Divergence and Remarkable Diversity of the Y Chromosome in Guppies

Pedro Almeida,*[,1] Benjamin A. Sandkam ⓘ,[2] Jake Morris,[1] Iulia Darolti,[2] Felix Breden,[3] and Judith E. Mank[1,2]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom
[2]Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada
[3]Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada

*Corresponding author: E-mail: pedro.almeida@ucl.ac.uk.
Associate editor: Rogers Rebekah

Abstract

The guppy sex chromosomes show an extraordinary diversity in divergence across populations and closely related species. In order to understand the dynamics of the guppy Y chromosome, we used linked-read sequencing to assess Y chromosome evolution and diversity across upstream and downstream population pairs that vary in predator and food abundance in three replicate watersheds. Based on our population-specific genome assemblies, we first confirmed and extended earlier reports of two strata on the guppy sex chromosomes. Stratum I shows significant accumulation of male-specific sequence, consistent with Y divergence, and predates the colonization of Trinidad. In contrast, Stratum II shows divergence from the X, but no Y-specific sequence, and this divergence is greater in three replicate upstream populations compared with their downstream pair. Despite longstanding assumptions that sex chromosome recombination suppression is achieved through inversions, we find no evidence of inversions associated with either Stratum I or Stratum II. Instead, we observe a remarkable diversity in Y chromosome haplotypes within each population, even in the ancestral Stratum I. This diversity is likely due to gradual mechanisms of recombination suppression, which, unlike an inversion, allow for the maintenance of multiple haplotypes. In addition, we show that this Y diversity is dominated by low-frequency haplotypes segregating in the population, suggesting a link between haplotype diversity and female preference for rare Y-linked color variation. Our results reveal the complex interplay between recombination suppression and Y chromosome divergence at the earliest stages of sex chromosome divergence.

Key words: Poecilia reticulata, sex chromosomes, Y haplotypes, recombination suppression, linked-reads.

Introduction

Sex chromosomes diverge from each other once recombination between the X and the Y chromosomes is halted (Bergero and Charlesworth 2009; Bachtrog et al. 2011). Despite the prevalence and repeated origin of sex chromosomes (Bachtrog et al. 2014), we still know little about the initial stages of X–Y divergence. In particular, it remains unclear exactly how recombination is suppressed between emerging sex chromosomes. Classic models assume that recombination is instantly and completely arrested when one chromosome undergoes an inversion (Charlesworth et al. 2005). However, empirical studies have suggested that the earliest stages of recombination suppression may be due to shifts in local recombination hotspots (Sun et al. 2017), which could, at least initially, provide an incomplete barrier to recombination. Moreover, inversions are rare events, and fixation of a Y-linked inversion as a mechanism to achieve recombination suppression would lead to a limited number of Y haplotypes. If recombination is curtailed through means other than an inversion, variation at neutral sites can still be maintained in the population and substantial initial haplotype diversity can persist within the nonrecombining region. The mechanism of recombination suppression can therefore leave fundamentally different patterns of diversity on emerging Y chromosomes.

Observations of Y-linked male color traits in guppies (Poecilia reticulata) (Winge 1922, 1927; Haskins and Haskins 1951, 1961; Nayudu 1979), helped inspire theories of sex chromosome formation (Fisher 1931), and observations that Y-linkage of color varies by population (Haskins and Haskins 1961; Lindholm and Breden 2002; Gordon et al. 2012, 2017) have fueled speculation of a link between sex chromosome divergence and female preference for Y-linked male color combinations (Wright et al. 2017; Bergero and Charlesworth 2019; Bergero et al. 2019; Wright et al. 2019). Curiously, there appears to be a surprising diversity of Y-linked color haplotypes (Brooks and Endler 2001;

Tripathi, Hoffmann, Willing, et al. 2009) which is somewhat counterintuitive as the combination of purifying selection and linkage effects on Y chromosomes typically combine to remove diversity from these regions (Bachtrog 2013).

Following their initial role in inspiring early ideas about sex chromosome formation, guppies have resurfaced as a model for the genomic study of sex chromosomes. Recent work has suggested that *P. reticulata* shows evidence of early Y degeneration at the distal end of Chromosome 12, in a region that is also ancestral to *P. wingei*, its sister species (Stratum I) (Wright et al. 2017; Morris et al. 2018; Darolti et al. 2019; Wright et al. 2019; Darolti et al. 2020). Stratum I is defined by reduced mapping of male sequence against the female reference genome (Wright et al. 2017; Darolti et al. 2019), and this nonrecombining region is consistent with previous cytogenetic and linkage studies (Winge 1922, 1927; Winge and Ditlevsen 1947; Yamamoto 1975; Traut and Winking 2001; Lisachov et al. 2015), and a genetic map of the sex-determining region (SDR) (Tripathi, Hoffmann, Weigel, et al. 2009).

In addition, the extent of X–Y divergence varies substantially among populations of *P. reticulata* and between *P. reticulata* and *P. wingei* (Wright et al. 2017; Darolti et al. 2019, 2020), with some populations showing evidence of elevated male:female SNP density (Stratum II) without an associated drop in male read mapping. This suggests that recombination is also rare or absent beyond Stratum I in some populations, either because recombinants are selected against in these natural populations, recombination has recently been abolished, or is sufficiently low that does not fully counter the accumulation of Y-specific mutations. However, it is important to note that others have not recovered support for this region of recombination suppression using slightly different genomic approaches (Bergero et al. 2019), and it has been suggested that the guppy Y chromosome lacks any discernible divergence from the X, despite cytogenetic work showing X–Y differentiation (Nanda et al. 1992, 1993, 2014).

The guppy sex chromosome system is surprisingly old, as it is shared with *P. picta* (Darolti et al. 2019), suggesting it originated at least 20 Ma (Meredith et al. 2010). The age of the sex chromosome system, coupled with the fact that the sex chromosomes in *P. picta* are highly diverged and possess a mechanism of complete X chromosome dosage compensation in males (Darolti et al. 2019), indicates that other forces are counteracting the degeneration normally associated with nonrecombining regions to maintain the overall integrity of the Y in *P. reticulata*.

In the northern range mountains of Trinidad, downstream and upstream river populations are known to differ in male color patterns and other important life-history traits due to differences in predator and food abundance (Houde and Endler 1990; Endler 1995), and these complex phenotypes have been shown to evolve rapidly and repeatedly (Reznick et al. 1990). In order to understand sex chromosome divergence in this system, identify Y-specific sequence, and determine the mechanism of recombination suppression, we generated linked-read sequences from multiple males and females from both upstream and downstream population

pairs from three rivers in Trinidad, allowing us to build high-quality population-specific genome assemblies and generate phased haplotype sequences in Stratum I.

Given the controversy over the extent of X–Y divergence in this species (Wright et al. 2017; Bergero and Charlesworth 2019; Bergero et al. 2019; Wright et al. 2019), we first independently replicated our previous results (Wright et al. 2017; Darolti et al. 2019). We again recovered evidence of both a region of X–Y divergence shared among populations (Stratum I), and a convergently evolved region of greater X–Y divergence in upstream compared with downstream populations (Stratum II). We then use linked-read assemblies to show a significant accumulation of male-specific sequence and male-linked SNPs in Stratum I, and reveal an astonishing diversity in Y haplotype sequence both among and within populations. We show that this diversity is not associated with an inversion, and instead suggest that the lack of a structural mechanism of recombination suppression allows for the maintenance of large numbers of Y haplotypes. Taken together, our results reveal the initial stages of sex chromosome divergence and the evolutionary processes that allow Y diversity to be maintained.

Results

We collected and obtained whole-genome sequencing reads for 120 wild-caught individuals, an average of 20 male and 20 female *P. reticulata* samples for each of the Aripo, Quare and Yarra rivers in Trinidad, equally divided between downstream and upstream populations. We used a combination of 10x Genomics linked-read sequencing and paired-end Illumina sequencing (see Materials and Methods and supplementary table S1, Supplementary Material online), and after filtering alignments to the reference genome (see Materials and Methods), we recovered an average effective coverage of $\sim$30$\times$ for each male and $\sim$20$\times$ for each female.

In order to remove potentially confounding effects of the underlying genetic variation between rivers, we constructed river-specific reference genomes using the best female-linked reads' de novo assembly, based on scaffold N50 and phase block N50 values (supplementary table S2, Supplementary Material online). In all cases, these assemblies showed a high completeness and contiguity with N50 of >1, 3, and 8 Mb for Yarra, Aripo, and Quare genomes, respectively (Künstner et al. 2016) (supplementary table S3, Supplementary Material online). We anchored our assemblies to the published guppy genome, and identified a clear inverted segment of the first 10 Mb on the sex chromosome (Chromosome 12) (supplementary fig. S1, Supplementary Material online). This rearrangement is shared across the three rivers we collected from, as well as related outgroup species and our laboratory guppy population (Darolti et al. 2019, 2020), suggesting that it may be confined to the population used to scaffold the reference genome.

