

1 **A comparison of methodological approaches to the study of young sex chromosomes: A case study in**
2 ***Poecilia***

3 Iulia Darolti¹, Pedro Almeida², Alison E. Wright³, Judith E. Mank^{1,4}

4 1 Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver,
5 Canada

6 2 Department of Genetics, Evolution and Environment, University College London, London, United
7 Kingdom

8 3 Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Sheffield, United
9 Kingdom

10 4 Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter,
11 Cornwall, United Kingdom

12 Co-corresponding author: Judith E. Mank, e-mail: mank@zoology.ubc.ca

13 Co-corresponding author: Iulia Darolti, e-mail: darolti@zoology.ubc.ca

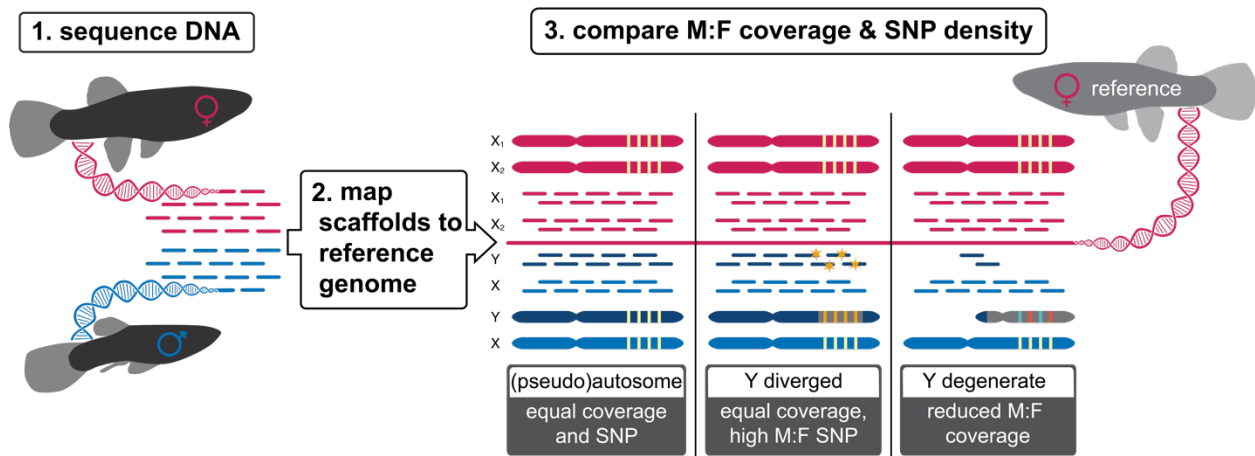
14 **Abstract**

15 Studies of sex chromosome systems at early stages of divergence are key to understanding the initial
16 process and underlying causes of recombination suppression. However, identifying signatures of
17 divergence in homomorphic sex chromosomes can be challenging due to high levels of sequence similarity
18 between the X and the Y. Variations in methodological precision and underlying data can make all the
19 difference between detecting subtle divergence patterns or missing them entirely. Recent efforts to test
20 for X-Y sequence differentiation in the guppy have led to contradictory results. Here we apply different
21 analytical methodologies to the same dataset to test for the accuracy of different approaches in
22 identifying patterns of sex chromosome divergence in the guppy. Our comparative analysis reveals that
23 the most substantial source of variation in the results of the different analyses lies in the reference
24 genome used. Analyses using custom-made *de novo* genome assemblies for the focal species successfully
25 recover a signal of divergence across different methodological approaches. By contrast, using the distantly
26 related *Xiphophorus* reference genome results in variable patterns, due to both sequence evolution and
27 structural variations on the sex chromosomes between the guppy and *Xiphophorus*. Changes in mapping
28 and filtering parameters can additionally introduce noise and obscure the signal. Our results illustrate how
29 analytical differences can alter perceived results and we highlight best practices for the study of nascent
30 sex chromosomes.

31 **Introduction**

32 Substantial recent attention in sex chromosome research has focused on the earliest stages of X-Y
33 divergence in order to glean the initial processes of recombination suppression (Wright et al. 2016).
34 Studies of nascent sex chromosome divergence will by definition result in subtle patterns of X-Y sequence
35 differentiation as substantial differences have not yet sufficiently accumulated. Given the expected
36 subtlety, methodology and underlying data can be quite important, and small changes may make all the
37 difference between identifying a delicate pattern or missing it entirely.

