc) made available for conducting the research reported in this paper. Samples were processed in accordance with University of Maryland IACUC protocol #R-10-73.

Data Availability

All reads have been deposited in the NCBI SRA under Bioproject accession number PRJNA432420.

Authors' contributions

WJG helped conceive the study, extracted and quantified the DNA, constructed the libraries for sequencing, carried out the bioinformatic approaches and wrote the manuscript. MAC aided in the bioinformatic approaches. BAS helped quantify the DNA. DJP collected the finclips for *C. zillii*. TDK helped conceive the study, aided in library construction and revised the manuscript.

Chapter 6: Synthesis

Historically, sex chromosome research has focused on old, heteromorphic and highly heterochromatic sex chromosomes because they were easy to identify in karyotypes. These sex chromosomes formed much of the foundation for the evolutionary theory on the evolution of sex chromosomes[14,15,167]. However, with the emergence of next generation sequencing, younger, homomorphic sex chromosomes are also beginning to be studied and the theoretical framework regarding the earliest stages of sex chromosome evolution are being tested. The work presented in this thesis has focused on these early stages by characterizing sex chromosome systems and quantifying the decay of genes that reside within them[109,126,169].

Size of the Sex-Determination Network

Some have argued that the number of genes capable of being the master sex-determination locus is limited because the sex chromosome systems characterized to date involve a small set of genes[162]. However, this conclusion is likely a reflection of confirmation bias, because researchers search the regions of sex differentiation for candidate genes already known to be critically involved in the sex-determination network. When one of these candidate genes is found, follow-up studies that analyze the functional effects of these candidate genes usually confirm their role in the sex-determination network. However, our knowledge of the genes that could be involved in the sex-determination network is still limited and therefore constrains our search criteria when looking for genes in regions where sex is mapping to in a given species. Thus, this leaves researchers without candidate genes when their regions do not contain genes known to be involved in the sex-determination network.

A recent study analyzed the mammalian sex-determination network using a multi-step Boolean model and for simplicity of modeling limited the size of the network of genes to 21 genes[125]. While the sex-determination network is undoubtedly larger than what was modeled, this study begins to illustrate that the size of the network is more than the handful of genes that have previously been implicated. Furthermore, while analyzing the work presented here, the genes from the regions that control sex do not harbor many of the genes in this network and thus push forward the notion that the sex-determination network is likely even larger than the 21 genes used in the Boolean model and certainly larger than the handful of genes hypothesized by others.

Landscape of sexual antagonism

As the community continues to collect transcriptome data sets, it becomes increasingly clear that males and females differ in expression at many loci at different life stages across the genome. Additionally, the notion that sexually antagonistic traits are dispersed broadly across the genome has previously been synthesized[174].

Two kinds of sexually antagonistic loci exist in the genome: polymorphic loci that are kept in intermediate frequency by a combination of balancing selection and drift, and monomorphic loci that are constrained to an expression level which is a compromise between the optimal expression for males and females. Sex differences in gene expression are common and, depending on the tissue, organ and time point selected, can represent more than half of the genome[93,175,176]. While we currently do not know how many sexually antagonistic loci are polymorphic at any given time, it seems reasonable to believe that the landscape of sexually antagonistic selection

might depend on the degree of sexual differentiation, history of sexual antagonism, and the molecular mechanisms that have been selected to differentiate gene expression and reduce sexually antagonistic selection. A recent study evaluated the levels of sexual antagonism in hemi-clonal lines of *Drosophila melanogaster* and estimated that 8% of the genes in the genome met their criteria for sexual antagonism[177]. If cichlids have a similar proportion of sexually antagonistic genes, we could expect that, on average, there are several genes experiencing sexual antagonistic selection in every megabase across the cichlid genome.

Interactions between the sex-determination network and the landscape of sexual antagonism

With the presence of a landscape of sexually antagonism across the genome and the large scale of the sex-determination network in mind, we can now ask how these two would interact with each other. Figure 23a depicts a hypothetical genome superimposing the polymorphic, sexually antagonistic landscape onto the sex-determination network. From this we can see that some regions, such as the right end of chromosome one and the middle of chromosome two, have islands of polymorphic, sexually antagonistic alleles. Furthermore, some of these regions also have nodes in the sex-determination network nearby.

From this assessment, we can make predictions about the evolution of sex-determination and sex chromosomes within these regions. Each one of the nodes in the sex-determination network has a probability of becoming the novel master sex-determination gene, but those with stronger and closer polymorphic sexual antagonistic selection in the landscape surrounding them will become fixed more

rapidly then the others. Each chromosome is part of a probability mass function for being the next chromosome to harbor the master sex-determination locus. The probability of a particular locus becoming fixed as the master sex-determination system is a function of the combined strength of the polymorphic, sexually antagonistic alleles in the region, the recombination distance between those alleles and the sex-determination locus, the probability of a mutation creating a sex-determination locus in the node, and the probability of mutations occurring which convert monomorphic sexually antagonistic loci into polymorphic sexually antagonistic loci (Figure 23b).

One might argue that this model favors the notion that some sex-determination mechanisms will be independently recycled[162]. However, the key insight in this model is that many of the chromosomes have some probability of being co-opted as sex chromosomes, and that many of them will be used for this purpose at some point. Therefore, the shape of the probability mass function in Figure 23b for a given group of organisms determines how many chromosomes could potentially be sex chromosomes. Furthermore, this probability mass function is dynamic over temporal and spatial scales as selection pressures change. This could limit or expand where sex chromosome transitions occur.

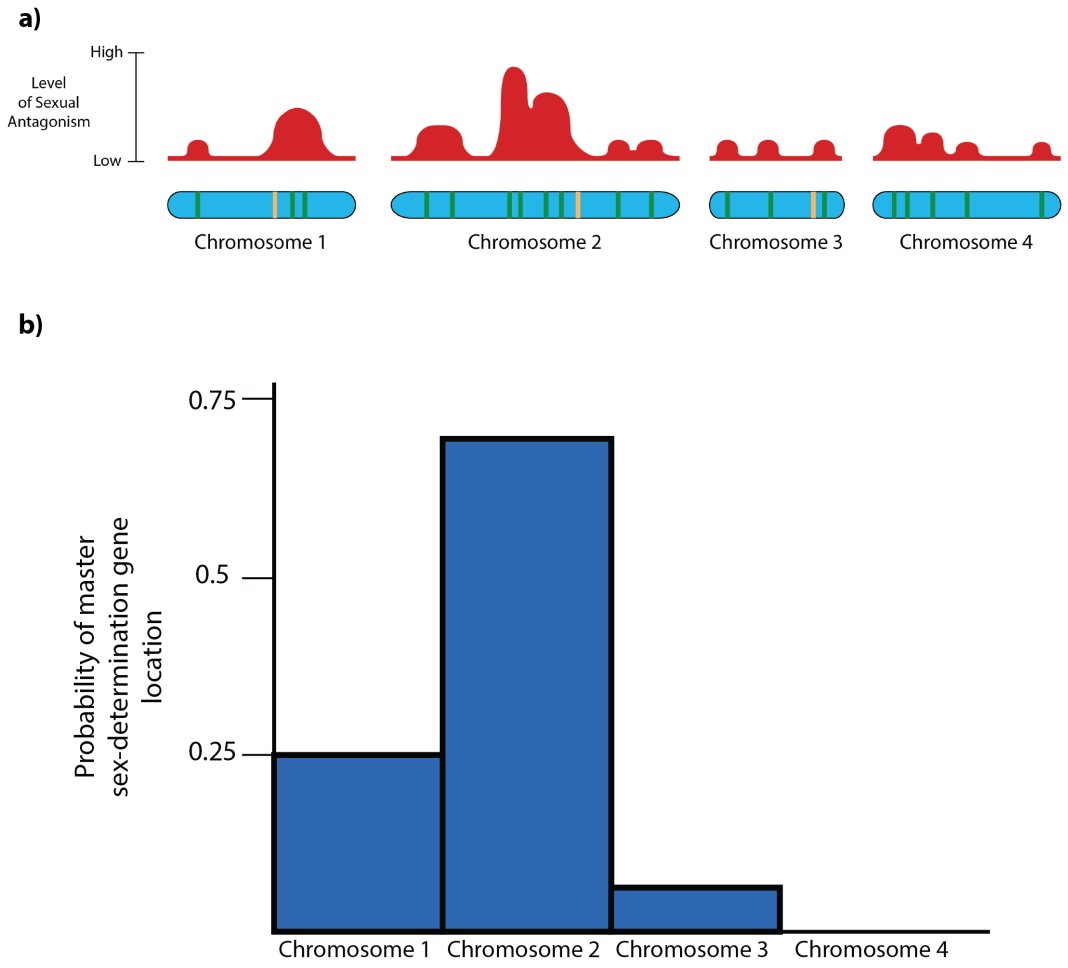


Figure 23. *a) Hypothetical genome with green bars on the chromosomes representing sexually antagonistic loci and tan bars representing nodes in the sex-determination network. The density of the sexual antagonistic alleles determines the level of sexual antagonism in the genome. b) Probability mass function reflecting the probability of each chromosome being the next sex chromosome in the genome.*

Why do cichlids have so many sex chromosomes?

The diversity of sex chromosomes within African cichlids is astounding (Figure 24), but cichlids also seem to possess a wide-variety of traits that seem likely to drive sexual antagonism, for example, parental care behaviors, size dimorphism and pigmentation. Interestingly, the patterns of diversity of sex-determination mechanisms appeared to be mirrored in another adaptive radiation, in lizards[168,178]. Lizards are have large sexual size dimorphisms in their heads and abdomens which are hypothesized to be linked with traits that are presumably related to sexual antagonism and likely to be polygenic. These traits include larger abdomens in females for producing larger clutch sizes while males have larger heads to aid in male-to-male competition combats and grasping of females[179]. As a result, I propose that the diversity of cichlid, and possibly lizard as well, sex chromosomes is a reflection of a rather uniform probability mass function for sex chromosome turnover.