Sex Chromosome Divergence

Degeneration or significant divergence of the Y chromosome results in reduced male coverage when mapped to a female reference genome, and therefore, the ratio of male to female
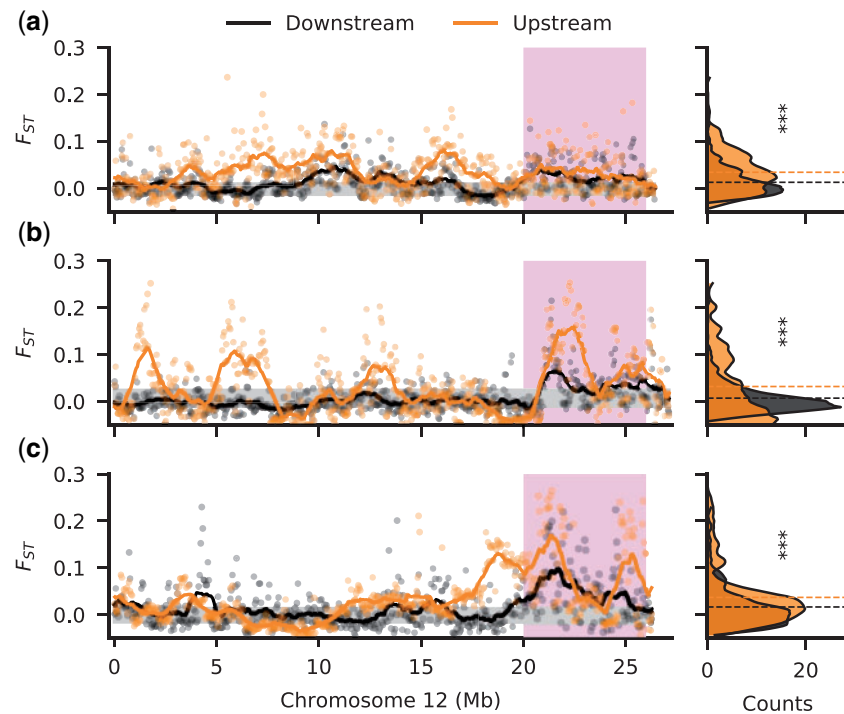
Fig. 1. Increased divergence between sexes in the upstream populations. Male–female F_{ST} along Chromosome 12 for downstream (black) and upstream (orange) populations in the Aripo (a), Quare (b), and Yarra (c) rivers. Stratum I is shaded in pink. F_{ST} was calculated in nonoverlapping windows of 50 kb. The 95% confidence interval, inferred from bootstrapping autosomal regions, is shaded in gray. The right-side density plots show the frequency (counts) of windows for the F_{ST} values. Dashed horizontal lines indicate the mean F_{ST} in each population. All upstream populations showed a significant increase in F_{ST} relative to the downstream populations (***P value < 0.001).

mapped reads can be used to identify regions where the Y chromosome has significantly degraded compared with the X (Vicoso and Bachtrog 2013; Vicoso et al. 2013; Vicoso and Bachtrog 2015; Darolti et al. 2019; Palmer et al. 2019). We have previously used this approach to identify a small region of significant Y divergence in all of the populations we assess here (Wright et al. 2017), which was present in the common ancestor with *P. wingei* (Darolti et al. 2019, 2020), designated as Stratum I. Supporting these previous findings, we again find evidence for this region in all our populations at the distal end of Chromosome 12, largely due to a drop in male mapping at 21–22 and 25–26 Mb (supplementary figs. S2 and S3, Supplementary Material online). Stratum I also exhibits elevated male:female F_{ST} in all our assessed populations (fig. 1). Together, these results are consistent with previous cytogenetic evidence (Winge and Ditlevsen 1947; Traut and Winking 2001; Lisachov et al. 2015) suggesting that Stratum I contains the male SDR, most likely within the intervals of 21–22 or 25–26 Mb.

We previously also observed elevated male SNP density (Wright et al. 2017) in each of our upstream populations of *P. reticulata* across a larger proportion of the sex chromosome compared with downstream populations, which we designated as Stratum II. However, this observation was not corrected for the large inversion on this chromosome that is present in the reference genome but absent from our study populations (supplementary fig. S1, Supplementary Material online). The region of elevated male SNP density is also apparent in *P. wingei* (Darolti et al. 2019), where it formed

independently (Darolti et al. 2020). We expect that the accumulation of male-specific mutations on the Y will lead to increased allelic difference (F_{ST}) between males and females. We observe significant increases in F_{ST} in replicate upstream compared with downstream populations along the length of Chromosome 12 ($P < 0.001$, Kruskal–Wallis test; fig. 1). Once correcting for the inversion, our results suggest that Stratum II in fact encompasses a larger extent of Chromosome 12 than we previously observed.

Male-Linked Polymorphisms

In a male heterogametic system, Y-linked alleles are only transmitted through male gametes. Thus, Y-linked regions that still retain homology to the X chromosome will show greater heterozygosity in males (XY) compared with females (XX). With a high number of Y haplotypes, explained below, we do not expect all Y-linked SNPs to be present in all males, and we therefore examined SNPs absent in all females but present in a subset of males. We observe an excess of such male-linked SNPs on Chromosome 12 (supplementary table S4, Supplementary Material online), as this chromosome was the only one in the genome where the observed number of male-linked SNPs was higher than the false positive rate (obtained from random resampling in all populations, 10,000 iterations) accounting for chromosome size. Furthermore, the distribution of male-linked SNPs on Chromosome 12 was strongly skewed, with significantly more sex-linked SNPs than expected by chance ($P < 0.001$, χ^2 test) in Stratum I based on its total length and SNP content
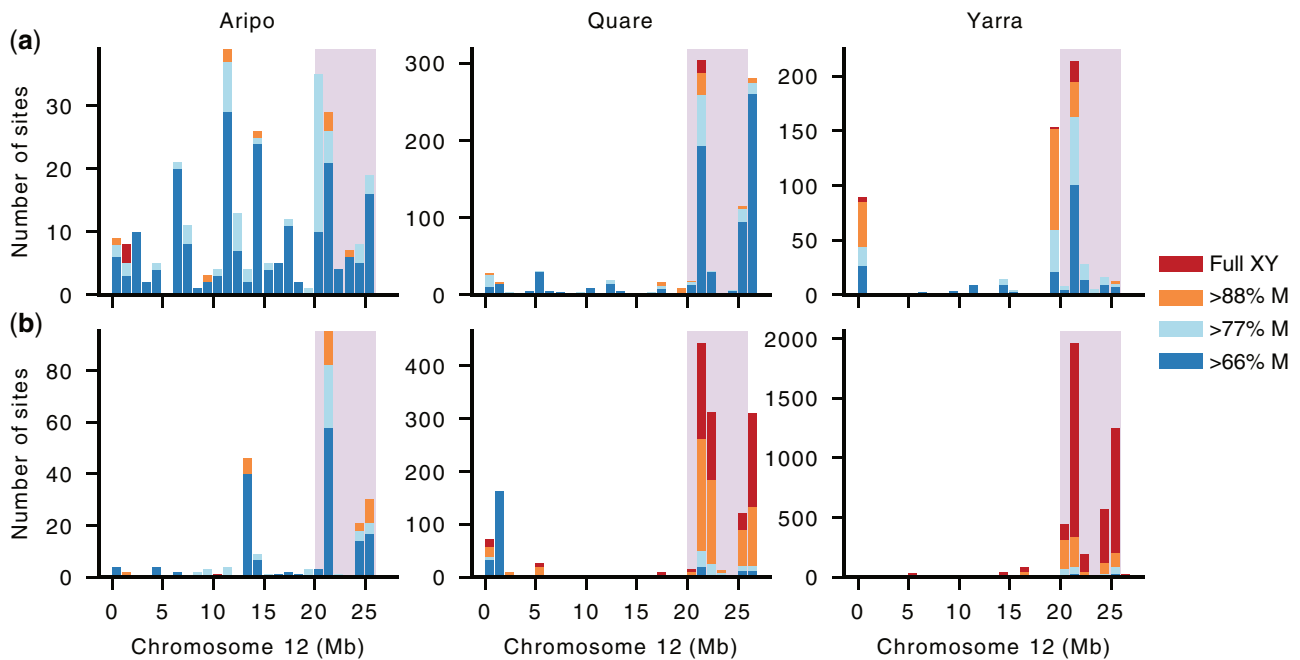
FIG. 2. Chromosome 12 shows a skew distribution of male-linked SNPs in Stratum I, particularly for upstream populations. The number of male-linked SNPs is shown for downstream (*a*) and upstream (*b*) populations. The number of SNPs compatible with complete male-linkage (in 100% of males) is shaded in red, male-linked in >88% (all but one male) in orange, >77% (all but two males) in light blue, and >66% (all but three males) in dark blue. Note that the scale of the Y axis differs between populations. SNPs are plotted in 1-Mb windows. Stratum I is shaded in pink.

(fig. 2). Within Stratum I, we generally observe increased levels of male-linked SNPs at 21–22 and 25–26 Mb, consistent with the area of greater Y divergence (fig. 1 and supplementary figs. S2 and S3, Supplementary Material online).

We then analyzed fully male-linked SNPs, those heterozygous in all males and homozygous in all females, on Chromosome 12 and observed marked differences across rivers. Consistent with a previous genotyping study of a downstream population from Aripo (Bergero et al. 2019), we also failed to recover fully male-linked SNPs in Stratum I for this watershed. These findings, however, contrast markedly with the other populations. In both Quare and Yarra populations, the distribution of fully male-linked SNPs was again significantly skewed ($P < 0.001$, χ^2 test), with either all (Quare downstream) or most (>62% in Quare upstream and both Yarra populations) of fully male-linked SNPs present in Stratum I. Interestingly, genetic diversity on the sex chromosome and autosomes, estimated as the proportion of segregating sites (Watterson's theta), was highest in Aripo compared with the other watersheds (supplementary table S5, Supplementary Material online). Genetic diversity is influenced by many factors, and is known to show a strong positive correlation with variation in recombination rates (Wallberg et al. 2015), suggesting that there may be genetic variation in this watershed for higher overall recombination rates.

The total proportion of all male-linked SNPs found in Stratum I also varied extensively between populations of the same river. In particular, upstream populations showed 1.4× (Aripo), 1.9× (Quare), and 15.6× (Yarra) more male-linked SNPs in Stratum I than their corresponding downstream populations. This difference was even more striking when considering only fully male-linked SNPs, with the upstream populations showing 21.5× (Quare) and 178.8× (Yarra) more full sex-linked SNPs than the respective downstream populations (fig. 2). Altogether, our results suggest higher sex-linkage in the distal end of Chromosome 12 and a much stronger association of sex-linkage in upstream populations, which is consistent with our previous study on these guppy populations (Wright et al. 2017). Despite these differences, a PCA analysis of all SNPs from Stratum I does not fully result in the clustering of samples by sex (supplementary fig. S4, Supplementary Material online). Although this pattern would be expected if there were ongoing genetic exchange, we think that this is unlikely given the evidence of recombination suppression predating the split between *P. reticulata* and *P. wingei* (Darolti et al. 2020), coupled with the accumulation of male-specific SNPs and repetitive sequence in this area (figs. 3 and 4). Instead, this pattern is more likely the result of incomplete lineage sorting of many Y alleles.