38 For example, several recent tests for divergence between the guppy X and Y chromosomes have revealed
39 contradictory results. Full genomic analysis of the *Poecilia reticulata* sex chromosomes was originally
40 presented in Wright et al. (2017) based on comparisons between male and female genomes (Fig. 1). This
41 approach can be used to identify what, if any, regions of the Y chromosome are diverged from the X, and
42 to compare across populations to determine intra-specific variation. Wright et al. (2017) found a relatively
43 small region (10Mb) of significant Y degeneration, designated Stratum I. This region was characterized by
44 a reduction in the number of male reads that mapped compared to females, consistent with the concept
45 of Y degeneration. Moreover, the same pattern was observed in all six of the natural populations assayed
46 as well as a captive lab population, and the rules of parsimony therefore suggest that Stratum I is ancestral
47 to the colonization of Trinidad. Wright et al. (2017) also observed evidence of a second region of nascent
48 divergence, Stratum II, that appeared to have formed independently in three upstream populations, but
49 was smaller in downstream populations. This region was characterized by an increase in male single
50 nucleotide polymorphism (SNP) density compared to females but no degradation of the Y. This pattern is
51 consistent with either greatly reduced or complete loss of male recombination in this region, or selection
52 against recombinant males.



53

54 **Figure 1. Genomic comparisons of male and female DNA data can be used to identify X-Y divergence.**

55 Step (1) Multiple males and females are sequenced, and female reads (red) are assembled, with resulting
56 scaffolds ordered and oriented to the nearest available full reference genome. Step (2) Male (blue) and
57 female (red) reads are mapped to this assembly. Step (3) Y divergence leads to male-specific SNPs, and
58 therefore elevated male:female SNP density. As the Y degenerates, Y reads will no longer map to the X
59 chromosome assembly, leading to reduced male:female coverage. Method adapted from Vicoso &
60 Bachtrog 2013 and Vicoso and Bachtrog 2015. Figure courtesy of Jacelyn Shu (jacelyndesigns.com).

61 Almeida et al. (2021) built on these initial findings with a greatly expanded dataset, again recovering
62 concordant patterns of Stratum I across the same six natural populations of *P. reticulata*. The expanded
63 dataset incorporated 10X genomics linked-reads, allowing for far more sophisticated analyses. Namely by
64 phasing X and Y haplotypes, it was possible to discern that Stratum I is comprised of two smaller separate
65 regions of reduced male:female read depth. This region is also enriched for male-specific sequences, male-
66 specific SNPs, and repetitive elements, the presence of which necessitate recombination suppression
67 from the X chromosome. Importantly, there was also evidence of phylogenetic clustering of phased Y
68 sequence in this region, indicating ancestral recombination suppression. Finally, this replication recovered
69 evidence of parallel expansion of Stratum II in upstream populations.

70 Expanding phylogenetically, Darolti et al (2019) uncovered consistent patterns of sex chromosome
71 topology in *P. wingei*. Initial karyotype analysis suggested that the X and Y chromosomes are ancestral to
72 the common guppy (*P. reticulata*) and Endler's Guppy (*P. wingei*) (Nanda et al. 1993). Furthermore, Darolti
73 et al. (2019) found the same small region of Y chromosome degeneration consistent with Stratum I,
74 although somewhat more pronounced in the degree of divergence from the X than *P. reticulata*. The
75 region of degeneration matched nearly perfectly with *P. reticulata*, suggesting Stratum I was in fact
76 present in the common ancestor of *P. wingei* and *P. reticulata*. Consistent with this, Morris et al (2018)
77 found evidence of male-specific sequence shared between *P. reticulata* and *P. wingei*, possible only if
78 recombination between the X and Y was halted in the common ancestor of these species. Moreover,
79 Darolti et al. (2020) used SNP segregation patterns from RNA-seq data across pedigrees to determine X
80 and Y sequence, and found four genes that showed phylogenetic evidence of recombination suppression
81 in the ancestor of *P. wingei* and *P. reticulata*. Although the bootstrap values for any one locus were not
82 excessively high, it is telling that all four were in Stratum I. The ancestral origin of Stratum I was further
83 supported by conserved patterns of male-hypomethylation within this region in both species (Metzger et
84 al. 2020), consistent with sexualization of gene regulation. Finally, Darolti et al. (2019) found evidence for
85 another independent origin of Stratum II based on SNP data in *P. wingei*. Work in outgroup species
86 revealed the same chromosome is a sex chromosome in *P. picta* and *P. parae* (Darolti et al. 2019; Sandkam
87 et al. 2021), although diverged to a far greater degree in both these species.