When a clade of organisms possesses this rather uniform probability mass function for sex chromosome turnover, then it allows new sex chromosomes in many different parts of the genome to emerge. As new sex–determination mechanisms emerge they are also fundamentally altering the wiring of the sex-determination network and this rewiring of the sex-determination network directly impacts the molecular interactions between two rival sex-determination systems. Thus, when these sex-determination systems interact with each other, their sex-determination networks are incompatible with each other and it might create hybrids with low fitness values due to poor gonad differentiation, infertility and populations with large sex-ratio distortions. There might be large post-zygotic barriers to gene flow between

organisms bearing different sex-determination systems and this process could promote speciation.

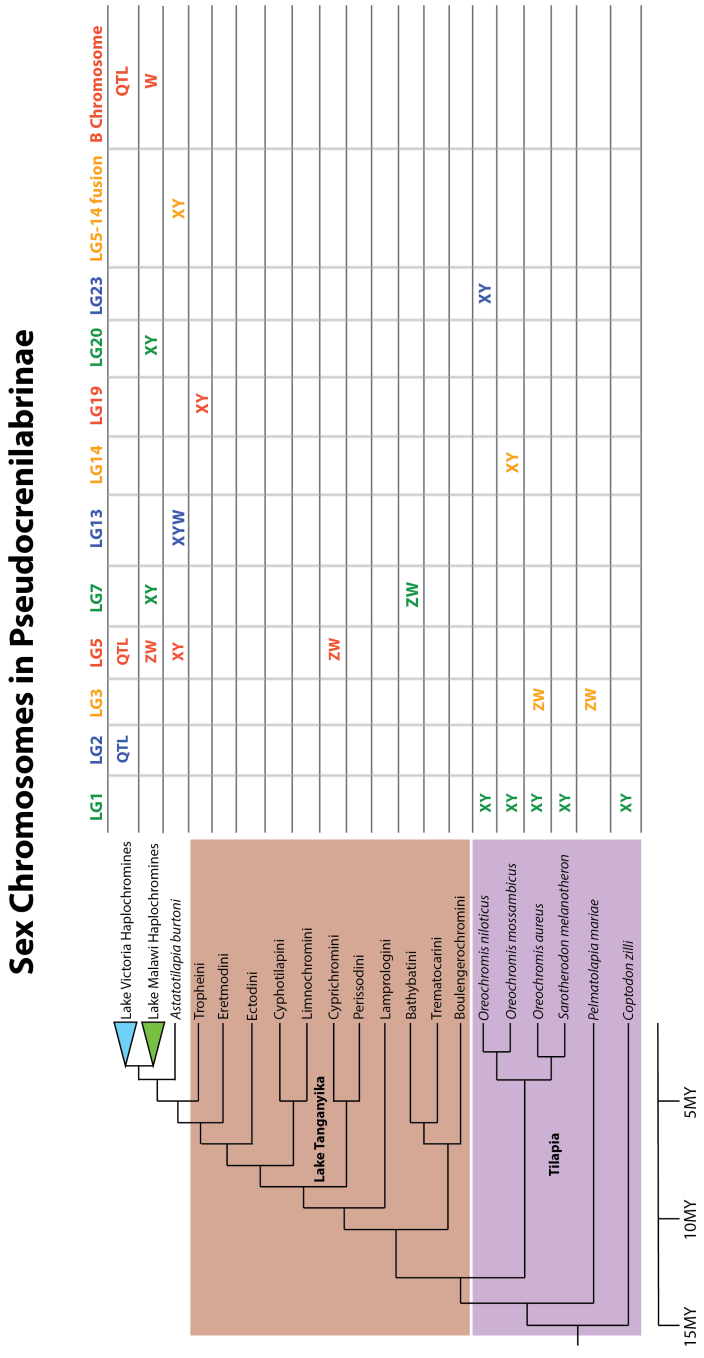


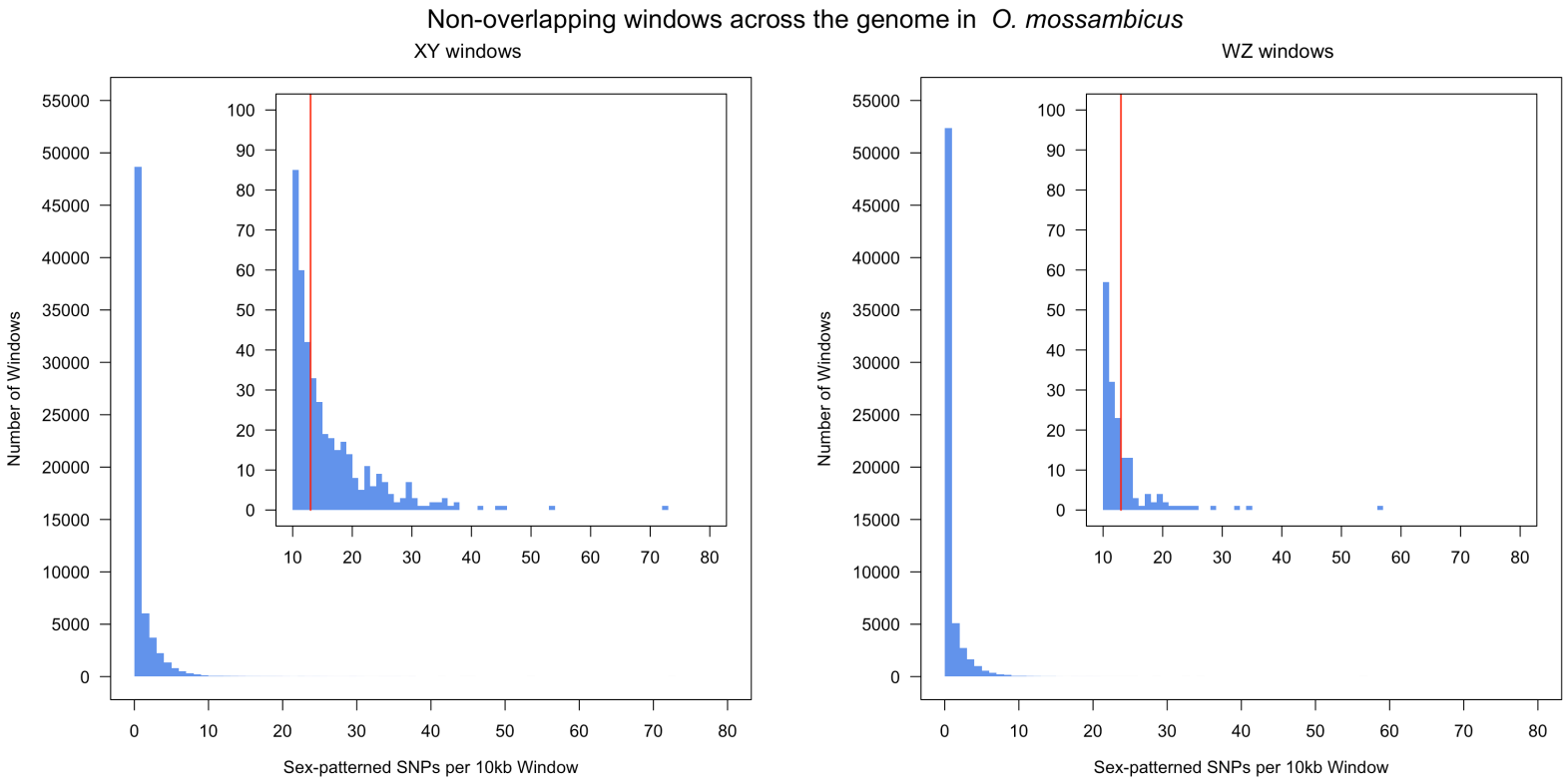
Figure 24. *Overview of sex chromosome systems within African cichlids including the work present in this dissertation.*

While it seems unlikely that this hypothesis explains all of the variation observed in the adaptive radiation of cichlids, it could be at least one source for generating the rich variety of cichlid species. A variety of other mechanisms are almost certainly also aiding in the speciation of cichlids and includes, but is not limited to, traits such as habitat adaptation and jaw morphology[52]. However, the role of sex chromosomes in the speciation of cichlids is likely an additional key factor.

In conclusion, I hypothesize that the rich diversity of sex chromosomes within cichlids is the result of a relatively uniform distribution of sexually antagonistic loci across the genome and that there is usually at least one sexually antagonistic locus locus near one of the dozens of genes that could potentially be co-opted to become a novel sex-determination locus. As a result, new sex-determination systems can swiftly emerge and simultaneously contribute to the rapid and fantastic diversity of cichlids.

Appendices

Appendix A. Distribution of sex-patterned windows both the XY and ZW direction for *O. mossambicus*. The vertical red line denotes the threshold of 13 sex-patterned SNPs per non-overlapping 10kb window.



Appendix B. A list of all sex-patterned missense mutations, "high-impact" mutations and genes within the regions of high differentiation in both *O. mossambicus* and *C. zillii*.