Lack of Inversion between X and Y Chromosomes

Inversions are often implicated as a major driver of recombination suppression in sex chromosomes (Lahn and Page 1999; Bergero et al. 2007; Vicoso et al. 2013), and can accelerate lineage sorting by creating a strong bottleneck when a rare inversion haplotype becomes fixed in a population. Given the curious lack of lineage sorting for the Y chromosome (supplementary fig. S4, Supplementary Material online), we used the linked-read information from 10x Genomics sequencing to search for differences between males and females in the aligned distance of barcodes. This method has considerably
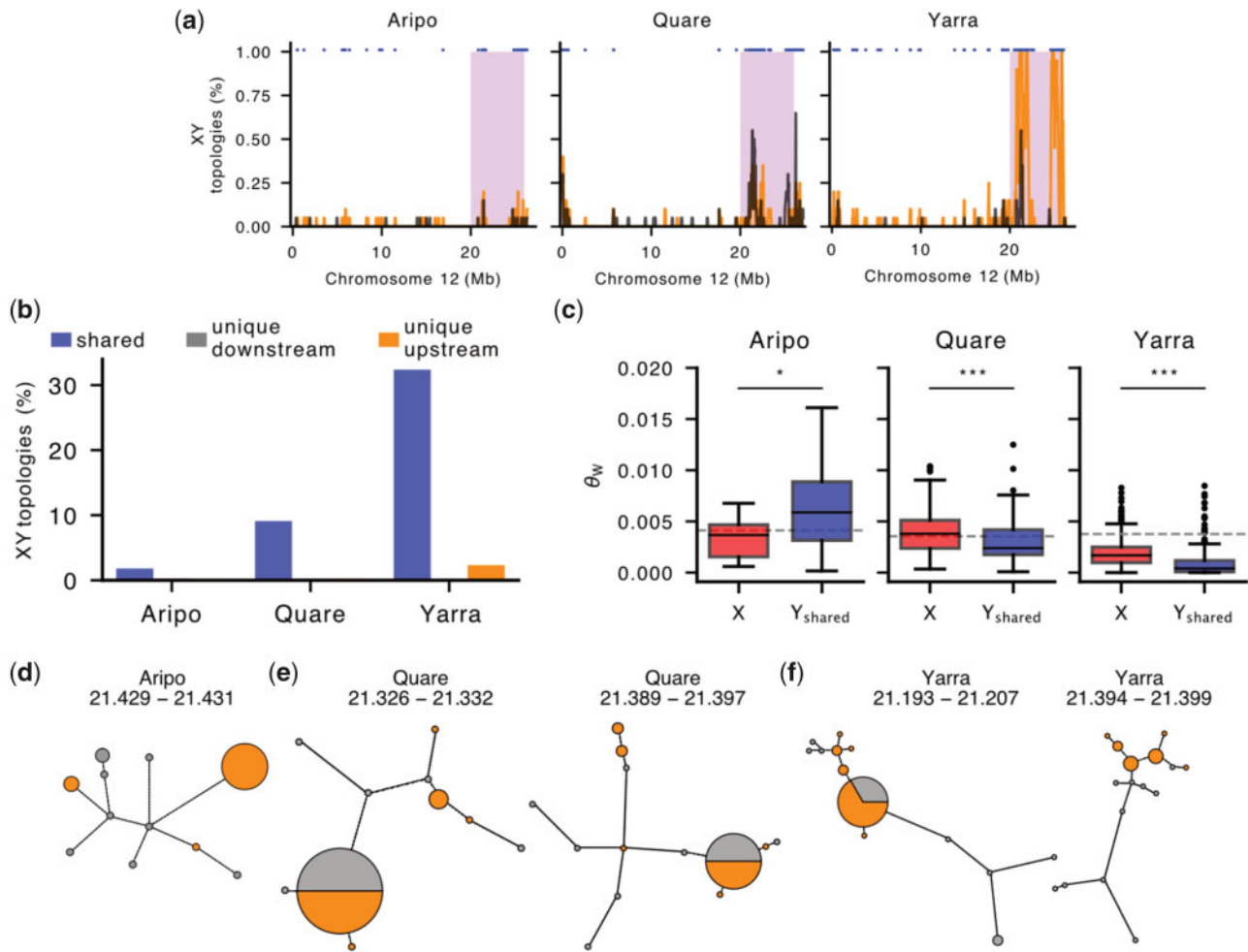
Fig. 3. The guppy Y-chromosome shows exceptional genetic diversity. (*a*) Distribution of tree topologies with separate clustering of X and Y SNPs (X–Y topology) on Chromosome 12 for downstream (black) and upstream (orange) populations in Aripo, Quare, and Yarra watersheds. Stratum I is shaded in pink. Trees were inferred from phased alignments in nonoverlapping windows of 100 SNPs with the neighbor-joining method, see Materials and Methods for more details. The blue markers at the top indicate shared Y haplotypes between populations. (*b*) Barplots showing the proportion of trees compatible with an XY sex chromosome system in Stratum I. Shared Y haplotypes were present in both populations and in >66% of males in at least one population. (*c*) Genetic diversity (Watterson's theta estimator) of X and Y haplotypes within Stratum I in the Aripo, Quare, and Yarra rivers. Y haplotypes were included only if found in both populations (Y_{shared}) and were consistent with an XY topology in at least one of them. The dashed line indicates the estimated diversity for 1,000 randomly sampled autosomal locations. ***P value < 0.001, *P value < 0.05. (*d–f*) Networks of Y haplotypes in Aripo (*d*), Quare (*e*), and Yarra (*f*) rivers. Only networks for which the maximum number of Y haplotypes could be identified in both populations are shown. The circle area is proportional to haplotype frequency with the smallest circles representing single haplotypes, and branch lengths connecting haplotypes are proportional to the number of SNPs between haplotypes. Haplotypes from downstream populations are in gray and from upstream populations in orange. The approximate location (in Mb) of each region is indicated above the network.

more power than short-read sequencing to detect rearrangements because the long molecules are more likely to span the rearranged breakpoints.

If an inversion has formed between the X and the Y chromosomes, we would expect males to be heterozygous for this structural variation and that it would be absent in females because the reference genome is also from a female. Although we could detect distinctive barcode overlaps between distant positions of Stratum I across the three rivers, we failed to identify any consistent difference between females and males in any of the populations (supplementary fig. S5, Supplementary Material online). This suggests that any potential rearrangement involving Stratum I is shared between the sexes in all populations, and thus not Y-specific (supplementary fig. S1a, Supplementary Material online). We therefore found no evidence of an inversion on Stratum I, possibly explaining the unusual Y diversity and lack of complete lineage sorting in this region.

Sequence Analysis of Y Haplotypes

To further investigate the genetic diversity of the guppy Y chromosome, we used our linked-read sequencing data to computationally phase all genotypes in our populations (see Materials and Methods). We estimated a phase switching error of <5% for all linked-read samples, with the exception of one Quare male and one Yarra female

(supplementary fig. S6, Supplementary Material online), indicating that this approach provides high phasing accuracy. We then used gene trees in nonoverlapping windows of 100 SNPs to infer Y-linked haplotypes. Briefly, if a region of the Y chromosome is linked to the SDR and recombination suppression has led to complete lineage sorting of X and Y alleles, it will only be present in males and, therefore, will form a monophyletic clade composed exclusively of male samples.

Regions where >66% of males formed a monophyletic clade largely clustered in Stratum I (fig. 3a), again predominantly between 20–22 and 24–26 Mb. Although we also observed some XY topologies at low frequency in the pseudoautosomal region (PAR), these are presumably caused by an overall reduced recombination rate (heterochiasmy) in male guppies (Bergero et al. 2019). The frequency of XY topologies in Stratum I was highest in Yarra, intermediate in Quare and lowest in Aripo (fig. 3b), in line with the distribution of sex-linked SNPs in these rivers (fig. 2). Interestingly, most of the Y haplotypes detected were shared between populations, with very few (four regions within Stratum I of Aripo, three in Quare, and 24 in Yarra Rivers, corresponding to 12.5%, <3%, and <7% of the total number of XY topologies in this region) unique to a single population (fig. 3b). This is expected if recombination on Stratum I is ancestral to *P. reticulata* populations in Trinidad and most Y haplotypes are monophyletic, and it is also consistent with our previous work showing that recombination was suppressed in this region in the common ancestor with *P. wingei* (Darolti et al. 2019).

We isolated shared male X- and Y-linked haplotypes from the gene trees described above and estimated nucleotide diversity for each aligned region. In comparison with the X chromosome, Y diversity was marginally higher in the Aripo watershed ($\theta_X = 0.003301$, $\theta_Y = 0.006854$, $P = 0.010$, Kruskal–Wallis test), and was significantly lower in Quare and Yarra (Quare $\theta_X = 0.003976$, Quare $\theta_Y = 0.003143$, Yarra $\theta_X = 0.001786$, Yarra $\theta_Y = 0.001327$, $P < 0.001$, Kruskal–Wallis test) (fig. 3c). Under a neutral model, diversity is expected to be proportional to the relative number of each chromosome in the population, therefore, X diversity is expected to be 3/4 of autosomal diversity and Y diversity is expected to be 1/4 of that in the autosomes. Furthermore, linkage effects will deplete nonrecombining regions of diversity, as Y chromosomes typically exhibit far less than 1/4 autosomal diversity (Bachtrog 2013). Both the X to autosome (X/A) as well as Y to autosome (Y/A) estimates depart from these neutral expectations. Y/A diversity was higher than expected in all three rivers, approaching ~0.35 of that observed in the autosomes in Yarra and up to 1.66 in Aripo. In contrast, the X/A diversity was close to the expected 0.75 in Aripo (0.80), but was considerably higher in Quare (1.12) and considerably lower in Yarra (0.47), suggesting that selection and/or demography may be operating differentially in males and females.