88 Crucially, all of these analyses were based on custom genome or transcriptome assemblies generated
89 bespoke from the underlying data (Wright et al. 2017; Darolti et al. 2019; Darolti et al. 2020; Almeida et
90 al. 2021), although they did use existing related reference genomes to physically place and orient
91 scaffolds. This is in contrast to other studies which have used existing resources derived from different
92 populations or species, resulting in potential mismatches between the underlying data and the genome
93 to which it is compared. Taking a bespoke approach is critical as it reduces the phylogenetic distance
94 between the sequence reads and the reference to which they are mapped, which can increase the
95 proportion of reads that are accurately mapped and reduce issues arising from structural variation and
96 repetitive sequence. Secondly, an important step in identifying diverged regions in sex chromosomes is
97 ensuring stringent mapping parameters (Carvalho and Clark 2013; Smeds et al. 2015; Vicoso and
98 Bachtrog 2013; Vicoso and Bachtrog 2015; Palmer et al. 2019). This is particularly relevant for
99 homomorphic sex chromosomes as they still retain sequence orthology between the X and Y, and
100 incorrectly mapped reads can mask coverage differences between the sexes and lead to the
101 misclassification of sex-linked sequences as autosomal. Wright et al. (2017), Darolti et al. (2019) and
102 Almeida et al. (2021) used stringent mapping limits, removed minor alleles with low frequency, which
103 likely represent sequencing errors, and focused on coding sequence to minimize issues with repetitive
104 elements (Table 1). This was based on the reasoning that young sex chromosomes would exhibit subtle
105 divergence signatures, and stringency would be required to detect it (Palmer et al. 2019; Vicoso and
106 Bachtrog 2015).

107

Table 1. A comparison of methods and findings for *P. reticulata* sex chromosome strata

Study	Evidence for Stratum I from M:F read depth	Phylogenetic clustering of Y sequences in Stratum I	Evidence for Stratum II from M:F F_{ST}	Genome Assembly	Read depth analysis	SNP analysis
Wright et al. (2017)	Yes	n/a	Yes	Bespoke (<i>P. reticulata</i>)	Uniquely mapping reads	Limited to coding sequence; Site coverage > 10; SNP frequency > 0.3 x site coverage
Bergero et al. (2019)	No	n/a	Yes	Kunstner et al. (2016) (<i>P. reticulata</i>)	Default mismatch parameters, duplicate reads excluded	Quality score > 30; Minimum coverage 20; Biallelic SNPs
Darolti et al. (2019)	Yes	n/a	Yes	Bespoke (<i>P. reticulata</i> and <i>P. wingei</i>)	Uniquely mapping read	Limited to coding sequence; Site coverage > 10; SNP frequency > 0.3 x site coverage
Charlesworth et al. (2020)	No	n/a	Yes	Kunstner et al. (2016) (<i>P. reticulata</i>)	Not reported*	Not reported*
Fraser et al. (2020)	Yes	n/a	No	Bespoke (<i>P. reticulata</i>)	Default mismatch parameters	MAF > 0.05
Kirkpatrick et al. (2020)	No	Yes	No	Schartl et al. (2013) (<i>Xiphophorus maculatus</i>)	Default read mapping parameters with local argument, duplicate reads excluded	Quality score > 20; Minimum read depth 3; Biallelic SNPs
Almeida et al. (2021)	Yes	Yes	Yes	Bespoke (<i>P. reticulata</i> , river-specific)	Uniquely mapping reads, duplicate reads excluded	MAF > 0.1; excluding extremely high coverage sites

*Mapping criteria not reported in methods and bioinformatic code not publicly available. Defaults assumed. MAF = minor allele frequency.

108

109 Despite observing remarkable concordance of these patterns across multiple datasets, species and
 110 analytical methods, other recent studies have differed substantially in their approach and reported some
 111 different results (Table 1). For example, Bergero et al. (2019) did not report evidence for Stratum I in their
 112 own *P. reticulata* data, and although they did uncover a pattern that is broadly consistent with Stratum II,
 113 it was not statistically different across populations (Charlesworth et al. 2020). Fraser et al. (2020)
 114 identified small male-specific regions largely consistent with the regions identified by Almeida et al.
 115 (2021), although because of scaffold orientation differences and population-specific inversions, they are

116 in different physical locations. Finally, Kirkpatrick et al. (2020), reanalyzing data from Darolti et al. (2019)
117 found evidence of Stratum I and Stratum II in *P. wingei*, but not in *P. reticulata*. Notably, they did find
118 phylogenetic evidence of recombination suppression in the ancestor of these two species in Stratum I,
119 consistent with Darolti et al. (2020) and Almeida et al. (2021).

120 Importantly Bergero et al. (2019), Charlesworth et al. (2020) and Kirkpatrick et al. (2020) relied on existing
121 reference genomes for all their analyses and did not use genomic reads to build custom-made assemblies
122 for the target species. Importantly, Kirkpatrick et al. (2020) mapped reads from *P. wingei* and *P. reticulata*
123 to *Xiphophorus*, which last shared a common ancestor 40 mya (Kumar et al. 2017). Additionally, Bergero
124 et al. (2019) used less stringent mapping criteria, and Charlesworth et al. (2020) and Fraser et al. (2020)
125 used default mapping parameters. Together, this produces two sources of potential methodological noise
126 in replication efforts. First, noise can arise from accumulated mutations due to phylogenetic distance
127 between the samples used to generate sequence reads and the genome that they are mapped to. Second,
128 permissive default mapping parameters allow for mismapping, and therefore potentially result in
129 significant noise in genomic comparisons between males and females. In addition to this, many of these
130 studies used different underlying datasets that varied in sample origin, number and read depth, and so it
131 is difficult to distinguish the role of sample variation from methodological differences in these
132 discrepancies.