SnpEff "High Impact" Sex Mutations in *O. mossambicus*						
contig/LG	**Position**	**Impact**	**EFFECT**	**mRNA Accession**	**Protein Accession**	**Gene**
contig201	791849	splice_acceptor_variant& intron_variant	HIGH	XM_019354168.1	XP_019209713.1	PREDICTED: uncharacterized protein LOC102082939 isoform X1 [Oreochromis niloticus]
lg14	6320084	splice_acceptor_variant& intron_variant	HIGH	XM_019367204.1	XP_019222749.1	PREDICTED: neuroblast differentiation-associated protein AHNAK [Oreochromis niloticus]
lg3b	39718419	splice_acceptor_variant& intron_variant	HIGH	XM_019357127.1	XP_019212672.1	PREDICTED: protein NLRC3-like [Oreochromis niloticus]

PROVEAN Deleterious Sex Missense Mutations in *O. mossambicus*					
contig/LG	**Position**	**PROVEAN SCORE**	**AA CHANGE**	**Protein Accession**	**Gene**
contig555	141677	-9.033	C180Y	XP_019211342.1	PREDICTED: uncharacterized protein C7orf43 homolog [Oreochromis niloticus]
lg14	4162412	-2.673	L211F	XP_019200944.1	PREDICTED: uncharacterized protein LOC106098495 isoform X3 [Oreochromis niloticus]

lg14	4162412	-2.673	L228F	XP_019200943.1	PREDICTED: uncharacterized protein LOC106098495 isoform X2 [Oreochromis niloticus]
lg14	4162412	-2.673	L247F	XP_019200942.1	PREDICTED: uncharacterized protein LOC106098495 isoform X1 [Oreochromis niloticus]
lg14	4281400	-3.134	R735Q	XP_005474915.1	PREDICTED: integrator complex subunit 4 [Oreochromis niloticus]
lg14	4283137	-3.994	T859M	XP_005474915.1	PREDICTED: integrator complex subunit 4 [Oreochromis niloticus]
lg14	5664637	-3.21	R1351W	XP_019222748.1	PREDICTED: remodeling and spacing factor 1 [Oreochromis niloticus]
lg14	5904697	-5.554	R426W	XP_019200630.1	PREDICTED: coiled-coil domain-containing protein 9 isoform X3 [Oreochromis niloticus]
lg14	5904697	-5.554	R427W	XP_019200628.1	PREDICTED: coiled-coil domain-containing protein 9 isoform X1 [Oreochromis niloticus]
lg14	5904697	-5.554	R427W	XP_019200629.1	PREDICTED: coiled-coil domain-containing protein 9 isoform X2 [Oreochromis niloticus]
lg14	5904697	-5.754	R412W	XP_019200631.1	PREDICTED: coiled-coil domain-containing protein 9 isoform X4 [Oreochromis niloticus]
lg14	6038733	-4.516	N368S	XP_003456442.2	PREDICTED: kinesin light chain 1 [Oreochromis niloticus]
lg14	6312099	-3.889	L1534F	XP_019223254.1	PREDICTED: neuroblast differentiation-associated protein AHNAK-like [Oreochromis niloticus]
lg14	6326287	-4.65	G1114R	XP_019222749.1	PREDICTED: neuroblast differentiation-associated protein AHNAK [Oreochromis niloticus]
lg14	6326481	-2.976	I1049N	XP_019222749.1	PREDICTED: neuroblast differentiation-associated protein AHNAK [Oreochromis niloticus]
lg14	9670595	-3.009	P648S	XP_003450231.2	PREDICTED: splicing factor, arginine/serine-rich 15 [Oreochromis niloticus]
lg18	11358328	-4.263	H116Q	XP_003438250.2	PREDICTED: uncharacterized protein LOC100699978 isoform X1 [Oreochromis niloticus]

lg18	11358328	-4.363	H115Q	XP_013126154.1	PREDICTED: uncharacterized protein LOC100699978 isoform X2 [Oreochromis niloticus]
lg20	5605572	-4.112	E128A	XP_019204940.1	PREDICTED: uncharacterized protein LOC109196119 isoform X1 [Oreochromis niloticus]
lg20	5605572	-4.152	E128A	XP_019204942.1	PREDICTED: uncharacterized protein LOC109196119 isoform X3 [Oreochromis niloticus]
lg20	5605572	-4.212	E128A	XP_019204941.1	PREDICTED: uncharacterized protein LOC109196119 isoform X2 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_005448488.2	PREDICTED: membrane cofactor protein isoform X1 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_005448489.1	PREDICTED: membrane cofactor protein isoform X1 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_005448490.1	PREDICTED: membrane cofactor protein isoform X2 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_005448492.1	PREDICTED: C4b-binding protein alpha chain isoform X5 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_013130661.1	PREDICTED: membrane cofactor protein isoform X3 [Oreochromis niloticus]
lg20	9361517	-4.122	L8Q	XP_013130666.1	PREDICTED: C4b-binding protein alpha chain isoform X4 [Oreochromis niloticus]

SnpEff "High Impact" Sex Mutations in *C. zillii*						
contig/LG	**Position**	**Impact**	**EFFECT**	**mRNA Accession**	**Protein Accession**	**Gene**
lg1	23928083	splice_donor_variant& intron_variant	HIGH	XM_005458931.3	XP_005458988.1	PREDICTED: heterogeneous nuclear ribonucleoprotein H3 [Oreochromis niloticus]

PROVEAN Deleterious Sex Missense Mutations in *C. zillii*					
contig/LG	**Position**	**PROVEAN SCORE**	**AA CHANGE**	**Protein Accession**	**Gene**
contig393	90584	-2.714	L136Q	XP_019210581.1	PREDICTED: uncharacterized protein LOC106097127 isoform X2 [Oreochromis niloticus]
contig393	90584	-2.714	L136Q	XP_019210582.1	PREDICTED: uncharacterized protein LOC106097127 isoform X3 [Oreochromis niloticus]
contig393	90991	-2.857	N40S	XP_019210581.1	PREDICTED: uncharacterized protein LOC106097127 isoform X2 [Oreochromis niloticus]
contig393	90991	-2.857	N40S	XP_019210582.1	PREDICTED: uncharacterized protein LOC106097127 isoform X3 [Oreochromis niloticus]
contig598	84187	-2.786	S91N	XP_019211502.1	PREDICTED: protein phosphatase 1 regulatory subunit 3G [Oreochromis niloticus]
lg1	15490488	-4.776	C21Y	XP_003445796.2	PREDICTED: haptoglobin [Oreochromis niloticus]
lg1	15567236	-2.689	N20Y	XP_005466978.1	PREDICTED: U3 small nucleolar ribonucleoprotein protein MPP10 [Oreochromis niloticus]
lg1	15641325	-2.734	R867G	XP_005466994.1	PREDICTED: kinesin-like protein KIF18A isoform X2 [Oreochromis niloticus]
lg1	15641325	-2.751	R868G	XP_005466993.1	PREDICTED: kinesin-like protein KIF18A isoform X1 [Oreochromis niloticus]

lg1	15705860	-2.976	T93A	XP_003445738.1	PREDICTED: probable methyltransferase-like protein 15 [Oreochromis niloticus]
lg1	15705860	-2.976	T93A	XP_005466987.1	PREDICTED: probable methyltransferase-like protein 15 [Oreochromis niloticus]
lg1	15705860	-2.976	T93A	XP_005466988.1	PREDICTED: probable methyltransferase-like protein 15 [Oreochromis niloticus]
lg1	15705860	-2.976	T93A	XP_005466989.1	PREDICTED: probable methyltransferase-like protein 15 [Oreochromis niloticus]
lg1	15705860	-2.976	T93A	XP_005466990.1	PREDICTED: probable methyltransferase-like protein 15 [Oreochromis niloticus]
lg1	15996731	-4.195	N281T	XP_005467000.1	PREDICTED: myosin-binding protein C, cardiac-type isoform X4 [Oreochromis niloticus]
lg1	15996731	-4.262	N362T	XP_005466998.1	PREDICTED: myosin-binding protein C, cardiac-type isoform X2 [Oreochromis niloticus]
lg1	15996731	-4.262	N362T	XP_013127145.1	PREDICTED: myosin-binding protein C, cardiac-type isoform X5 [Oreochromis niloticus]
lg1	15996731	-4.262	N371T	XP_005466997.1	PREDICTED: myosin-binding protein C, cardiac-type isoform X1 [Oreochromis niloticus]
lg1	15996731	-4.262	N371T	XP_005466999.1	PREDICTED: myosin-binding protein C, cardiac-type isoform X3 [Oreochromis niloticus]
lg1	18556187	-3.458	Y229F	XP_005467231.1	PREDICTED: protein RIC-3 [Oreochromis niloticus]
lg1	19085366	-3.749	T240A	XP_003442330.1	PREDICTED: lymphatic vessel endothelial hyaluronic acid receptor 1 [Oreochromis niloticus]
lg1	19609803	-3.826	A1848V	XP_005467084.1	PREDICTED: polycystic kidney disease protein 1-like 2 isoform X1 [Oreochromis niloticus]
lg1	19609803	-3.826	A1912V	XP_019214067.1	PREDICTED: polycystic kidney disease protein 1-like 2 isoform X2 [Oreochromis niloticus]
lg1	20768581	-2.998	R93C	XP_013125183.1	PREDICTED: protein GRINL1A [Oreochromis niloticus]