To further explore the diversity and evolutionary history of the guppy Y chromosome, we built haplotype networks for the Y-linked regions for which Y haplotypes were shared across both populations (fig. 3d–f and supplementary figs. S7 and S8, Supplementary Material online). We focused on the regions with haplotypes present in all or most males (fig. 3d–f) as these are the most informative. For both Quare and Yarra, this included two regions on Stratum I, between 21.3 and 21.4 Mb, and one region in Aripo at ~21.4 Mb. Except in Aripo where the most frequent haplotype is only present in upstream males, dominant haplotypes are shared between downstream and upstream populations. These dominant haplotypes include less than half of the male sequences sampled, and many of the Y haplotypes are observed in only one or two males, emphasizing the genetic diversity of Y chromosomes and the presence of low-frequency Y-linked haplotypes in the guppy.

The number of mutational steps separating haplotypes of the same population was considerably higher in downstream populations with an average of 10.67, 17.14, and 14.5 mutations in Aripo, Quare, and Yarra, respectively. In contrast, in the upstream populations, there were 2.8 and 1.67 mutations for Quare and Yarra, respectively (the three upstream haplotypes observed in Aripo are not directly connected). These findings indicate higher haplotype diversity in the downstream populations, which is in line with the distribution of male-linked SNPs between the two populations. Some of the downstream Y haplotypes branch directly from the upstream population and may represent episodes of downstream migration.

Characterization of Male-Specific Sequence

To identify the ancestral region of the Y chromosome in *P. reticulata*, we searched for 21-bp k-mer sequences unique to males and absent in all females (hereafter Y-mers) in each of our populations. We have previously used a Y-mer-based approach to identify a small ancestral Y region across several *Poecilia* species (Morris et al. 2018; Darolti et al. 2019). In order to evaluate the false positive rate (type-I error) for our sample sizes, we compared the number of Y-mers with female-specific k-mers (supplementary fig. S9a, Supplementary Material online). The latter can be used as a control because sequence unique to females is theoretically absent in a male heterogametic species. The false positive rate was <5% in all populations for Y-mers found in at least six males (supplementary fig. S9b, Supplementary Material online), so these thresholds were used for subsequent analyses.

Altogether, the number of population-specific Y-mers comprised ~93% of the total number of Y-mers found in our data set (fig. 4a), which suggests two nonmutually exclusive scenarios in the evolution of the Y chromosome in the guppy. The Y chromosomes have been evolving independently in each population since their split, expected if there were a large number of Y haplotypes at colonization that vary in abundance across watersheds (Fraser et al. 2015). Supporting this hypothesis, we also recovered only a very small number of Y-mers shared across the watersheds (13 Y-mers; fig. 4b), indicating a very dynamic nature of the Y chromosome. Alternatively, Y-specific sequence could also be accumulating separately within each population after recombination was suppressed in the ancestor. In line with this, most Y-mers map to repetitive rather than unique sequence
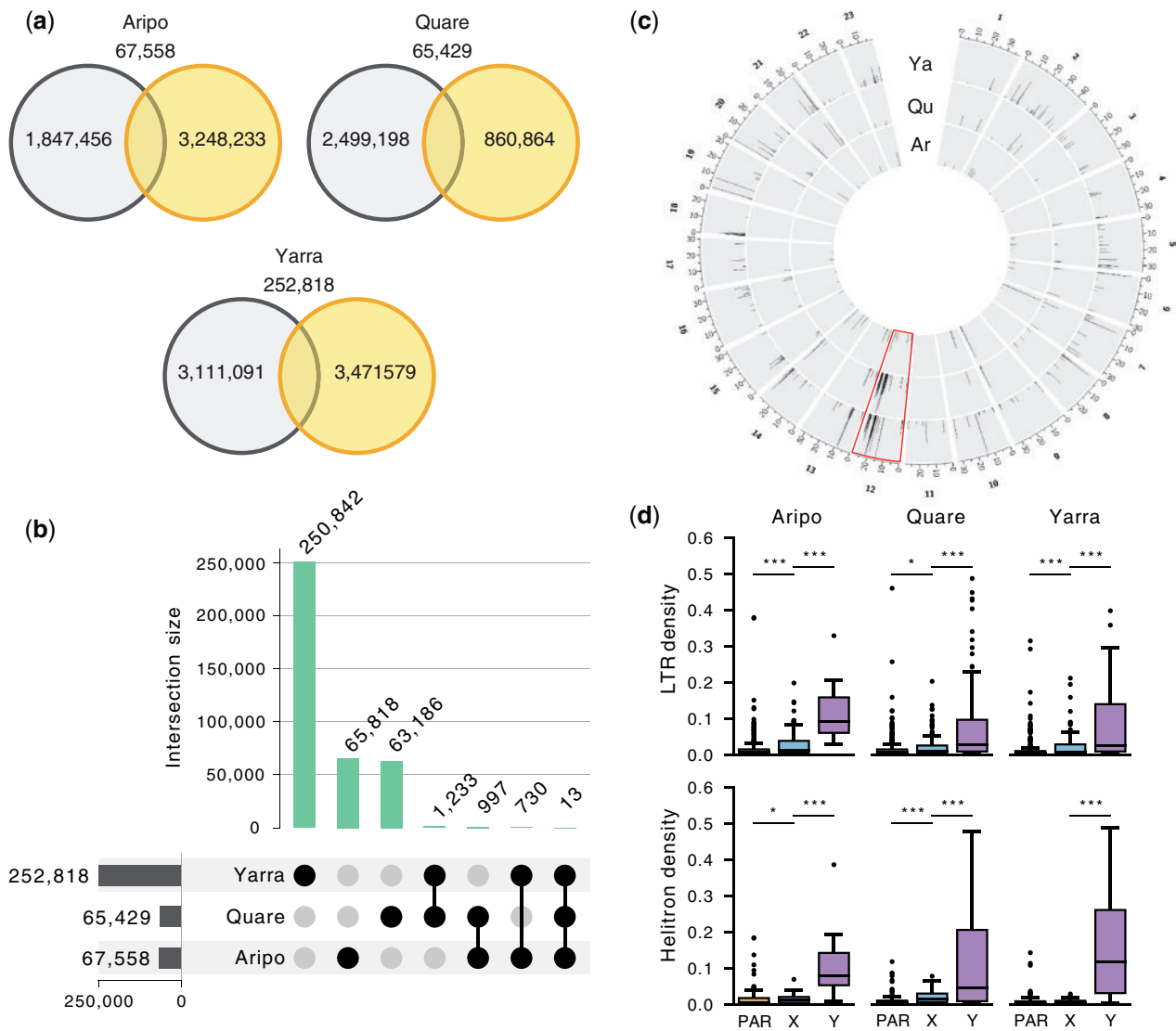
FIG. 4. Characterization of Y-linked sequence in the guppy. (*a*) Venn diagrams showing the number and distribution of Y-mers among upstream (orange) and downstream (black) populations, along with overlapping Y-mers, (*b*) Y-mers in the guppy are mostly river-specific, with only a limited number shared between watersheds. (*c*) Circos plot showing the alignment of Y-mers to the reference genome of each population. Only Y-mers with unique alignments are shown. The sex chromosome (Chromosome 12) is highlighted in red. The region with the highest number of Y-mers aligned on Chromosome 12 overlaps with Stratum I (20–26 Mb). Ar, Aripo; Qu, Quare; Ya, Yarra. (*d*) Transposable elements (TEs) have accumulated significantly more on the Y chromosome. Boxplots showing the density of LTRs and helitrons (total sequence of TEs in every nonoverlapping 50-kb window) in scaffolds enriched with Y-mers. Statistics were calculated for the Y-linked scaffolds (Y), the region of the X chromosome homologous to the SDR (X), and to the pseudoautosomal region (PAR). ***P value < 0.001, **P value < 0.01, *P value < 0.05.

in the reference genome (supplementary fig. S10, Supplementary Material online). This suggests that the build-up of repetitive sequence could be a major driver of sex chromosome divergence in the guppy, as it has been observed cytogenetically in *P. wingei* and to a less extent also in *P. reticulata* (Traut and Winking 2001; Nanda et al. 2014). Nonetheless, Stratum I was enriched for uniquely aligned Y-mers (fig. 4c), reinforcing again that this region is likely linked to the SDR and represents the region of ancestral recombination suppression.

We extracted 10x Genomics Y haplotypes (megabubbles) enriched for Y-mers (see Materials and Methods) to further characterize Y-specific sequence. The density of transposable

elements (TEs), measured as base pairs of TE per 50-kb non-overlapping windows, was significantly higher in the candidate Y scaffolds, in comparison with both the PAR and the X-linked region of the sex chromosome (*P* < 0.001, Kruskal–Wallis test; supplementary fig. S11*a*, Supplementary Material online). In particular, TE enrichment was largely driven by long-terminal repeats (LTRs) and helitrons (DNA transposons with a rolling-circle replication mechanism) that were markedly abundant in the Y scaffolds of all watersheds (fig. 4c and supplementary fig. S11*b*, Supplementary Material online). In addition, we also annotated a total of 193 potentially Y-linked genes, most of which are single copy (supplementary table S6, Supplementary Material online). This

estimate is, however, likely to include an unknown but probably considerable proportion of hitchhiking genes in these scaffolds. We could identify a homolog in the female guppy reference annotation for 69% of the genes, including in 19 out of the 22 genes common in at least two rivers. Six of these genes (htr1a-B, C6, C7, GHR, Nim1K, and UNC13B) have been previously associated with the guppy SDR and mapped to a duplicated segment between Chromosome 9 and Chromosome 12 (Dor et al. 2019).

Discussion

We used linked-read sequencing to assemble genomes for multiple individuals in downstream and upstream population pairs across three rivers in order to assess the divergence and diversity of the guppy Y chromosome. After replicating our previous identification of an ancestral nonrecombining Y region shared across all watersheds, as well as greater overall sex chromosome divergence in replicate upstream compared with downstream populations (Wright et al. 2017), we identified Y-specific sequence in order to assess the degree of male haplotype diversity. We find that, contrary to the expected depletion of genetic variation that accompanies Y chromosome divergence due to the combined effects of linkage and purifying selection (Bachtrog 2013), the Y chromosome in guppies displays a remarkable level of diversity.

Recombination Suppression

By necessity, studying the initial stages of sex chromosome divergence requires studying sex chromosome systems with very low differentiation between the X and the Y. This can be difficult given the subtlety of molecular signals that have barely begun to accumulate. It is perhaps not surprising then that our initial findings (Wright et al. 2017; Darolti et al. 2019) have been challenged by others (Bergero and Charlesworth 2019; Bergero et al. 2019). In particular, Bergero et al. (2019) neither observed any signal of Y degradation in Stratum I nor increased Y divergence in upstream compared with downstream populations (Stratum II).