133 The proliferation of studies on this system with different levels of analytical sophistication allow for a
134 remarkable comparison of the role of genomic methodology in pattern discovery. We tested various
135 methods on the same underlying data with the goal of determining the methodological reasons for
136 inconsistent findings across these studies, and to develop best practices moving forward to the genomic
137 study of nascent sex chromosome systems.

138 **Methods**

139 *Datasets*

140 Using the *P. reticulata* data from Almeida et al. (2021) and the *P. wingei* data from Darolti et al. (2019),
141 we ran multiple analyses of guppy sex chromosome evolution, following the various analytical methods
142 used by Wright et al. (2017), Bergero et al. (2019) and Kirkpatrick et al. (2020), as summarized in Table 1
143 and detailed below. We were unable to include the methodology of Charlesworth et al. (2020) as mapping
144 criteria was not reported in methods and bioinformatic code is not publicly available. The datasets for *P.*
145 *reticulata* and *P. wingei* included paired-end DNA-seq reads from three males and three females from the

146 Quare upstream population (EBI ENA under BioProject PRJEB39998) and from our lab population (NCBI
147 SRA under BioProject PRJNA528814), respectively. We assessed read quality using FastQC v0.11.9
148 (www.bioinformatics.babraham.ac.uk/projects/fastqc/, last accessed 8 November 2021), trimmed using
149 Trimmomatic v0.36 (Bolger et al. 2014) and concatenated reads as in Darolti et al. (2019) and Almeida et
150 al. (2021). To replicate previous studies, all analyses were repeated using several different genomes and
151 their respective gene annotations, which included the *P. reticulata* Quare *de novo* genome assembly from
152 Almeida et al. (2021), the *P. reticulata* reference genome from Kunstner et al. (2016) (NCBI accession
153 GCF_000633615.1), the *P. wingei* *de novo* genome assembly from Darolti et al. (2019) and the *Xiphophorus*
154 *maculatus* reference genome from Scharl et al. (2013) (NCBI accession GCF_002775205.2, v5.0).

155 *Coverage analysis*

156 For each focal species, we used three separate methodological pipelines to map and filter reads and to
157 estimate read depth. The first method followed the analysis in Wright et al. (2017), which used bwa
158 v0.7.15 aln/sampe (Li and Durbin 2009) to map reads, removed reads that were not uniquely mapping
159 and estimated coverage with soap.coverage v2.7.7 (<http://soap.genomics.org.cn>, last accessed 1 April
160 2019). The second method followed the pipeline in Kirkpatrick et al. (2020), which mapped reads using
161 bowtie2 v2.2.9 with default parameters and the -local argument (Langmead and Salzberg 2012), removed
162 PCR duplicates using Picard v2.0.1 (<http://broadinstitute.github.io/picard>, last accessed 8 November
163 2021) and calculated coverage with BEDtools v2.26 (Quinlan and Hall 2010). Lastly, the third method
164 followed the analysis in Bergero et al. (2019), which mapped reads with bwa mem and the -M argument
165 (Li and Durbin 2009), removed PCR duplicates with BEDtools (Quinlan and Hall 2010) and estimated
166 coverage using SAMtools v1.3.1 (Li et al. 2009).

167 For all three methodological pipelines, average coverage values were calculated separately for males and
168 females, and average male:female coverage for each non-overlapping window was calculated as
169 $\log_2(\text{average male coverage}) - \log_2(\text{average female coverage})$. A window size of 50kb was used for all *P.*
170 *reticulata* analyses and *P. wingei* analyses based on the *X. maculatus* genome, while 10kb windows were
171 used for *P. wingei* analyses using the more fragmented *de novo P. wingei* genome. Moving averages of
172 coverage were plotted in R v4.0.5 (R Core Team 2019) based on sliding window analyses using the
173 *roll_mean* function. Ninety-five percent confidence intervals for the moving average plots were obtained
174 by randomly sampling autosomal values 1,000 times without replacement.

175 *SNP density analysis*

176 To further assess patterns of Y divergence, for both *P. reticulata* and *P. wingei*, we compared three
177 methodological approaches of estimating SNP density differences between males and females.

178 First, based on Wright et al. (2017), we mapped reads to each genome using bowtie2 with default
179 parameters (Langmead and Salzberg 2012). After file sorting, we used bow2pro v0.1
180 (<http://guanine.evolbio.mpg.de>, last accessed 8 November 2021) to generate a profile for each sample,
181 representing counts for each of the four nucleotide bases at each site. We then applied a minimum site
182 coverage threshold of 10 and kept SNPs with a frequency of 0.3 times the site coverage. We further used
183 gene annotation information to remove SNPs from the analysis if they were not located within coding
184 sequences. For each sample, we calculated average SNP density for each gene as the sum of all SNPs
185 divided by the sum of filtered sites in that gene, excluding those with zero filtered sites.