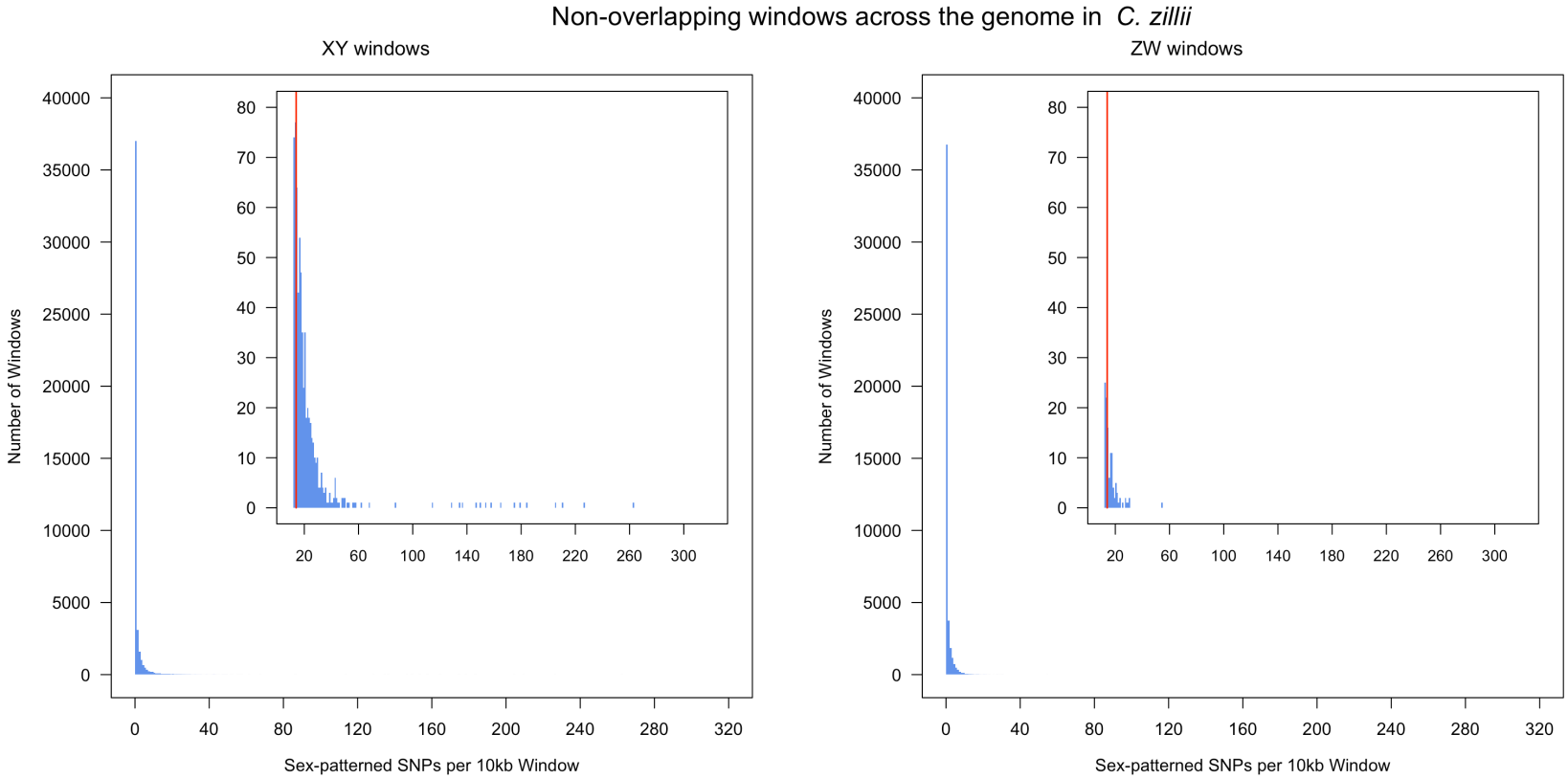
lg1	20982195	-2.507	D103N	XP_019214376.1	PREDICTED: DNA-binding protein RFX7 isoform X2 [Oreochromis niloticus]
lg1	20982195	-2.507	D199N	XP_005467130.1	PREDICTED: DNA-binding protein RFX7 isoform X1 [Oreochromis niloticus]
lg1	21179829	-5.728	L92P	XP_005467151.1	PREDICTED: prolyl 3-hydroxylase OGFOD1 [Oreochromis niloticus]
lg1	21258552	-2.574	H206R	XP_003442385.1	PREDICTED: putative sodium-coupled neutral amino acid transporter 8 [Oreochromis niloticus]
lg1	21368727	-2.587	H685Y	XP_005467237.2	PREDICTED: stereocilin [Oreochromis niloticus]
lg1	21656349	-2.92	D152H	XP_005467174.1	PREDICTED: transcription factor 25 isoform X2 [Oreochromis niloticus]
lg1	21656349	-2.937	D152H	XP_003442375.1	PREDICTED: transcription factor 25 isoform X1 [Oreochromis niloticus]
lg1	21703903	-3.699	A384T	XP_019214648.1	PREDICTED: AFG3-like protein 1 isoform X2 [Oreochromis niloticus]
lg1	21703903	-3.699	A385T	XP_005467180.1	PREDICTED: AFG3-like protein 1 isoform X1 [Oreochromis niloticus]
lg1	21703903	-3.699	A385T	XP_013125207.1	PREDICTED: AFG3-like protein 1 isoform X3 [Oreochromis niloticus]
lg1	22950448	-7.864	P117S	XP_003442359.1	PREDICTED: elongator complex protein 4 [Oreochromis niloticus]
lg1	23796110	-6.162	D657G	XP_003454556.1	PREDICTED: hexokinase-1 [Oreochromis niloticus]
lg1	23836869	-3.406	V98A	XP_005458982.1	PREDICTED: tetraspanin-15 [Oreochromis niloticus]
lg1	23914641	-2.608	S572C	XP_005458984.1	PREDICTED: DNA replication ATP-dependent helicase/nuclease DNA2 isoform X1 [Oreochromis niloticus]
lg1	23914641	-2.684	S572C	XP_019214957.1	PREDICTED: DNA replication ATP-dependent helicase/nuclease DNA2 isoform X2 [Oreochromis niloticus]
lg1	23914991	-3.521	T600I	XP_005458984.1	PREDICTED: DNA replication ATP-dependent helicase/nuclease DNA2 isoform X1 [Oreochromis niloticus]

lg1	23914991	-3.528	T599I	XP_019214957.1	PREDICTED: DNA replication ATP-dependent helicase/nuclease DNA2 isoform X2 [Oreochromis niloticus]
lg1	23923966	-2.678	V89M	XP_003454514.1	PREDICTED: phenazine biosynthesis-like domain-containing protein [Oreochromis niloticus]
lg1	24025079	-7.436	G433E	XP_005458993.1	PREDICTED: delta-like protein 1 [Oreochromis niloticus]
lg1	24055261	-3.536	T40S	XP_003454525.2	PREDICTED: protein MIS12 homolog [Oreochromis niloticus]
lg1	24315473	-6.595	E74G	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24315473	-6.595	E74G	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24315481	-3.937	D77N	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24315481	-3.937	D77N	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24315940	-4.123	T110I	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24315940	-4.123	T110I	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316076	-2.895	K122R	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316076	-2.895	K122R	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316700	-3.526	Q200R	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316700	-3.526	Q200R	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316717	-2.532	S206R	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316717	-2.532	S206R	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316735	-3.099	A212P	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24316735	-3.099	A212P	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24317071	-6.026	R290C	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24317071	-6.026	R290C	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24317164	-2.718	L321V	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24317164	-2.718	L321V	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24319444	-2.766	A727D	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24319444	-2.766	A727D	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24328668	-2.631	S2074Y	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24328668	-2.631	S2074Y	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]

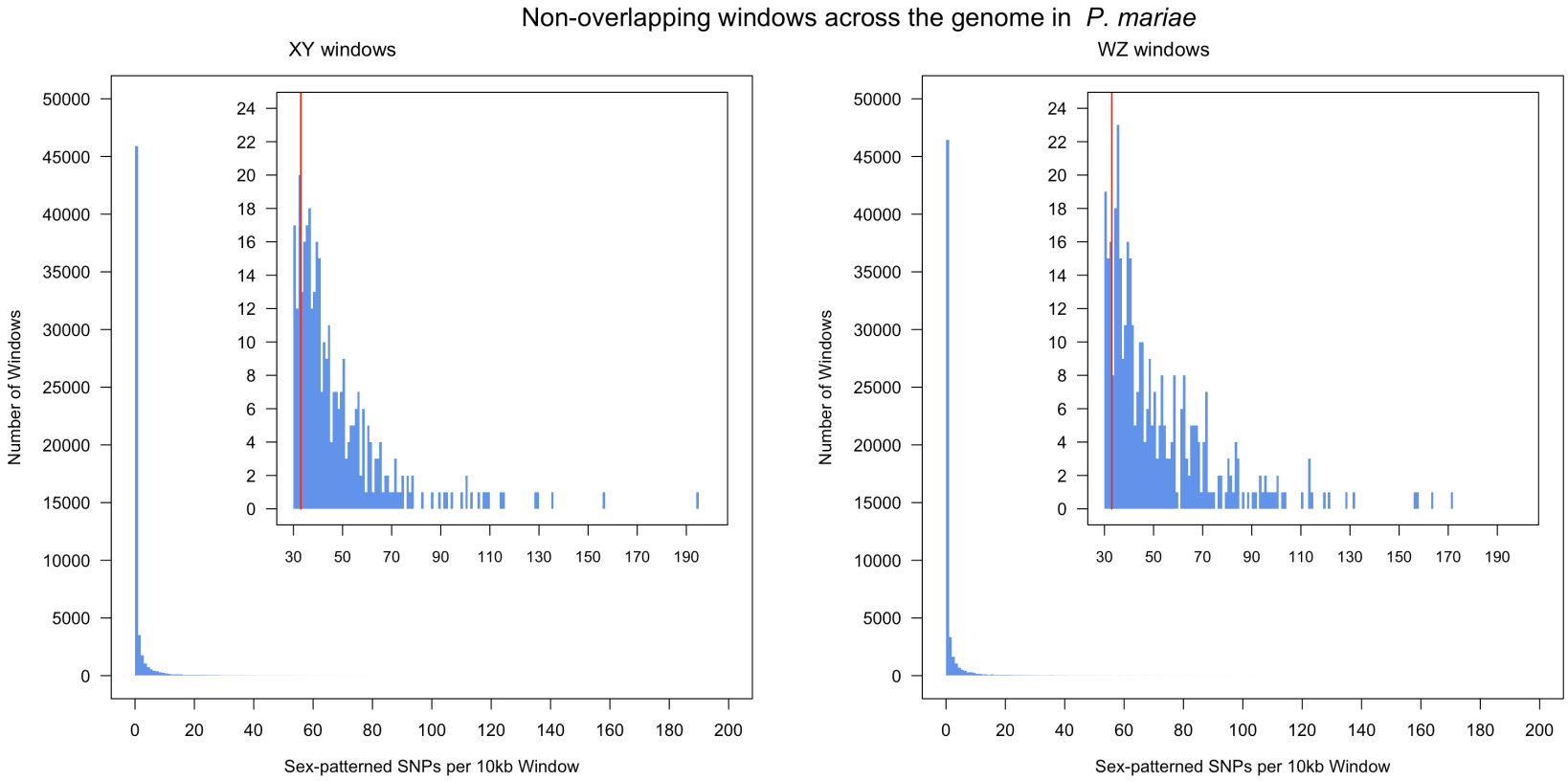
lg1	24334821	-2.843	H2835Y	XP_005459003.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24334821	-2.843	H2835Y	XP_005459004.1	PREDICTED: centromere protein F [Oreochromis niloticus]
lg1	24352086	-2.709	A175S	XP_005459005.1	PREDICTED: potassium channel subfamily K member 2 [Oreochromis niloticus]
lg1	24908095	-6.969	G3354R	XP_005459028.3	PREDICTED: LOW QUALITY PROTEIN: usherin [Oreochromis niloticus]
lg1	30555881	-4.366	Y438H	XP_013130441.1	PREDICTED: semaphorin-7A [Oreochromis niloticus]
lg1	30567873	-3.153	V417A	XP_019216233.1	PREDICTED: semaphorin-7A [Oreochromis niloticus]
lg1	34781823	-3.138	N3121S	XP_019216754.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X9 [Oreochromis niloticus]
lg1	34781823	-3.138	N3133S	XP_019216742.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X3 [Oreochromis niloticus]
lg1	34781823	-3.255	N3171S	XP_019216747.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X6 [Oreochromis niloticus]
lg1	34781823	-3.305	N3075S	XP_013128853.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X8 [Oreochromis niloticus]
lg1	34781823	-3.305	N3113S	XP_013128804.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X5 [Oreochromis niloticus]
lg1	34781823	-3.305	N3125S	XP_013128772.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X4 [Oreochromis niloticus]
lg1	34781823	-3.305	N3159S	XP_019216740.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X2 [Oreochromis niloticus]
lg1	34781823	-3.305	N3171S	XP_019216737.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X1 [Oreochromis niloticus]
lg1	34781823	-3.555	N2657S	XP_019216750.1	PREDICTED: vacuolar protein sorting-associated protein 13C isoform X7 [Oreochromis niloticus]
lg4	3784385	-7.16	P275L	XP_019213141.1	PREDICTED: transmembrane protease serine 9-like [Oreochromis niloticus]