We have both replicated our previous results (supplementary fig. S2, Supplementary Material online; Wright et al. 2017; Darolti et al. 2019) using the same F_{ST} approach advocated by Bergero et al. (2019), and expanded it using several other lines of evidence in conjunction with our population-specific genome assemblies. We find consistent evidence of sex chromosome divergence within Stratum I, both due to reduced read mapping in males (supplementary figs. S2 and S3, Supplementary Material online) as well as the accumulation of male-linked SNPs (fig. 2), Y-mers, and repetitive elements (fig. 4). We observe sequence divergence particularly in two regions, 21–22 and 25–26 Mb, where we also observe an excess of male-specific SNPs as well as an increase in gene tree topologies compatible with an XY sex chromosome system where Y chromosomes form a single cluster (fig. 3). These multiple concordant signatures of Y degeneration are common to all six wild populations analyzed here, as well as recent comparative studies suggesting that the region of recombination suppression is ancestral to *P. reticulata* and its sister species, *P. wingei*, and therefore predates the colonization of

Trinidad (Morris et al. 2018; Darolti et al. 2019, 2020). Interestingly, we observe differences in these patterns across all three watersheds, with the greatest effect in Yarra and least in Aripo. At this point, the cause of this variation is unclear, as we would not necessarily expect a pattern matching the phylogenetic structure of populations (Suk and Neff 2009) for this stratum, as recombination suppression predates the colonization of Trinidad (Darolti et al. 2020). The population-level differences may relate to the lack of an inversion, which in turn can lead to gradual recombination suppression, discussed in more detail below.

We also again find evidence of Stratum II (fig. 1) (Wright et al. 2017; Darolti et al. 2019), with greater male–female F_{ST} in all three replicate upstream populations compared with their downstream pair, consistent with convergence evolution of recombination suppression. After correcting for the structural difference between the reference genome and our populations (supplementary fig. S1, Supplementary Material online), elevated male–female F_{ST} in upstream populations extends over a greater length of Chromosome 12 than our previous estimates (Wright et al. 2017), and encompasses nearly the entirety of Chromosome 12 proximal to Stratum I.

At this point, it is not clear what causes the difference in divergence between upstream and downstream populations. It may be that recombination rates are similar, but recombinant males are selected against and are effectively removed before reproducing. Alternatively, recombination may be curtailed due to differences in methylation in this region between populations (Gorelick 2003; Metzger and Mank 2020), or there may be environmental effects on recombination rate (Lloyd and Jenczewski 2019). Testing these alternatives remains an exciting area for further work.

Y Chromosome Diversity

Within Stratum I, which achieved recombination suppression prior to the colonization of Trinidad (Morris et al. 2018; Darolti et al. 2020), we observe a remarkable diversity of Y chromosome haplotypes, which is counter to the general expectations of strong sweep effects and low haplotype diversity expected of nonrecombining Y regions. Instead, Y-specific sequence varies extensively across populations (fig. 4), and is supported by estimates of nucleotide variation (figs. 2 and 3c) as well as by haplotype networks (fig. 3d–f). Nucleotide diversity of Y haplotypes was also higher than expected from neutral models (fig. 3c). However, these models assume complete hemizygosity in males (for an XY system) which does not seem to be the case in the guppy Y chromosome (this study; Wright et al. 2017; Bergero and Charlesworth 2019; Bergero et al. 2019; Darolti et al. 2019).

This diversity is initially puzzling, as nonrecombining Y regions are typically depleted of variation via linkage effects and background selection. However, upon reflection, our observation is consistent with organismal reports of Y haplotype diversity within the degenerated region. Specifically, YY males are viable only when Y chromosomes from different lineages are combined (Winge and Ditlevsen 1947; Haskins et al. 1970), suggesting both that many Y chromosome haplotypes contain recessive lethal variants and that the exact Y

chromosome complement of these recessive lethal mutations varies across male lineages. Such high genetic diversity also matches well with the extraordinary phenotypic variation in male guppy coloration. Numerous reports in natural guppy populations suggest Y-linkage of many male color traits and a variety of Y-linked color combinations within populations (Houde 1997; Brooks and Endler 2001; Lindholm and Breden 2002), which together paint a picture of substantial Y diversity. Finally, it is also possible that segregating recessive lethal mutations in the PAR, which are shared between the X and the Y chromosomes, could generate associative overdominance, resulting in elevated genetic diversity on the Y (Gilbert et al. 2020).

Recombination suppression is often thought to result once a large inversion fixes on the X or the Y chromosome (Charlesworth et al. 2005; Wright et al. 2016). However, we do not observe an inversion (supplementary fig. S5, Supplementary Material online) in the nonrecombining region of the Y chromosome, even though our use of long linked-read sequence from multiple individuals offers vastly increased power to detect chromosomal rearrangements relative to more traditional short-read sequencing. This lack of inversion could account for the marked differences between populations and rivers as well as the remarkable diversity that we observe in Y haplotypes. Because inversions are rare events, fixation of a Y inversion as a mechanism to achieve recombination suppression would lead to a strong selective sweep and a limited number of Y haplotypes (Smith and Haigh 1974). If recombination is curtailed through means other than an inversion, this bottleneck would not occur, leading to substantial initial haplotype diversity within the nonrecombining region.

Several recent studies have indeed suggested that recombination suppression can proceed via means other than inversions (Nicolas et al. 2004; Chibalina and Filatov 2011; Bergero et al. 2013; Natri et al. 2013) and also that inversions may follow recombination arrest in cases where they are not the cause of the initial recombination suppression (Sun et al. 2017). Our data are consistent with this. For example, we find a significant enrichment of male-linked SNPs in Stratum I, but relatively few are fixed in all downstream populations (fig. 2). This suggests that recombination suppression can be achieved via means other than an inversion, such as through differences in sex-specific recombination hotspots (heterochiasmy) (Burt et al. 1991; Lenormand 2003; Lenormand and Dutheil 2005) or from epigenetic variation (Zhang et al. 2008). This, in turn, is expected to leave far greater Y chromosome diversity in the population.

Interestingly, the lack of an inversion as a mechanism of recombination suppression may explain some of the variance we observe in apparent stratum boundaries across populations. Without an inversion, recombination suppression will be a more gradual and incomplete process resulting in fuzzy, rather than strictly discrete boundaries between strata (Furman et al. 2020).

Despite the initial maintenance of Y haplotype diversity in the absence of an inversion, we would still expect the steady depletion of Y variation due to linkage effects and sweeps (Bachtrog 2013) without some other countering mechanism. Negative frequency-dependent selection might act to maintain multiple male color haplotypes on the Y chromosomes. Female guppies have long been known for their preference for rare or novel coloration phenotypes in males (Farr 1977; Hughes et al. 1999), and several experimental studies suggest that rare male coloration polymorphisms could be maintained by sexual selection via negative frequency-dependent selection (Endler 1988; Olendorf et al. 2006; Hughes et al. 2013; Kemp et al. 2018). Our observation of an abundance of low-frequency Y haplotypes (fig. 3) is consistent with this hypothesis. Given that at least some genes controlling male coloration in the guppy are thought to be Y-linked (Lindholm and Breden 2002), negative frequency-dependent selection could contribute to an increase in Y chromosome diversity and to the prevalence of multiple Y-linked genetic polymorphisms in the population. Future work linking specific color combinations and Y haplotypes will be interesting to test the possible relationship between female preference and Y chromosome variation.

Finally, this and our previous work (Wright et al. 2017) supports a stronger association of sex-linked polymorphisms in upstream populations (figs. 1 and 2). Both the number of Y haplotypes unique to the upstream populations (not observed in downstream populations of the same watershed), and the genetic distance between upstream Y haplotypes were generally low (fig. 3b). This suggests that most of the upstream Y-linked variation is derived from the ancestral downstream populations, in agreement with other population genetic surveys (Alexander et al. 2006; Suk and Neff 2009).

Conclusion

At the initial stages of sex chromosome divergence, population variation in the degree of sex-linkage is likely to maintain Y-linked alleles segregating in the male population (Furman et al. 2020), but the extent of such variation has been difficult to quantify in natural populations. Using phased X and Y haplotypes, our results show a remarkable population variation in the degree of sex-linkage in natural guppy populations, possibly due to the fact that recombination suppression is not based on an inversion. This, combined with other population-specific processes, such as frequency-dependent selection, act to maintain multiple Y polymorphisms segregating in natural guppy populations.

Materials and Methods

Sample Collection

We collected wild *P. reticulata* samples from three rivers (Yarra, Quare, Aripo) in the Northern Range Mountains of Trinidad in December 2016. A description of the habitats can be found in Sandkam et al. (2015). From each river, we caught between 19 and 21 males and 20 females from both upstream and downstream populations. After dispatching samples, we immediately minced and placed heads in ethanol and flash froze tubes in liquid nitrogen. All samples were collected in accordance with national and institutional ethical guidelines.

DNA Extraction and Sequencing

We prepared high-molecular weight DNA samples from all males and females from each population. We extracted DNA from ∼25 mg of head tissue following an adapted 10x Genomics HMW DNA Extraction Protocol (HMW DNA Extraction from Fresh Frozen Tissue, CG000072, Rev B. 2017) and assessed molecular weight with the Agilent Femto Pulse system prior to library preparation. We then selected ten males from each population with the highest molecular weights for 10x Genomics Chromium Sequencing. Because we were focused on phasing the X and the Y chromosomes, linked reads were not required from female samples, however, we still selected three females from each population with the highest molecular weights for 10x Chromium Sequencing. We used the 10x Genomics Chromium Genome Library Preparation Kit according to the manufacturer's protocol (No. CG00043 Chromium Genome Reagent Kit v2 User Guide), reducing the amount of starting DNA from the recommended 1.25–0.6 ng to account for the smaller genome size of *P. reticulata* (∼700 Mb) compared with the human genome, for which the protocol was optimized. The libraries were sequenced on an Illumina HiSeqX with 2×150-bp cycles using the v2.5 sequencing chemistry, resulting in an average of 346 M reads per sample (range between 159 and 494 M), representing an estimated average coverage of $68 \times$ ($32 \times$–$98 \times$) (supplementary table S1, Supplementary Material online).