186 Second, following Kirkpatrick et al. (2020), we called variants from files previously filtered for PCR
187 duplicates (see *Coverage analysis* section above) using BCFtools v.1.3.1 (Li 2011). We then filtered variants
188 using VCFtools v0.1.12b (Danecek et al. 2011), removing indels and variants with a quality score lower
189 than 20, and selecting for biallelic SNPs and a minimum read depth of 3. For each sample, we then used
190 BEDtools counts to count the number of SNPs within 50kb windows across the genome.

191 Third, we used the pipeline in Bergero et al. (2019) to call SNPs from the PCR duplicates filtered files (see
192 *Coverage analysis* section above) using GATK HaplotypeCaller v4.1.9 (Poplin et al. 2017) with the
193 parameters --emit-ref-confidence GVCF and -stand-call-conf 30. Further genotyping was done with GATK
194 GenotypeGVCFs with default parameters and SelectVariants to keep SNPs with a minimum coverage of
195 20, minimum quality of 30 and selecting for biallelic SNPs only. For each sample, we then used BEDtools
196 counts to count the number of SNPs within 50kb windows across the genome.

197 Lastly, in each of these three methodological approaches, average SNP density across all males and across
198 all females was calculated separately. For each gene or window, we calculated male:female SNP density
199 as $\log_2(\text{average male SNP density}) - \log_2(\text{average female SNP density})$. We then divided male:female SNP
200 density estimates into autosomal and sex-linked based on chromosomal position. The distributions of
201 male:female SNP density for the autosomes and the sex chromosomes were plotted in R (R Core Team
202 2019) and differences between them were tested using Wilcoxon rank sum tests.

203 *Pairwise synteny analyses*

204 We used LAST v1256 (Kielbasa et al. 2011) to perform pairwise synteny analyses between the *P. reticulata*
205 sex chromosome (chromosome 12) from the reference genome (Kunstner et al. 2016), the *P. reticulata*
206 sex chromosome from the Quare de novo assembly (Almeida et al. 2021) and the *X. maculatus* syntenic
207 chromosome 8. For alignments involving the *X. maculatus* sequence, we used LAST with the HOXD70
208 seeding scheme designed for a higher rate of substitution, whereas for alignments involving *P. reticulata*
209 sequences only we used the uNEAR seeding scheme for aligning sequences with lower rate of
210 substitutions.

211 **Results**

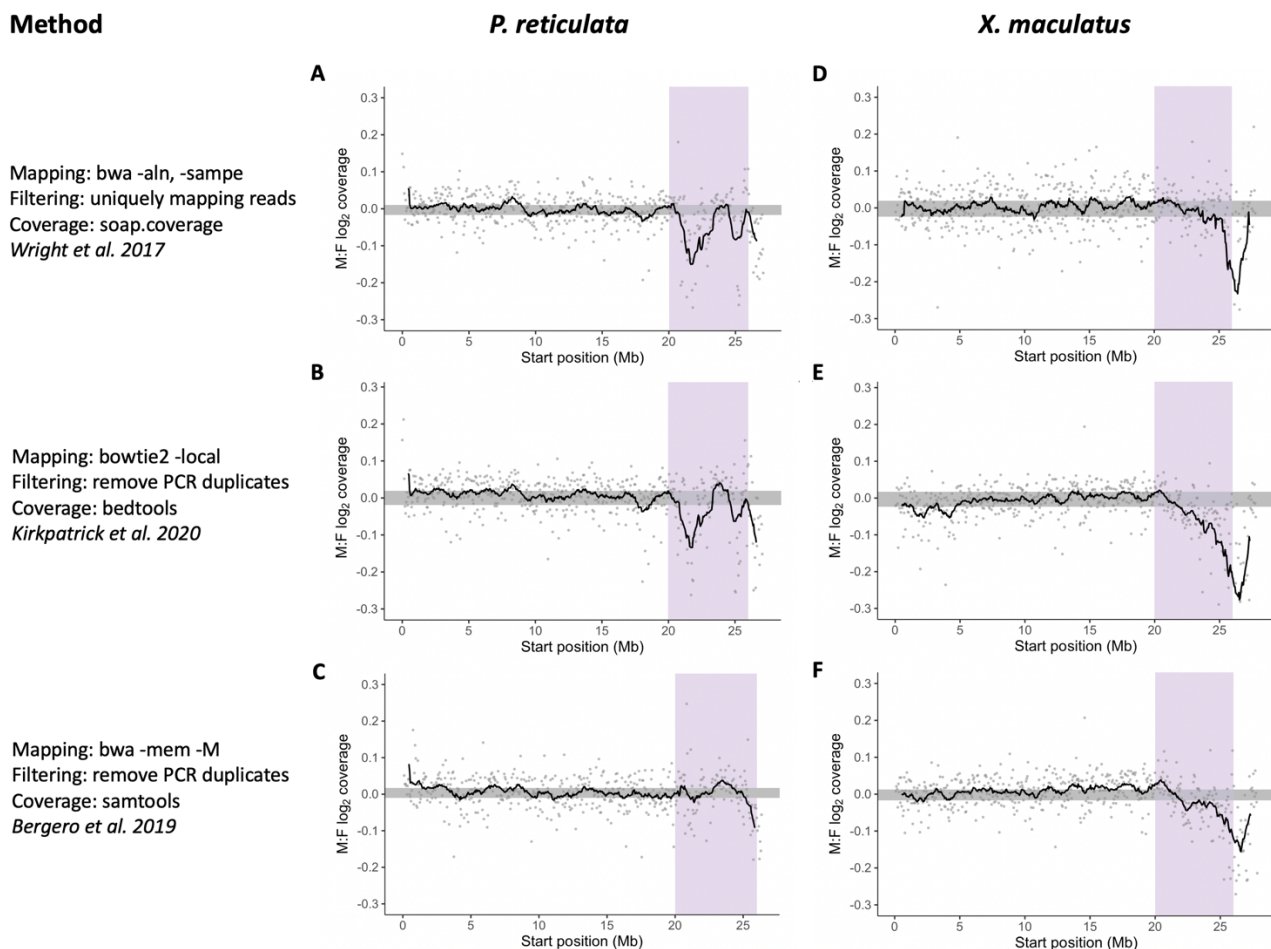
212 Using the same dataset across different genomes and methods, we first assessed the role of various
213 genomic analysis parameters (Table 1) in detecting Stratum I on the *P. reticulata* and *P. wingei* sex
214 chromosomes, previously reported in Wright et al. (2017), Darolti et al. (2019) and Almeida et al. (2021),
215 summarized in Fig. 2 and Fig. 3.