lg4	17021346	-2.904	D223N	XP_019213465.1	PREDICTED: nuclear GTPase SLIP-GC isoform X4 [Oreochromis niloticus]
lg4	17021346	-2.904	D233N	XP_019213464.1	PREDICTED: nuclear GTPase SLIP-GC isoform X3 [Oreochromis niloticus]
lg4	17021346	-2.904	D236N	XP_019213463.1	PREDICTED: nuclear GTPase SLIP-GC isoform X2 [Oreochromis niloticus]
lg4	17021346	-2.904	D250N	XP_019213461.1	PREDICTED: nuclear GTPase SLIP-GC isoform X1 [Oreochromis niloticus]
lg4	17021346	-2.904	D250N	XP_019213462.1	PREDICTED: nuclear GTPase SLIP-GC isoform X1 [Oreochromis niloticus]
lg12	21938944	-3.844	E70G	XP_019220805.1	PREDICTED: beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase-like [Oreochromis niloticus]
lg22	28006697	-3.39	H242R	XP_013132566.1	PREDICTED: hepatic lectin-like [Oreochromis niloticus]
lg22	31363827	-4.752	V88E	XP_019206259.1	PREDICTED: major histocompatibility complex class I-related gene protein isoform X2 [Oreochromis niloticus]
lg22	31363827	-4.752	V128E	XP_005464783.1	PREDICTED: major histocompatibility complex class I-related gene protein isoform X1 [Oreochromis niloticus]
lg22	31363827	-4.752	V128E	XP_019206258.1	PREDICTED: major histocompatibility complex class I-related gene protein isoform X1 [Oreochromis niloticus]

Appendix C. Distribution of sex-patterned windows both the XY and ZW direction for *C. zillii*. The vertical red line denotes the threshold of 14 sex-patterned SNPs per non-overlapping 10kb window.



Appendix D. Distribution of sex-patterned windows both the XY and ZW direction for *P. mariae*.



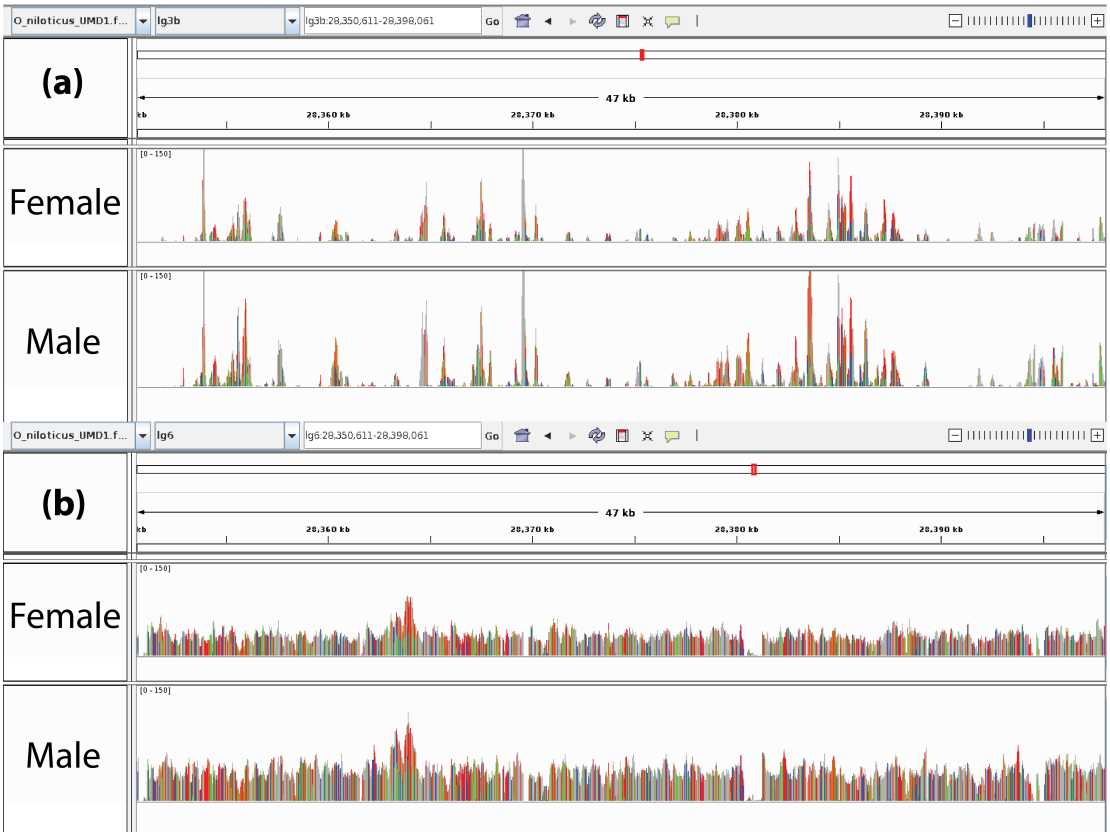
Appendix E. ***Mean and standard deviations for coverage on each linkage group in both sexes for each species.***

***O. mossambicus* Males**			***O. mossambicus* Females**		
Linkage Group	Mean Coverage	SD Coverage	Linkage Group	Mean Coverage	SD Coverage
lg1	15.072881	7.510289	lg1	17.500682	8.565871
lg2	14.411598	8.455265	lg2	16.658423	9.589168
lg3a	12.989705	13.935358	lg3a	15.031033	15.669524
lg3b	9.569823	11.772444	lg3b	11.02999	13.216034
lg4	13.485938	9.182332	lg4	15.570901	10.436702
lg5	14.716345	7.261532	lg5	17.057815	8.372717
lg6	13.904631	8.298762	lg6	16.073179	9.445153
lg7	14.681233	7.645773	lg7	17.014533	8.73492
lg8	14.494453	8.852899	lg8	16.766209	9.945082
lg9	14.244507	9.255573	lg9	16.493706	10.474941
lg10	14.255769	7.646688	lg10	16.507364	8.782459
lg11	14.345793	8.485071	lg11	16.559127	9.880164
lg12	14.20905	8.356593	lg12	16.411751	9.56297
lg13	14.305325	7.655967	lg13	16.553779	8.792714
lg14	14.388751	9.729656	lg14	16.645303	10.888705
lg15	14.579453	8.325055	lg15	16.866448	9.364484
lg16	13.968714	8.279962	lg16	16.271255	9.284333
lg17	14.335105	7.657539	lg17	16.572024	8.686164
lg18	13.838513	8.814262	lg18	15.962353	10.141897
lg19	14.616929	7.486629	lg19	16.840556	8.542024
lg20	14.573164	8.51288	lg20	16.791762	9.787684
lg22	13.297469	8.647439	lg22	15.363818	9.897297
lg23	13.914922	8.303437	lg23	16.079671	9.466752

***C. zillii M*ales**			***C. zillii M*ales**		
Linkage Group	Mean Coverage	SD Coverage	Linkage Group	Mean Coverage	SD Coverage
lg1	10.596521	6.969284	lg1	21.149188	13.104874
lg2	9.942709	7.876346	lg2	19.742896	13.713872
lg3a	8.257237	12.736156	lg3a	16.204721	21.39834
lg3b	6.287113	11.883804	lg3b	12.221796	19.829991
lg4	9.387764	7.752483	lg4	18.678136	14.312553
lg5	10.458641	7.565799	lg5	20.819012	13.564884
lg6	9.318728	8.020963	lg6	18.515503	14.539393
lg7	10.383966	7.435652	lg7	20.680368	13.360413
lg8	10.171474	8.526527	lg8	20.194699	15.00114
lg9	10.035454	9.133731	lg9	19.823036	15.409222
lg10	9.997079	7.554817	lg10	19.838817	13.47781
lg11	10.037083	7.076011	lg11	19.936534	13.049322
lg12	9.853281	8.024726	lg12	19.667894	14.063837
lg13	10.152573	7.458746	lg13	20.245349	13.199944
lg14	9.790472	8.502943	lg14	19.432543	14.57086
lg15	10.171278	7.409991	lg15	20.188845	13.352734
lg16	9.505427	7.621908	lg16	18.888128	14.081751
lg17	9.697447	7.194324	lg17	19.288522	13.262691
lg18	9.512667	7.834449	lg18	18.909868	14.452037
lg19	10.346381	9.194247	lg19	20.519808	14.969985
lg20	10.222036	7.787426	lg20	20.357663	14.069334
lg22	9.183239	8.241179	lg22	18.23304	14.832648
lg23	9.645256	7.502549	lg23	19.148916	13.513136