For the remaining seven female samples in each population with lower molecular weight DNA, we generated additional Illumina sequencing data after extracting genomic DNA from ∼25 mg of head tissue using the DNeasy Blood & Tissue Kit (Qiagen) following the standard protocol. We prepared sequencing libraries from 1 μg DNA with the TruSeq PCRfree DNA Sample Preparation Kit, targeting an insert size of ∼350 bp according to the manufacturers' instructions (Guide No. 1000000039279). We sequenced resulting libraries on a Illumina HiSeqX with 2×150-bp cycles using the v2.5 sequencing chemistry, resulting in an average of 137 M reads per sample (93–218 M), representing $29 \times$ coverage on an average ($20 \times$–$40 \times$) (supplementary table S1, Supplementary Material online).

10x Genomics De Novo Assembly of Genomes

To construct a reference genome specific for each river, we assembled linked reads de novo for all our females with linked-read sequencing using Supernova v2.1.1 (10x Genomics). We then chose the best female assembly in each river based on scaffold N50 and phase block N50. Because Supernova can generate nearly identical haplotypes for the same sequence, we created a nonredundant assembly by removing smaller scaffolds with evidence of sequence overlap with longer scaffolds. For this, we aligned each assembly to itself with LAST v926 (Kielbasa et al. 2011), using the NEAR DNA seeding scheme and postmasking of repeats (R11). To avoid false matches caused by repetitive sequences and paralogous scaffolds, we generated orthologous alignments with "last-split," and discarded alignments with high proportions of masked sequence with "last-postmask." We

designated scaffolds as allelic variations in the assembly if they showed >90% sequence overlap and >95% sequence identity with other longer scaffolds.

We then used ARKS v1.0.2 (Coombe et al. 2018) to increase the contiguity of each of our nonredundant assemblies. ARKS uses a k-mer approach to infer graph edges by determining the Chromium barcodes associated with the best-matching contig end for each read and selecting the contig end with the largest fraction of k-mer overlap. We ran ARKS with LINKS v1.8.6 (Warren et al. 2015) using default parameters and a Jaccard index = 0.5. Because different starting assemblies can have different optimal k-mers, we tested a range of k-mer values from 40 to 100 with increases of 20. As all k-mers tested produced a longer assembly N50 than the original, we chose the k-mer value that generated fewer scaffold misassemblies relative to the guppy reference genome (Künstner et al. 2016), as determined by QUAST v5.0.2 (Gurevich et al. 2013).

Finally, we scaffolded each of our best ARKS assemblies, using the guppy reference genome (NCBI accession GCF_000633615.1) (Künstner et al. 2016) as backbone, with RaGOO v1.02 (Alonge et al. 2019). This process orders and orients assembled scaffolds relative to the chromosome-level reference genome. We ran RaGOO with a gap length of 100 and a minimum grouping confidence score = 0.3 instead of the default 0.2 in order to increase the precision of localized scaffolds. In general, only ∼1% (∼7 Mb) of the total genome size for each assembly could not be localized in the reference genome, which indicates a near-complete placement of the assembly scaffolds.

In order to verify the scaffolding made by RaGOO, we performed pairwise synteny analyses between the assemblies and the NCBI guppy reference genome used in the scaffolding procedure. We made alignments with LAST v926 (Kielbasa et al. 2011) using the NEAR DNA seeding scheme, postmasking repeats (R11) using a sensitive search with parameters "-m50 -C2." We identified orthologous alignments with "last-split -m1," discarding alignments comprised mostly of masked sequence with "last-postmask." Although all of the three genomes showed strong colinearity with the guppy NCBI reference genome, we identified an inverted segment on Chromosome 12 (between 0 and ∼9.9 Mb) that appears to have been translocated to ∼10.8 Mb. In order to verify that this rearrangement was not an artifact created during the genome assemblies, we independently aligned the 10x Genomics barcoded reads from each of the three female genomes, representing each river, to the guppy reference genome (Künstner et al. 2016) using Long Ranger v2.2.2. From these read alignments, we identified clear breakpoints in all three female genomes for which the reads surrounding the breakpoints lack the expected pair orientation and/or are soft-clipped at the breakpoint positions (supplementary fig. S1b and c, Supplementary Material online). For consistency and direct comparison between the three rivers, we repeated the assembly scaffolding step above for the Quare and Yarra genomes using the Aripo genome as backbone, as the long scaffolds in the latter were correctly assembled for this inversion. This approach resulted in highly contiguous alignments

without any evidence for an inversion in any of the female genomes.

Preprocessing of Sequencing Reads

We used FastQC v0.11.5 (http://www.bioinformatics.babra-ham.ac.uk/projects/fastqc, last accessed October 4, 2018) to assess read quality and used BBTools v38.34 "bbduk" (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/, last accessed January 25, 2019) to remove adapter sequences, trim regions with average quality scores <Q10 and remove PhiX-174 spike-in control reads. After filtering, we excluded read-pairs from downstream analyses if either read had an average quality score <Q10 or was <35 bp. In order to make both 10x Genomics and Illumina data sets comparable, we also preprocessed 10x Genomics reads as above after first trimming barcodes and performing barcode error correction with the Long Ranger v2.2.2 Basic pipeline.

Read Alignment and Genotype Calling

We aligned all preprocessed reads with BWA v0.7.15-r1140 using the MEM algorithm (Li 2013) and default options and processed the resulting SAM/BAM files with SAMtools v1.9 (Li et al. 2009), flagging duplicated reads with biobambam v2.0.87 (Tischler and Leonard 2014) after alignment.

We called genotypes within each river using Freebayes v1.3.1-16-g85d7bfc (Garrison and Marth 2012) with the following parameters: –min-repeat-entropy 1 –no-partial-observations –use-mapping-quality –min-mapping-quality 3 –min-base-quality 13 –skip-coverage 40,000 –genotype-qualities –strict-vcf. Reads marked as duplicated and secondary alignments are excluded by default in Freebayes. We used "vcfallelicprimitives" from vcflib v09df564 (https://github.com/vcflib/vcflib, last accessed June 19, 2019) to decompose complex variants into canonical SNP and indel representations and Vt v0.5772-60f436c3 (Tan et al. 2015) to normalize the resulting variants. We then used vcflib and BCFtools v1.9 (Li 2011) to perform a series of filters in order to exclude low-quality variants and variants likely arising from copy number variations or paralogous sequences not present in the reference genomes (Li 2014). Specifically, we excluded variant sites with an alternate allele observation on the forward and reverse strands supported by <3 reads, variants with quality <30, heterozygous variants with depth <4 and homozygous variants with quality <50 and depth <4. If heterozygous, a variant was also excluded based on a maximum depth filter if the variant quality was <2× average depth and the maximum genotype depth was higher than the average depth + 3√(average depth) (Li 2014). Indels and single nucleotide polymorphisms (SNPs) within 3 bp of an indel were also excluded as the latter are difficult to ascertain with confidence. Finally, we filtered variants with an allele balance in the tenth percentile, but included fixed variants, and masked individual sample heterozygous genotypes if the corresponding genotype had <3 supporting reads for the reference and alternate allele. For all downstream analyses, we only considered SNP variants with ≤10% missing genotypes.

Degeneration and Divergence Analysis

Using our female genome assemblies, we expect that any regions of Y degeneration will result in female-biased read depth, as highly diverged Y reads will not map to the X chromosome, an approach previously implemented for guppies (Wright et al. 2017; Darolti et al. 2019). We therefore assessed M:F read depth for each population separately by aligning preprocessed Illumina and 10x Genomics reads to the respective female river-specific genome assembly. We calculated per-site coverage with the SAMtools "depth" command after applying stringent filtering criteria to exclude duplicated reads, reads with secondary alignments, nonunique alignments (with the XA or SA tags in the BAM file), and reads with a high number of mismatches to the reference genome (mapping quality <Q10). We restricted coverage calculations to the 23 assigned linkage groups in the guppy genome. We then calculated the effective coverage value as the median per site coverage in nonoverlapping windows of 50 kb. To account for differences in the overall coverage between individuals, we normalized coverage data based on the median genomic coverage of each individual.

For regions of the X and the Y chromosomes that have diverged, but still show little evidence of Y degeneration, we expect an increase in male compared with female diversity, as Y-linked reads will map to the female genome assembly, but will carry male-specific SNPs (Darolti et al. 2019; Palmer et al. 2019). We previously observed elevated male SNP density (Wright et al. 2017) in each of our upstream populations of *P. reticulata* across a large proportion of the sex chromosome compared with downstream populations. To investigate the extent of male to female divergence, we used Hudson's method (Hudson et al. 1992), as implemented in the Python library scikit-allel v1.2.1 (https://scikit-allel.readthe-docs.io/en/stable/index.html, last accessed June 28, 2019), to calculate male:female F_{ST} for biallelic sites in nonoverlapping windows of 50 kb. Windows of 10 kb produced qualitatively similar results (data not shown). Prior to F_{ST} calculation, we excluded singleton sites and sites with ≥20% of missing data in each sex.

Phasing and Analysis of Y Haplotypes

We used the phasing information from linked-reads sequencing of ten males and three females from each population to aid in phasing all our samples. 10x Genomics samples were assembled and phased using the 10x Genomics Long Ranger v.2.1.2 suite with the wgs option and using FreeBayes for variant calling. Briefly, reads that share the same barcode were sequenced from the same original long input DNA molecule, and this information is used to link the reads into large haplotype blocks. This is expected to generate a high-quality phasing for each individual sample. To make the variation data between the 10x Genomics single-sample VCF and the pooled variants from each watershed as close as possible, we regenotyped all individuals using the BAM files generated by Long Ranger. Genotyping was performed with FreeBayes and filtered for SNPs as detailed above.