216 Analyses on *P. reticulata* sequencing data that used the custom-made *de novo P. reticulata* genome
217 assembly show a significantly lower male to female coverage, indicative of X-Y degeneration, at the distal
218 end of the chromosome, in the previously estimated location of Stratum I (Fig. 2A, B). This pattern is
219 evident from both the analysis that followed the pipeline from Wright et al. (2017) (Fig. 2A) and the
220 analysis based on the methodology in Kirkpatrick et al. (2020) (Fig. 2B). All three analyses that relied on
221 the *X. maculatus* reference genome also show a region with decreased male coverage relative to that in
222 females, however, this region is shifted closer to the end of the chromosome and only partially overlaps
223 with the syntenic region of the estimated location of *P. reticulata* Stratum I (Fig. 2D, E, F). Pairwise
224 alignments revealed several structural rearrangements between the *P. reticulata* sex chromosome
225 (chromosome 12) and the syntenic *X. maculatus* chromosome 8, particularly in the region of the predicted
226 guppy Stratum I (Fig. 4), which may explain the shifted position of the region with low male coverage in
227 analyses that use the *X. maculatus* genome. In addition, different methodological parameters can have a
228 significant impact on the proportion of reads mapped. Mapping efficiency is substantially reduced when
229 using the *X. maculatus* reference (Table 2), which decreases power to detect a signal of X-Y differentiation.

230 We find no clear pattern of Stratum I when mapping reads to the *P. reticulata* reference genome based
231 on the methodology in Bergero et al. (2019) (Fig. 2C). While we cannot disentangle between the reference
232 genome used and the methodology in this analysis, our other data suggests that the absence of a Stratum
233 I signal is largely due to the *P. reticulata* reference genome. Specifically, when mapping *P. reticulata* reads

234 to the *X. maculatus* genome we recover qualitatively the same coverage pattern across all three
 235 methodological approaches (Fig. 2D, E, F). Similarly, our *P. wingei* analyses detailed below reveal that,
 236 when using the same genome, the Bergero et al. (2019) pipeline produces very similar patterns to the
 237 other two methodological approaches (Fig. 3). Previous work has reported several inversions and
 238 assembly errors on the sex chromosome of the first draft of the *P. reticulata* reference genome (Bergero
 239 et al. 2019; Darolti et al. 2020; Charlesworth et al. 2020; Fraser et al. 2020), which may be obscuring a
 240 signal of Stratum I.

Method



241

242 **Figure 2. Signal for *P. reticulata* Stratum I using comparative methodological approaches.** *P. reticulata* DNA-seq
 243 reads were mapped in turn to a *P. reticulata* genome assembly and the *X. maculatus* reference genome assembly
 244 (Schartl et al. 2013). For replicating previous studies, the *P. reticulata* reference genome from Kunstner et al. (2016)
 245 was used in the analysis based on the methods from Bergero et al. (2019), while the high quality Quare *de novo*
 246 assembly from Almeida et al. (2021) was used in the other two analyses. Moving average plots represent male to
 247 female coverage differences across the guppy sex chromosome (*P. reticulata* chromosome 12, and syntenic *X.*
 248 *maculatus* chromosome 8) in non-overlapping windows of 50kb. 95% confidence intervals, based on bootstrapping
 249 autosomal values, are shown in grey, and predicted boundaries for Stratum I from Almeida et al. (2021) are
 250 highlighted in purple.