***P. mariae M*ales**			***P. mariae Fem*ales**		
Linkage Group	Mean Coverage	SD Coverage	Linkage Group	Mean Coverage	SD Coverage
lg1	27.586089	17.051862	lg1	23.250376	14.65457
lg2	26.338618	17.122994	lg2	22.199543	14.67755
lg3a	21.560837	23.93878	lg3a	18.292408	20.693019
lg3b	15.726941	24.386729	lg3b	13.387841	21.125724
lg4	24.635426	19.183665	lg4	20.80295	16.480717
lg5	27.032834	16.866936	lg5	22.716314	14.445622
lg6	24.368306	18.770332	lg6	20.575059	16.02062
lg7	27.041577	17.415682	lg7	22.829617	15.0585
lg8	26.233266	19.329527	lg8	22.148098	16.618172
lg9	25.45253	18.365228	lg9	21.486999	15.844075
lg10	25.75609	17.719893	lg10	21.753811	15.350561
lg11	26.004551	17.815674	lg11	21.958308	15.397963
lg12	25.738709	18.81583	lg12	21.737333	16.205874
lg13	26.635576	17.801566	lg13	22.473435	15.396319
lg14	26.074113	20.573101	lg14	22.018032	17.768346
lg15	26.416052	17.673805	lg15	22.295102	15.27328
lg16	24.659373	18.754825	lg16	20.759072	15.903618
lg17	25.366745	17.629497	lg17	21.429662	15.14619
lg18	24.860921	19.344797	lg18	20.964719	16.527913
lg19	26.279588	17.222373	lg19	22.221394	14.83664
lg20	26.531783	18.263556	lg20	22.424746	15.821081
lg22	23.905549	19.903309	lg22	20.210227	17.169371
lg23	25.152719	18.62353	lg23	21.265467	15.969692

Appendix F. Coverage in *P. mariae* from males and females on (a) LG3 and (b) LG6.



Bibliography

1. Bull, J. J. & Vogt, R. C. Temperature-dependent sex determination in turtles. *Science.* **206,** 1186–1188 (1979).
2. Ferguson, M. Temperature of egg incubation determines sex in *Alligator mississippiensis*. *Nature* **296,** 850–853 (1982).
3. Eicher, E. M. & Washburn, L. L. Genetic control of primary sex determination in mice. *Annu. Rev. Genet.* **20,** 327–360 (1986).
4. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. Male development of chromosomally female mice transgenic for Sry. *Nature* **351,** 117–121 (1991).
5. Conover, D. & Kynard, B. Environmental Sex Determination: Interaction of Temperature and Genotype in a Fish. *Science.* **213,** 577–579 (1981).
6. Strüssmann, C. A., Saito, T., Usui, M., Yamada, H. & Takashima, F. Thermal Thresholds and Critical Period of Thermolabile Sex Determination in Two Atherinid Fishes, *Odontesthes bonariensis* and *Patagonina hatcheri*. *J. Exp. Zool.* **177,** 167–177 (1997).
7. Ross, R. M. Growth and sexual strategies in the fish *Thalassoma duperrey* (Labridae), a protogynous hermaphrodite. *Environ. Biol. Fishes* **10,** 253–259 (1984).
8. Hewitt, J. & Hewiit, G. Inter-population sex chromosome polymorphism in the grasshopper *Podisma pedestris*. *Chromosoma* **31,** 291–308 (1970).
9. Bridges, C. The origin of variations in sexual and sex-limited characteristics. *Am. Nat.* **56,** 51–63 (1922).

10. Burtis, K. C. & Baker, B. S. Drosophila doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell* **56,** 997–1010 (1989).

11. Wilhelm, D., Palmer, S. & Koopman, P. Sex determination and gonadal development in Mammals. *Physiol. Rev.* **87,** 1–28 (2007).

12. Smith, C. a *et al.* The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature* **461,** 267–271 (2009).

13. Yoshimoto, S. *et al.* A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1–6 (2008).

14. Fisher, R. A. The evolution of dominance. *Biol. Rev.* **6,** 345–368 (1931).

15. Rice, W. R. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution.* **41,** 911–914 (1987).

16. van Doorn, G. S. & Kirkpatrick, M. Transitions between male and female heterogamety caused by sex-antagonistic selection. *Genetics* **186,** 629–645 (2010).

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

18. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14,** 113–124 (2013).

19. Muller, H. The relation of recombination to mutational advance. *Mutat. Res.* **1,** 2–9 (1964).

20. Green, M. Muller's Ratchet and the evolution of supernumerary chromosomes. *Genome* **33,** 818–824 (1990).

21. Blaser, O., Grossen, C., Neuenschwander, S. & Perrin, N. Sex-chromsome turnovers induced by deleterious mutation load. *Evolution.* **67,** 635–645 (2012).

22. Blaser, O., Neuenschwander, S. & Perrin, N. Sex-chromosome turnovers: the hot-potato model. *Am. Nat.* **183,** 140–146 (2014).

23. Lyon, M. F. Sex chromatin and gene action in the mammalian X-chromosome. *Am. J. Hum. Genet.* **14,** 135–148 (1962).

24. Barr, M. L. Sex chromatin and the phenotype in man. *Science.* **130,** 679–685 (1959).

25. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190,** 372–373 (1961).

26. Ohno, S., Kaplan, W. & Kinosita, R. X-chromosome behavior in germ and somatic cells of *Rattus norvegicus*. *Exp. Cell Res.* **22,** 535–544 (1961).

27. Davidson, R. G., Nitowsky, H. M. & Childs, B. Demonstration of two populations of cells in the Human female heterozygous for Glucose-6-Phosphate Dehydrogenase variants. *Proc. Natl. Acad. Sci.* **50,** 481–485 (1963).

28. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137 (1996).

29. Beletskii, A., Hong, Y.-K., Pehrson, J., Egholm, M. & Strauss, W. M. PNA interference mapping demonstrates functional domains in the noncoding RNA

Xist. *Proc. Natl. Acad. Sci.* **98,** 9215–9220 (2001).

30. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508,** 494–499 (2014).

31. Chen, C.-K. *et al.* Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science.* **354,** 468–472 (2016).

32. Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. & Davis, R. L. roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. *Cell* **88,** 445–457 (1997).

33. Kelley, R. L. *et al.* Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98,** 513–522 (1999).

34. Park, Y., Kelley, R. L., Oh, H., Kuroda, M. I. & Meller, V. H. Extent of chromatin spreading determined by roX RNA recruitment of MSL proteins. *Science.* **298,** 1620–1623 (2002).

35. Muller, H. Evidence of the precision of genetic adaptation. *Harvey Lect.* **43,** 165–229 (1950).

36. Dobzhansky, T. The X-chromosome in the larval salivary glands of hybrids *Drosophila insularis* × *Drosophila tropicalis*. *Chromosoma* **8,** 691–698 (1956).

37. Mukherjee, A. S. & Beermann, W. Synthesis of ribonucleic acid by the X-chromosomes of *Drosophila melanogaster* and the problem of dosage compensation. *Nature* **207,** 785–786 (1965).

38. Belote, J. M. & Lucchesi, J. C. Control of X chromosome transcription by the maleless gene in Drosophila. *Nature* **285,** 573–575 (1980).

39. Matsuda, M. *et al.* DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature* **417,** 559–563 (2002).

40. Nanda, I. *et al.* A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11778–11783 (2002).

41. Myosho, T. *et al.* Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*. *Genetics* **191,** 163–170 (2012).

42. Hattori, R. S. *et al.* A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 2955–2959 (2012).

43. Kamiya, T. *et al.* A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (fugu). *PLoS Genet.* **8,** e1002798 (2012).

44. Yano, A. *et al.* An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*. *Curr. Biol.* **22,** 1423–1428 (2012).

45. Li, M. *et al.* A tandem duplicate of anti-müllerian hormone with a missense SNP on the Y Chromosome is essential for male sex determination in Nile Tilapia, *Oreochromis niloticus*. *PLoS Genet.* **11,** e1005678 (2015).

46. Eshel, O. *et al.* Identification of male-specific amh duplication, sexually differentially expressed genes and microRNAs at early embryonic development of Nile tilapia (*Oreochromis niloticus*). *BMC Genomics* **15,** 774 (2014).

47. Koshimizu, E., Strüssmann, C. A., Okamoto, N., Fukuda, H. & Sakamoto, T. Construction of a genetic map and development of DNA markers linked to the sex-determining locus in the Patagonian pejerrey (*Odontesthes hatcheri*). *Mar. Biotechnol.* **12,** 8–13 (2010).

48. Kikuchi, K. & Hamaguchi, S. Novel sex-determining genes in fish and sex chromosome evolution. *Dev. Dyn.* **242,** 339–353 (2013).

49. Morinaga, C. *et al.* The hotei mutation of medaka in the anti-Müllerian hormone receptor causes the dysregulation of germ cell and sexual development. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 9691–9696 (2007).

50. Nakamoto, M. *et al.* Gonadal sex differentiation and expression of Sox9a2, Dmrt1, and Foxl2 in *Oryzias luzonensis*. *Genesis* **47,** 289–299 (2009).

51. Eshel, O., Shirak, A., Weller, J. I., Hulata, G. & Ron, M. Linkage and physical mapping of sex region on LG23 of Nile tilapia (*Oreochromis niloticus*). *G3* **2,** 35–42 (2012).

52. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5,** 288–298 (2004).

53. Lee, B.-Y., Hulata, G. & Kocher, T. D. Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). *Heredity.* **92,** 543–549 (2004).

54. Ser, J. R., Roberts, R. B. & Kocher, T. D. Multiple interacting loci control sex determination in lake Malawi cichlid fish. *Evolution.* **64,** 486–501 (2010).

55. Roberts, R. B., Ser, J. R. & Kocher, T. D. Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. *Science.* **326,** 998–1001 (2009).

56. Parnell, N. F. & Streelman, J. T. Genetic interactions controlling sex and color establish the potential for sexual conflict in Lake Malawi cichlid fishes. *Heredity.* **110,** 239–246 (2013).

57. Böhne, A., Wilson, C. A., Postlethwait, J. H. & Salzburger, W. Variations on a theme: Genomics of sex determination in the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* **17,** 883 (2016).

58. Roberts, N. B. *et al.* Polygenic sex determination in the cichlid fish, *Astatotilapia burtoni*. *BMC Genomics* **17,** 835 (2016).

59. Cnaani, A. *et al.* Genetics of sex determination in tilapiine species. *Sex. Dev.* **2,** 43–54 (2008).

60. Kudo, Y. *et al.* A microsatellite-based genetic linkage map and putative sex-determining genomic regions in Lake Victoria cichlids. *Gene* **560,** 156–164 (2015).

61. Yoshida, K. *et al.* B chromosomes have a functional effect on female sex determination in Lake Victoria cichlid fishes. *PLoS Genet.* **7,** e1002203 (2011).

62. Clark, F. E. *et al.* Dynamic sequence evolution of a sex-associated B chromosome in Lake Malawi cichlid fish. *J. Hered.* **108,** 53–62 (2017).

63. Peterson, E. N., Cline, M. E., Moore, E. C., Roberts, N. B. & Roberts, R. B. Genetic sex determination in *Astatotilapia calliptera*, a prototype species for the Lake Malawi cichlid radiation. *Sci. Nat.* **104,** 41 (2017).

64. Lee, B.-Y., Penman, D. J. & Kocher, T. D. Identification of a sex-determining region in Nile tilapia (*Oreochromis niloticus*) using bulked segregant analysis. *Anim. Genet.* **34,** 1–5 (2003).

65. Lee, B.-Y. *et al.* Genetic and physical mapping of sex-linked AFLP markers in Nile tilapia (*Oreochromis niloticus*). *Mar. Biotechnol.* **13,** 557–562 (2011).

66. Ezaz, M. T. *et al.* Isolation and physical mapping of sex-linked AFLP markers in nile tilapia (*Oreochromis niloticus* L.). *Mar. Biotechnol.* **6,** 435–445 (2004).

67. Palaiokostas, C. *et al.* Mapping and validation of the major sex-determining region in Nile tilapia (*Oreochromis niloticus* L.) using RAD sequencing. *PLoS One* **8,** e68389 (2013).

68. Eshel, O. *et al.* Fine-mapping of a locus on linkage group 23 for sex determination in Nile tilapia (*Oreochromis niloticus*). *Anim. Genet.* **42,** 222–224 (2011).

69. Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J. & Kocher, T. D. A high quality assembly of the Nile tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genomics* **18,** 341 (2017).

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

71. Rice, W. R. Evolution of the Y sex in animals: Y chromosomes evolve through the degeneration of autosomes. *Bioscience* **46,** 331–343 (1996).

72. Hedrick, P. *Genetics of Populations*. (Jones and Bartlett Publishers, 2011).

73. Mank, J. E. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* **29,** 677–683 (2013).

74. Vicoso, B. & Bachtrog, D. Progress and prospects toward our understanding of the evolution of dosage compensation. *Chromosome Res.* **17,** 585–602 (2009).

75. Heard, E. & Disteche, C. M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev.* **20,** 1848–1867 (2006).

76. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508,** 488–493 (2014).

77. Ross, J. A., Urton, J. R., Boland, J., Shapiro, M. D. & Peichel, C. L. Turnover of sex chromosomes in the stickleback fishes (Gasterosteidae). *PLoS Genet.* **5,** e1000391 (2009).

78. FASTQC. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

79. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513,** 376–381 (2014).

80. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2013).

81. Picard. Available at: http://picard.sourceforge.net.

82. DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

83. Kofler, R., Pandey, R. V. & Schlötterer, C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27,** 3435–3436 (2011).

84. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2012).

85. Cingolani, P. *et al.* A program for annotating and predicting the effects of

single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.* **6,** 80–92 (2012).

86. Cingolani, P. *et al.* Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3,** 35 (2012).

87. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

88. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).

89. Lee, B.-Y. *et al.* A second-generation genetic linkage map of tilapia (*Oreochromis* spp.). *Genetics* **170,** 237–244 (2005).

90. Oliveira, C. & Wright, J. M. Molecular cytogenetic analysis of heterochromatin in the chromosomes of tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). *Chromosome Res.* **6,** 205–211 (1998).

91. Chibalina, M. V & Filatov, D. A. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* **21,** 1475–1479 (2011).

92. Zhou, Q. & Bachtrog, D. Chromosome-wide gene silencing initiates Y degeneration in Drosophila. *Curr. Biol.* **22,** 522–525 (2012).

93. Chen, S. *et al.* Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat. Genet.* **46,** 253–260 (2014).

94. Mair, G. C., Abucay, J. S., Skibinski, D. O. F., Abella, T. A. & Beardmore, J. A. Genetic manipulation of sex ratio for the large-scale production of all-male

tilapia, *Oreochromis niloticus*. *Can. J. Fish. Aquat. Sci.* **54,** 396–404 (1997).

95. Lee, B. Y. & Kocher, T. D. Exclusion of Wilms tumour (WT1b) and ovarian cytochrome P450 aromatase (CYP19A1) as candidates for sex determination genes in Nile tilapia (*Oreochromis niloticus*). *Anim. Genet.* **38,** 85–86 (2007).

96. Massagué, J. How cells read TGF-β signals. *Nat. Rev. Mol. Cell Biol.* **1,** 169–178 (2000).

97. Patmore, D. M. *et al.* In vivo regulation of TGF-β by R-Ras2 revealed through loss of the RasGAP protein NF1. *Cancer Res.* **72,** 5317–5327 (2012).

98. Rhee, J.-S., Lee, Y.-M., Raisuddin, S. & Lee, J.-S. Expression of R-ras oncogenes in the hermaphroditic fish *Kryptolebias marmoratus*, exposed to endocrine disrupting chemicals. *Comp. Biochem. Physiol. Part C Toxicol. Pharmacol.* **149,** 433–439 (2009).

99. Trukhina, A. V, Lukina, N. A., Wackerow-Kouzova, N. D. & Smirnov, A. F. The variety of vertebrate mechanisms of sex determination. *Biomed Res. Int.* **2013,** 587460 (2013).

100. MOTIF. Available at: http://www.genome.jp/tools/motif/.

101. Swain, A., Narvaez, V., Burgoyne, P., Camerino, G. & Lovell-Badge, R. Dax1 antagonizes Sry action in mammalian sex determination. *Nature* **391,** 761–767 (1998).

102. Gallardo, T. D. *et al.* Genomewide discovery and classification of candidate ovarian fertility genes in the mouse. *Genetics* **177,** 179–194 (2007).

103. Mohammed, H. *et al.* Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. *Cell Rep.* **3,** 342–349 (2013).

104. Ohe, K., Lalli, E. & Sassone-Corsi, P. A direct role of SRY and SOX proteins in pre-mRNA splicing. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 1146–1151 (2002).

105. Dammann, R. *et al.* Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat. Genet.* **25,** 315–319 (2000).

106. Burbee, D. G. *et al.* Epigenetic inactivation of RASSF1A in lung and breast cancers and malignant phenotype suppression. *J. Natl. Cancer Inst.* **93,** 691–699 (2001).

107. Shivakumar, L., Minna, J., Sakamaki, T., Pestell, R. & White, M. A. The RASSF1A tumor suppressor blocks cell cycle progression and inhibits cyclin D1 accumulation. *Mol. Cell. Biol.* **22,** 4309–4318 (2002).

108. Graves, J. Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annu. Rev. Genet.* **42,** 565–586 (2008).

109. Gammerdinger, W. J., Conte, M. A., Acquah, E. A., Roberts, R. B. & Kocher, T. D. Structure and decay of a proto-Y region in tilapia, *Oreochromis niloticus*. *BMC Genomics* **15,** 975 (2014).

110. Nagl, S. *et al.* Classification and phylogenetic relationships of African tilapiine fishes inferred from mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* **20,** 361–374 (2001).

111. Harvey, S. C., Powell, S. F., Kennedy, D. D., Mcandrew, B. J. & Penman, D. J. Karyotype analysis of *Oreochromis mortimeri* (Trewavas) and *Sarotherodon melanotheron* (Rüppell). *Aquac. Res.* **33,** 339–342 (2002).

112. Sex_SNP_Finder_GA. Available at: https://github.com/Gammerdinger/sex-

SNP-finder.

113. Nei, M. *Molecular evolutionary genetics*. (Columbia university press, 1987).

114. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).

115. Quinlan, A. R. & Hall, I. M. Genome analysis BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

116. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44,** D110–D115 (2015).

117. Koboldt, D. C. *et al.* VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

118. Varscan_multiple. Available at: https://github.com/Gammerdinger/Varscan_multiple.

119. Kirkpatrick, M. & Guerrero, R. Signatures of sex-antagonistic selection on recombining sex chromosomes. *Genetics* **197,** 531–541 (2014).

120. Munger, S. C. *et al.* Elucidation of the transcription network governing mammalian sex determination by exploiting strain-specific susceptibility to sex reversal. *Genes Dev.* **23,** 2521–2536 (2009).

121. Miyamoto, Y., Taniguchi, H., Hamel, F., Silversides, D. W. & Viger, R. S. A GATA4/WT1 cooperation regulates transcription of genes required for mammalian sex determination and differentiation. *BMC Mol. Biol.* **9,** 1–18 (2008).