The full phasing of all genotypes was then performed in two steps. First, we used WhatsHap v0.19.dev156+g1564a9f

(Martin et al. 2016) for read-backed phasing. For the 10x Genomics samples, this simply maps the phase sets from the Long Ranger VCF to the river-specific genotyping VCF. In the case of the Illumina sequencing female samples, WhatsHap runs its full algorithm that uses the sequencing reads to reconstruct haplotypes. Secondly, we used SHAPEIT4 v4.1.2 (Delaneau et al. 2019) to computationally phase all samples. SHAPEIT4 was run to use the phase sets from WhatsHap with the recommended expected error rate of 0.01% and adjusting the default parameters for sequencing data (−use-PS 0.0001 −sequencing). To improve the phasing accuracy, we increased the number of iterations to "10b,1p,2b,1p,2b,1p,2b,1p,10m" and also increased the number of conditioning neighbors to 8 (−pbwt-depth 8).

Y haplotypes can be identified phylogenetically using gene trees from the phased genotypes if they form a single monophyletic clade. For this, we divided the genome into nonoverlapping windows of 100 SNPs and generated outputs in fasta format for the two inferred haplotypes of each individual with the BCFtools v1.9 consensus command. We used SNP windows for better resolution of regions with elevated SNP density. We then converted the fasta alignments at each window to the phylip format and built gene trees using FastME v 2.1.6.1 (Lefort et al. 2015) with the parameters −method=BIONJ −dna=F84 −spr. For a given gene tree, we identified the Y chromosome haplotypes if a clade was composed exclusively of male individuals and included >66% males (all but three males) in each population (downstream and upstream populations). Because we are using river-specific reference genomes, gene trees were computed separately in each river.

Nucleotide diversity was estimated using the Watterson's theta estimator for the male Y haplotypes and the correspondingly alternative haplotypes (inferred as the X haplotypes). To obtain an autosomal diversity estimate, we used a single arbitrarily chosen male haplotype from 1,000 randomly sampled autosomal gene trees (excluding Chromosome 12). To compute haplotype networks, we used the haploNet method with default parameters from the package pegas (Paradis 2010).

Y-Mer Mining

k-Mer refers to all the possible substrings of length k that are contained in a genome, and have been useful in identifying sex-specific (Y chromosome) sequence in a range of organisms (Akagi et al. 2014; Pucholt et al. 2017; Torres et al. 2018), including guppies (Morris et al. 2018) and other Poecilia species (Darolti et al. 2019; Sandkam et al. 2020), by comparing male and female k-mer profiles. Male-specific k-mers, referred to here as Y-mers, most likely represent Y chromosome sequence, and Y-mers can be used to distinguish reads, read pairs, or scaffolds likely to be Y-linked. We used JELLYFISH v2.2.6 (Marçais and Kingsford 2011) to count the number of 21-bp canonical k-mers in the trimmed and filtered preprocessed reads for the male and female samples for each of our six wild guppy populations. We sorted and filtered k-mers, rejecting k-mers with observed counts <3 in any individual, as these are likely sequencing errors. We then combined k-

mer profiles for all samples of the same sex within each population, and identified male-specific k-mers absent from all female samples (Y-mers) with >5 average counts across males.

To identify scaffolds enriched for Y-specific sequence, we used Bowtie v1.2.3, without allowing for mismatches and reporting all alignments (-f -v 0 −all -l 21), to map Y-mers to the Supernova assemblies of each male genome. We used the megabubbles output in Supernova because in this output style Supernova generates an individual FASTA record for each homologous phased haplotype without mixing maternal and paternal alleles in the same sequence. Scaffolds with ≥100 Y-mers and more Y-mers than the homologous haplotype (if identified) were selected for further analysis. To remove possible redundancy in the resulting scaffolds, as the same region could be identified in scaffolds from different males, we clustered scaffolds within watersheds with a sequence identity threshold ≥90% and ≥50% of sequence overlap as calculated from BlastN v2.5.0+ with parameters -dust yes -evalue 0.000001 -max_target_seqs 100,000. The longest scaffold of each cluster with at least five samples, assuming that a Y-linked region could be absent from one sample due to insufficient coverage for assembly, was used for annotation.

Annotation of Y-Linked Scaffolds

Candidate Y-linked scaffolds from the three watersheds were pooled together for annotation. Annotation was performed with MAKER v2.31.10 (Holt and Yandell 2011). We ran the MAKER pipeline twice: first based on a guppy-specific repeat library, protein sequence, EST, and RNA sequence data (later used to train ab initio software) and a second time combining evidence data from the first run and ab initio predictions. We created a repeat library for these scaffolds using de novo repeats identified by RepeatModeler v1.0.10 (http://www.repeatmasker.org, last accessed October 9, 2017) which we then combined with Actinopterygii-specific repeats to use with RepeatMasker v4.0.7 (http://www.repeatmasker.org, last accessed May 3, 2017). Annotated protein sequences were downloaded from Ensembl (release 95) (Howe et al. 2020) for eight fish species: *Danio rerio* (GRCz11), *Gasterosteus aculeatus* (BROADS1), *Oryzias latipes* (ASM223467v1), *P. latipinna* (1.0), *P. mexicana* (1.0), *P. reticulata* (1.0), *Takifugu rubripes* (FUGU5), and *Xiphophorus maculatus* (5.0). For EST, we used 10,664 tags from Dreyer et al. (2007) isolated from guppy embryos and male testis. Furthermore, to support gene predictions, we also used two publicly available libraries of RNA-seq data collected from guppy male testis and male embryos (Sharma et al. 2014) and assembled with StringTie 1.3.3b (Pertea et al. 2015). As basis for the construction of gene models, we combined ab initio predictions from Augustus v3.2.3 (Stanke et al. 2006), trained via BUSCO v3.0.2 (Seppey et al. 2019), and SNAP v2006-07-28 (Korf 2004). To train Augustus and SNAP, we first ran the MAKER pipeline a first time to create a profile using the protein and EST evidence along with RNA-seq data. Both Augustus and SNAP were then trained from this initial evidence-based annotation. Functional inference for genes and transcripts was performed

using the translated CDS features of each coding transcript. Protein sequences were searched with BLAST in the Uniprot/Swissprot reference data set in order to retrieve gene names and protein functions as well as in the InterProscan v5 database to retrieve additional annotations from different sources.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Data Availability

The data underlying this article are available in EBI's ENA at https://www.ebi.ac.uk/ena, and can be accessed with project number PRJEB39998. Scripts and data sets underlying this article will be shared on reasonable request to the corresponding author.

References

Akagi T, Henry IM, Tao R, Comai L. 2014. A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* 346(6209):646–650.

Alexander HJ, Taylor JS, Wu SS, Breden F. 2006. Parallel evolution and vicariance in the guppy (*Poecilia reticulata*) over multiple spatial and temporal scales. *Evolution* 60(11):2352–2369.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20(1):224.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet.* 14(2):113–124.

Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice W, Valenzuela N. 2011. Are all sex chromosomes created equal. *Trends Genet.* 27(9):350–357.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman TL, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it. *PLoS Biol.* 12(7):e1001899.

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

Bergero R, Charlesworth D. 2019. Reply to Wright et al.: how to explain the absence of extensive Y-specific regions in the guppy sex chromosomes. *Proc Natl Acad Sci U S A.* 116(26):12609–12610.

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

Bergero R, Qiu S, Forrest A, Borthwick H, Charlesworth D. 2013. Expansion of the pseudo-autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes. *Genetics* 194(3):673–686.

Bergero R, Gardner J, Bader B, Yong L, Charlesworth D. 2019. Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. *Proc Natl Acad Sci U S A.* 116(14):6924–6931.

Brooks R, Endler JA. 2001. Direct and indirect sexual selection and quantitative genetics of male traits in guppies (*Poecilia reticulata*). *Evolution* 55(5):1002–1015.

Burt A, Bell G, Harvey PH. 1991. Sex differences in recombination. *J Evol Biol.* 4(2):259–277.

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

Chibalina MV, Filatov DA. 2011. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol.* 21(17):1475–1479.

Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, Warren RL. 2018. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* 19(1):234.

Darolti I, Wright AE, Mank JE. 2020. Guppy Y chromosome integrity maintained by incomplete recombination suppression. *Genome Biol Evol.* 12(6):965–977.

Darolti I, Wright AE, Sandkam BA, Morris J, Bloch NI, Farré M, Fuller RC, Bourne GR, Larkin DM, Breden F, et al. 2019. Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proc Natl Acad Sci U S A.* 116(38):19031–19036.

Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat Commun.* 10(1):5436.

Dor L, Shirak A, Kohn YY, Gur T, Weller JI, Zilberg D, Seroussi E, Ron M. 2019. Mapping of the sex determining region on linkage group 12 of guppy (*Poecilia reticulata*.). *G3 (Bethesda)* 9:3867–3875.

Dreyer C, Hoffmann M, Lanz C, Willing EM, Riester M, Warthmann N, Sprecher A, Tripathi N, Henz SR, Weigel D. 2007. ESTs and EST-linked polymorphisms for genetic mapping and phylogenetic reconstruction in the guppy, *Poecilia reticulata*. *BMC Genomics* 8:269.

Endler JA. 1988. Frequency-dependent predation, crypsis and aposematic coloration. *Philos Trans R Soc Lond B Biol Sci.* 319(1196):505–523.

Endler JA. 1995. Multiple-trait coevolution and environmental gradients in guppies. *Trends Ecol Evol.* 10(1):22–29.

Farr JA. 1977. Male rarity or novelty, female choice behaviour, and sexual selection in the guppy, *Poecilia reticulata* Peters (Pisces: Poeciliidae). *Evolution* 31(1):162–168.

Fisher RA. 1931. The evolution of dominance. *Genetics* 6(4):345–368.

Fraser BA, Künstner A, Reznick DN, Dreyer C, Weigel D. 2015. Population genomics of natural and experimental populations of guppies (*Poecilia reticulata*). *Mol Ecol.* 24(2):389–408.

Furman BLS, Metzger DCH, Darolti I, Wright AE, Sandkam BA, Almeida P, Shu JJ, Mank JE. 2020. Sex chromosome evolution: so many exceptions to the rules. *Genome Biol Evol.* 12(6):750–763.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv 1207.3907.