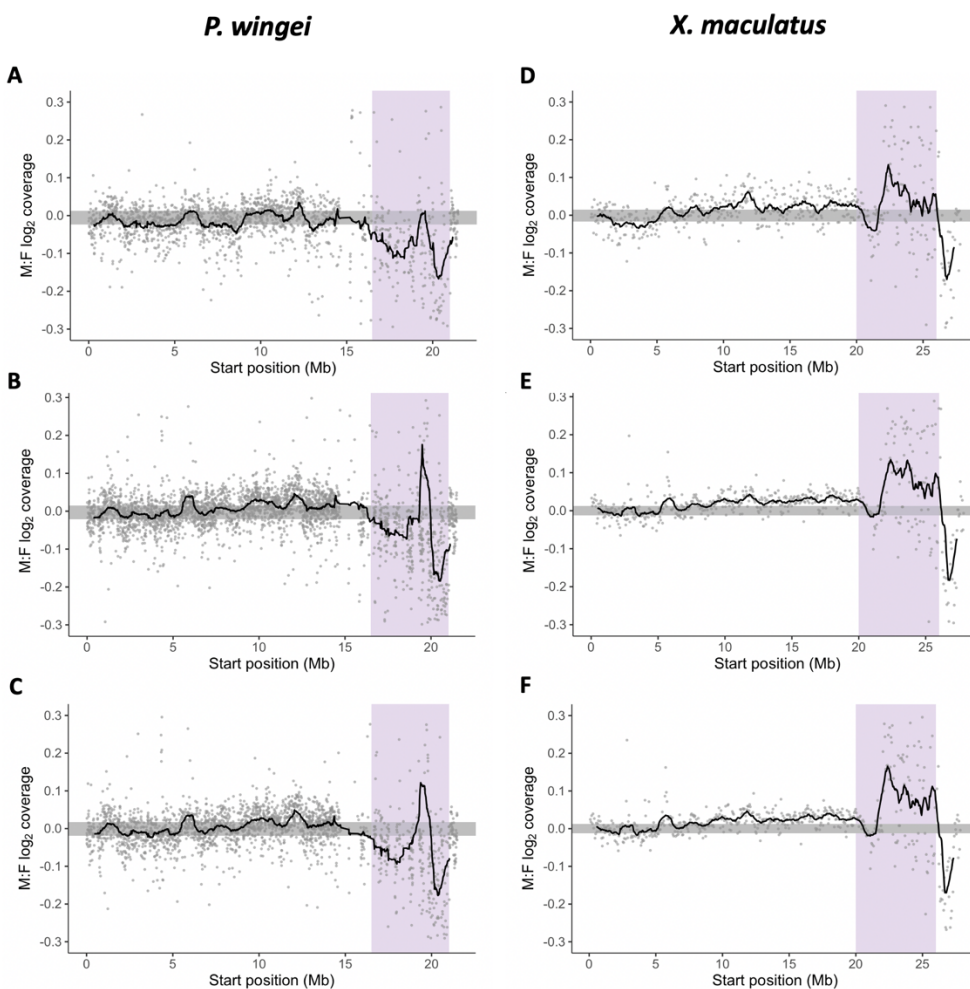
251 The analyses for *P. wingei* also reveal a lower male to female coverage at the distal end of the
252 chromosome, however, this pattern is only observed in analyses that mapped reads to the *de novo P.*
253 *wingei* assembly (Fig. 3A, B, C). By contrast, the analyses that used the *X. maculatus* genome all show a
254 significantly elevated read depth in males compared to females, similar to the results in Kirkpatrick et al.
255 (2020) (Fig. 3D, E, F). Previous cytogenetic work has shown that the *P. wingei* Y chromosome is the largest
256 chromosome in the genome, having accumulated a large heterochromatin block (Nanda et al. 2014).
257 However, in addition to the expansion of repetitive sequence, duplication events from the rest of the
258 genome could have also contributed to the remarkable size of the *P. wingei* Y chromosome. Duplications
259 from the X chromosome to the Y chromosome would explain a signal of elevated coverage in males
260 relative to females in this species.