122. Bachtrog, D. *et al.* Are all sex chromosomes created equal? *Trends Genet.* **27,** 350–357 (2011).

123. Rice, W. R. Sexually antagonistic genes: experimental evidence. *Science.* **256,** 1436–1439 (1992).

124. Bachtrog, D. *et al.* Sex determination: why so many ways of doing it? *PLoS Biol.* **12,** e1001899 (2014).

125. Sánchez, L. & Chaouiya, C. Primary sex determination of placental mammals: a modelling study uncovers dynamical developmental constraints in the formation of Sertoli and granulosa cells. *BMC Syst. Biol.* **10,** 37 (2016).

126. Gammerdinger, W. J., Conte, M. A., Baroiller, J.-F., D'Cotta, H. & Kocher, T. D. Comparative analysis of a sex chromosome from the blackchin tilapia, *Sarotherodon melanotheron*. *BMC Genomics* **17,** 808 (2016).

127. Takahashi, T. Systematics of Tanganyikan cichlid fishes (Teleostei: Perciformes). *Ichthyol. Res.* **50,** 367–382 (2003).

128. Takahashi, T. Greenwoodochromini Takahashi from Lake Tanganyika is a junior synonym of Limnochromini Poll (Perciformes: Cichlidae). *J. Fish Biol.* **84,** 929–936 (2014).

129. Koblmüller, S., Sefc, K. M. & Sturmbauer, C. The Lake Tanganyika cichlid species assemblage: recent advances in molecular phylogenetics. *Hydrobiologia* **615,** 5–20 (2008).

130. Meyer, B. S., Matschiner, M. & Salzburger, W. A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach. *Mol. Phylogenet. Evol.* **83,** 56–71 (2015).

131. Salzburger, W., Mack, T., Verheyen, E. & Meyer, A. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evol. Biol.* **5,** 17 (2005).

132. Koblmüller, S. *et al.* Age and spread of the haplochromine cichlid fishes in Africa. *Mol. Phylogenet. Evol.* **49,** 153–169 (2008).

133. Schwarzer, J. *et al.* Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proc. R. Soc. B Biol. Sci.* **279,** 4389–4398 (2012).

134. Salzburger, W., Meyer, A., Baric, S., Verheyen, E. & Sturmbauer, C. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Syst. Biol.* **51,** 113–135 (2002).

135. Koblmüller, S., Duftner, N., Katongo, C., Phiri, H. & Sturmbauer, C. Ancient divergence in bathypelagic Lake Tanganyika deepwater cichlids: mitochondrial phylogeny of the tribe Bathybatini. *J. Mol. Evol.* **60,** 297–314 (2005).

136. Takahashi, T. & Sota, T. A robust phylogeny among major lineages of the East African cichlids. *Mol. Phylogenet. Evol.* **100,** 234–242 (2016).

137. Konings, A. *Tropheus in their natural habitat*. (Cichlid Press, 2013).

138. Takahashi, T., Sota, T. & Hori, M. Genetic basis of male colour dimorphism in a Lake Tanganyika cichlid fish. *Mol. Ecol.* **22,** 3049–3060 (2013).

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

140. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

141. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).

142. Wright, W. W., Smith, L., Kerr, C. & Charron, M. Mice that express enzymatically inactive cathepsin L exhibit abnormal spermatogenesis. *Biol. Reprod.* **68,** 680–687 (2003).

143. Boujrad, N., Ogwuegbu, S. O., Garnier, M. & Lee, C.-H. Identification of a stimulator of steroid hormone synthesis isolated from testis. **268,** 1609–1612 (1995).

144. Nakamura, M., Kobayashi, T., Chang, X.-T. & Nagahama, Y. Gonadal sex differentiation in teleost fish. *J. Exp. Zool.* **281,** 362–372 (1998).

145. Gautier, T. *et al.* Human luteinized granulosa cells secrete apoB100-containing lipoproteins. *J. Lipid Res.* **51,** 2245–2252 (2010).

146. Kim, Y. *et al.* Fgf9 and Wnt4 act as antagonistic signals to regulate mammalian sex determination. *PLoS Biol.* **4,** e187 (2006).

147. Mazen, I. *et al.* Homozygous mutation of the FGFR1 gene associated with congenital heart disease and 46,XY disorder of sex development. *Sex. Dev.* **10,** 16–22 (2016).

148. Liang, N. *et al.* Steroidogenic factor-1 is required for TGF-β3-mediated 17β-estradiol synthesis in mouse ovarian granulosa cells. *Endocrinology* **152,** 3213–3225 (2011).

149. Nicol, B. & Guiguen, Y. Expression profiling of Wnt signaling genes during

gonadal differentiation and gametogenesis in rainbow trout. *Sex. Dev.* **5,** 318–329 (2011).

150. Wang, H.-X., Li, T. Y. & Kidder, G. M. WNT2 regulates DNA synthesis in mouse granulosa cells through beta-catenin. *Biol. Reprod.* **82,** 865–875 (2010).

151. Beales, P. L., Elcioglu, N., Woolf, A. S., Parker, D. & Flinter, F. A. New criteria for improved diagnosis of Bardet-Biedl syndrome: results of a population survey. *J. Med. Genet.* **36,** 437–446 (1999).

152. Böhne, A., Heule, C., Boileau, N. & Salzburger, W. Expression and sequence evolution of aromatase cyp19a1 and other sexual development genes in East African cichlid fishes. *Mol. Biol. Evol.* **30,** 2268–2285 (2013).

153. Böhne, A., Sengstag, T. & Salzburger, W. Comparative transcriptomics in East African cichlids reveals sex- and species-specific expression and new candidates for sex differentiation in fishes. *Genome Biol. Evol.* **6,** 2567–2585 (2014).

154. Gennotte, V. *et al.* Brief exposure of embryos to steroids or aromatase inhibitor induces sex reversal in Nile tilapia (*Oreochromis niloticus*). *J. Exp. Zool. Part A Ecol. Genet. Physiol.* **323,** 31–38 (2015).

155. Göppert, C. *et al.* Inhibition of aromatase induces partial sex change in a cichlid fish: Distinct functions for sex steroids in brains and gonads. *Sex. Dev.* **10,** 97–110 (2016).

156. Kwon, J. Y., McAndrew, B. J. & Penman, D. J. Cloning of brain aromatase gene and expression of brain and ovarian aromatase genes during sexual differentiation in genetic male and female Nile tilapia *Oreochromis niloticus*.

Mol. Reprod. Dev. **59,** 359–370 (2001).

157. Heldin, C.-H., Miyazono, K. & ten Dijke, P. TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature* **390,** 465–471 (1997).

158. Miller, C. & Sassoon, D. A. Wnt-7a maintains appropriate uterine patterning during the development of the mouse female reproductive tract. *Development* **125,** 3201–3211 (1998).

159. Bernard, P. & Harley, V. R. Wnt4 action in gonadal development and sex determination. *Int. J. Biochem. Cell Biol.* **39,** 31–43 (2007).

160. Siggers, P. *et al.* A novel mouse Fgfr2 mutant, hobbyhorse (hob), exhibits complete XY gonadal sex reversal. *PLoS One* **9,** e100447 (2014).

161. Bagheri-Fam, S. *et al.* FGFR2 mutation in 46,XY sex reversal with craniosynostosis. *Hum. Mol. Genet.* **24,** 6699–6710 (2015).

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

163. van Doorn, G. S. & Kirkpatrick, M. Turnover of sex chromosomes induced by sexual conflict. *Nature* **449,** 909–912 (2007).

164. Gontan, C. *et al.* Exportin 4 mediates a novel nuclear import pathway for Sox family transcription factors. *J. Cell Biol.* **185,** 27–34 (2009).

165. Potterf, S. B. *et al.* Analysis of SOX10 function in neural crest-derived melanocyte development: SOX10-dependent transcriptional control of dopachrome tautomerase. *Dev. Biol.* **237,** 245–257 (2001).

166. Dutton, K. A. *et al.* Zebrafish colourless encodes sox10 and specifies non-

ectomesenchymal neural crest fates. *Development* **128,** 4113–4125 (2001).

167. Abbott, J. K., Nordén, A. K. & Hansson, B. Sex chromosome evolution: historical insights and future perspectives. *Proc. R. Soc. B Biol. Sci.* **284,** 20162806 (2017).

168. Gamble, T. A review of sex determining mechanisms in geckos (Gekkota: Squamata). *Sex. Dev.* **4,** 88–103 (2010).

169. Gammerdinger, W. J. *et al.* Novel sex chromosomes in three cichlid fishes from Lake Tanganyika. *J. Hered.*

170. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S.* (Springer, 2002).

171. R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing. (2016).

172. Davis, T. A., Loos, B. & Engelbrecht, A. M. AHNAK: The giant jack of all trades. *Cell. Signal.* **26,** 2683–2693 (2014).

173. Lee, I. H. *et al.* Ahnak functions as a tumor suppressor via modulation of TGFβ/Smad signaling pathway. *Oncogene* **33,** 4675–4684 (2014).

174. Mank, J. E. Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *Am. Nat.* **173,** 141–150 (2009).

175. Ranz, J., Castillo-Davis, C., Meiklejohn, C. & Hartl, D. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science.* **300,** 1742–1745 (2003).

176. Ellegren, H. & Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* **8,** 689–698 (2007).

177. Innocenti, P. & Morrow, E. H. The sexually antagonistic genes of *Drosophila melanogaster*. *PLoS Biol.* **8,** e1000335 (2010).

178. Butler, M. A., Sawyer, S. A. & Losos, J. B. Sexual dimorphism and adaptive radiation in Anolis lizards. *Nature* **447,** 202–205 (2007).

179. Scharf, I. & Meiri, S. Sexual dimorphism of heads and abdomens: Different approaches to 'being large'in female and male lizards. *Biol. J. Linn. Soc.* **110,** 665–673 (2013).