Gilbert KJ, Pouyet F, Excoffier L, Peischl S. 2020. Transition from background selection to associative overdominance promotes diversity in regions of low recombination. *Curr Biol.* 30(1):101–107.e3.

Gordon SP, López-Sepulcre A, Reznick DN. 2012. Predation-associated differences in sex linkage of wild guppy coloration. *Evolution* 66(3):912–918.

Gordon SP, López-Sepulcre A, Rumbo D, Reznick DN. 2017. Rapid changes in the sex linkage of male coloration in introduced guppy populations. *Am Nat.* 189(2):196–200.

Gorelick R. 2003. Evolution of dioecy and sex chromosomes via methylation driving Muller's ratchet. *Biol J Linn Soc.* 80(2):353–368.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.

Haskins CP, Haskins EF. 1951. The inheritance of certain color patterns in wild populations of *Lebistes reticulatus* in Trinidad. *Evolution* 5(3):216–225.

Haskins CP, Haskins EF. 1961. Polymorphisms and population structure in *Lebistes reticulatus*, an ecological study. In: Blair WF, editor.

Vertebrate speciation. Austin (TX): University of Texas Press. p. 320–395.

Haskins CP, Young P, Hewitt RE, Haskins EF. 1970. Stabilised heterozygosis of supergenes mediating certain Y-linked colour patterns in populations of *Lebistes reticulatus*. *Heredity* 25(4):575–589.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12(1):491.

Houde AE, Endler JA. 1990. Correlated evolution of female mating preferences and male color patterns in the guppy *Poecilia reticulata*. *Science* 248(4961):1405–1408.

Houde AE. 1997. Sex, color, and mate choice in guppies. Princeton (NJ): Princeton University Press.

Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, Alvarez-Jarreta J, Barba M, Bolser DM, Cambell L, et al. 2020. Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48(D1):D689–D695.

Hudson RR, Boos DD, Kaplan NL. 1992. A statistical test for detecting geographic subdivision. *Mol Biol Evol.* 9(1):138–151.

Hughes KA, Du L, Rodd FH, Reznick DN. 1999. Familiarity leads to female mate preference for novel males in the guppy, *Poecilia reticulata*. *Anim Behav.* 58(4):907–916.

Hughes KA, Houde AE, Price AC, Rodd FH. 2013. Mating advantage for rare males in wild guppy populations. *Nature* 503(7474):108–110.

Kemp DJ, Batistic FK, Reznick DN. 2018. Predictable adaptive trajectories of sexual coloration in the wild: evidence from replicate experimental guppy populations. *Evolution* 72(11):2462–2477.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* 21(3):487–493.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5(1):59.

Künstner A, Hoffmann M, Fraser BA, Kottler VA, Sharma E, Weigel D, Dreyer C. 2016. The genome of the Trinidadian guppy, *Poecilia reticulata*, and variation in the Guanapo population. *PLoS One* 11(12):e0169087.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* 286(5441):964–967.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol.* 32(10):2798–2800.

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

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

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

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30(20):2843–2851.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Lindholm A, Breden F. 2002. Sex chromosomes and sexual selection in Poeciliid fishes. *Am Nat.* 160(S6):S214–S224.

Lisachov AP, Zadesenets KS, Rubtsov NB, Borodin PM. 2015. Sex chromosome synapsis and recombination in male guppies. *Zebrafish* 12(2):174–180.

Lloyd A, Jenczewski E. 2019. Modelling sex-specific crossover patterning in *Arabidopsis*. *Genetics* 211(3):847–859.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.

Martin M, Patterson M, Garg S, O Fischer S, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. WhatsHap: fast and accurate read-based phasing. bioRxiv 10.1101/085050

Meredith RW, Pires MN, Reznick DN, Springer MS. 2010. Molecular phylogenetic relationships and the evolution of the placenta in *Poecilia* (*Micropoecilia*) (Poeciliidae: Cyprinodontiformes). *Mol Phylogenet Evol.* 55(2):631–639.

Metzger DCH, Mank JE. 2020. Conserved sex-biased DNA methylation patterns target key developmental genes and non-recombining region of the guppy sex chromosome. bioRxiv 2020.08.21.261792. .

Morris J, Darolti I, Bloch NI, Wright AE, Mank JE. 2018. Shared and species-specific patterns of nascent Y chromosome evolution in two guppy species. *Genes* 9(5):238.

Nanda I, Schartl M, Epplen JT, Feichtinger W, Schmid M. 1993. Primitive sex chromosomes in poeciliid fishes harbor simple repetitive DNA sequences. *J Exp Zool.* 265(3):301–308.

Nanda I, Schartl M, Feichtinger W, Epplen JT, Schmid M. 1992. Early stages of sex chromosome differentiation in fish as analysed by simple repetitive DNA sequences. *Chromosoma* 101(5–6):301–310.

Nanda I, Schories S, Tripathi N, Dreyer C, Haaf T, Schmid M, Schartl M. 2014. Sex chromosome polymorphism in guppies. *Chromosoma* 123(4):373–383.

Natri HM, Shikano T, Merilä J. 2013. Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Mol Biol Evol.* 30(5):1131–1144.

Nayudu PL. 1979. Genetic studies of melanic color patterns and atypical sex determination in the guppy, *Poecilia reticulata*. *Copeia* 1979(2):225–231.

Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Monéger F. 2004. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS Biol.* 3(1):e4.

Olendorf R, Rodd FH, Punzalan D, Houde AE, Hurt C, Reznick DN, Hughes KA. 2006. Frequency-dependent survival in natural guppy populations. *Nature* 441(7093):633–636.

Palmer DH, Rogers TF, Dean R, Wright AE. 2019. How to identify sex chromosomes and their turnover. *Mol Ecol.* 28(21):4709–4724.

Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26(3):419–420.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 33(3):290–295.

Pucholt P, Wright AE, Conze LL, Mank JE, Berlin S. 2017. Recent sex chromosome divergence despite ancient dioecy in the willow *Salix viminalis*. *Mol Biol Evol.* 34(8):1991–2001.

Reznick D, Bryga H, Endler JA. 1990. Experimentally induced life-history evolution in a natural population. *Nature* 346(6282):357–359.

Sandkam B, Young CM, Breden F. 2015. Beauty in the eyes of the beholders: colour vision is tuned to mate preference in the Trinidadian guppy (*Poecilia reticulata*). *Mol Ecol.* 24(3):596–609.

Sandkam BA, Almeida P, Darolti I, Furman B, van der Bijl W, Morris J, Bourne G, Breden F, Mank JE. 2020. Extreme Y chromosome polymorphism corresponds to five male reproductive morphs. bioRxiv 2020.08.19.258434.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol.* 1962:227–245.

Sharma E, Künstner A, Fraser BA, Zipprich G, Kottler VA, Henz SR, Weigel D, Dreyer C. 2014. Transcriptome assemblies for studying sex-biased gene expression in the guppy, *Poecilia reticulata*. *BMC Genomics* 15:400.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.* 23(1):23–35.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34(Web Server):W435–W439.

Suk HY, Neff BD. 2009. Microsatellite genetic differentiation among populations of the Trinidadian guppy. *Heredity* 102(5):425–434.

Sun Y, Svedberg J, Hiltunen M, Corcoran P, Johannesson H. 2017. Large-scale suppression of recombination predates genomic rearrangements in *Neurospora tetrasperma*. *Nat Commun.* 8(1):1140.

Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinformatics* 31(13):2202–2204.

Tischler G, Leonard S. 2014. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol Med.* 9(1):2078.

Torres MF, Mathew LS, Ahmed I, Al-Azwani IK, Krueger R, Rivera-Nuñez D, Mohamoud YA, Clark AG, Suhre K, Malek JA. 2018. Genus-wide sequencing supports a two-locus model for sex-determination in *Phoenix. Nat Commun.* 9(1):3969.

Traut W, Winking H. 2001. Meiotic chromosomes and stages of sex chromosome evolution in fish: zebrafish, platyfish and guppy. *Chromosome Res.* 9(8):659–672.

Tripathi N, Hoffmann M, Weigel D, Dreyer C. 2009. Linkage analysis reveals the independent origin of Poeciliid sex chromosomes and a case of atypical sex inheritance in the guppy (*Poecilia reticulata*). *Genetics* 182(1):365–374.

Tripathi N, Hoffmann M, Willing EM, Lanz C, Weigel D, Dreyer C. 2009. Genetic linkage map of the guppy, *Poecilia reticulata*, and quantitative trait loci analysis of male size and colour variation. *Proc Biol Sci.* 276(1665):2195–2208.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* 499(7458):332–335.

Vicoso B, Bachtrog D. 2015. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol.* 13(4):e1002078.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biol.* 11(8):e1001643.

Wallberg A, Glémin S, Webster MT. 2015. Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera. PLoS Genet.* 11(4):e1005189.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I. 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* 4:35.

Winge Ø. 1922. One-sided masculine and sex-linked inheritance in *Lebistes reticulatus. J Genet.* 12:137–144.

Winge Ö. 1927. The location of eighteen genes in *Lebistes reticulatus. J Genet.* 18(1):1–43.

Winge Ø, Ditlevsen E. 1947. Colour inheritance and sex determination in *Lebistes. Genetics* 1(1):65–83.

Wright AE, Darolti I, Bloch NI, Oostra V, Sandkam B, Buechel SD, Kolm N, Breden F, Vicoso B, Mank JE. 2017. Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nat Commun.* 8(1):14251.

Wright AE, Darolti I, Bloch NI, Oostra V, Sandkam BA, Buechel SD, Kolm N, Breden F, Vicoso B, Mank JE. 2019. On the power to detect rare recombination events. *Proc Natl Acad Sci U S A.* 116(26):12607–12608.

Wright AE, Dean R, Zimmer F, Mank JE. 2016. How to make a sex chromosome. *Nat Commun.* 7(1):12087.

Yamamoto T. 1975. The Medaka, *Oryzias latipes*, and the Guppy, *Lebistes reticularis*. In: King RC, editor. Handbook of genetics. Boston: Springer. p. 133–149.

Zhang W, Wang X, Yu Q, Ming R, Jiang J. 2008. DNA methylation and heterochromatinization in the male-specific region of the primitive Y chromosome of papaya. *Genome Res.* 18(12):1938–1943.