Method

Mapping: `bwa -aln, -sampe`
Filtering: uniquely mapping reads
Coverage: `soap.coverage`
Wright et al. 2017

Mapping: `bowtie2 -local`
Filtering: remove PCR duplicates
Coverage: `bedtools`
Kirkpatrick et al. 2020

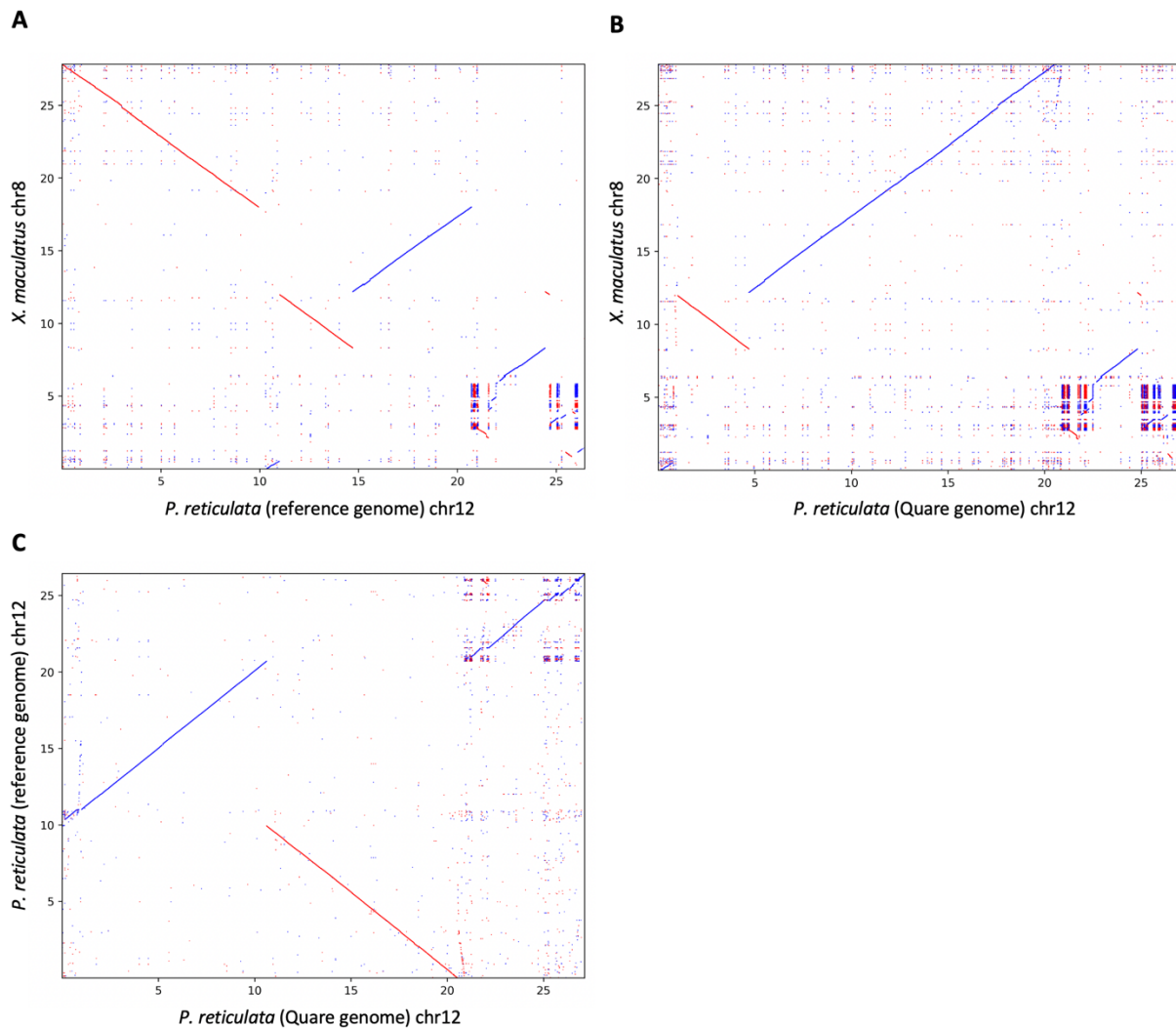
Mapping: `bwa -mem -M`
Filtering: remove PCR duplicates
Coverage: `samtools`
Bergero et al. 2019



261

262 **Figure 3. Signal for *P. wingei* Stratum I using comparative methodological approaches.** *P. wingei* DNA-seq reads
263 were mapped in turn to a *de novo P. wingei* genome assembly (Darolti et al. 2019) and the *X. maculatus* reference
264 genome assembly (Schartl et al. 2013). Moving average plots represent male to female coverage differences across
265 the sex chromosome (*P. wingei* chromosome 12, and syntenic *X. maculatus* chromosome 8) in non-overlapping

266 windows of 50kb for the analyses that rely on the *X. maculatus* genome and windows of 10kb for the analyses that
267 use the *de novo* *P. wingei* genome. The 95% confidence intervals, based on bootstrapping autosomal values, are
268 shown in grey, and predicted boundaries for Stratum I from Darolti et al. (2019) are highlighted in purple.



269

270 **Figure 4. Structural rearrangements and duplications between *P. reticulata* and *X. maculatus* genomes.** Dot-plots
271 of alignments between (A) *X. maculatus* chromosome 8 and *P. reticulata* chromosome 12 from the reference
272 genome assembly (Kunstner et al. 2016), (B) *X. maculatus* chromosome 8 and *P. reticulata* chromosome 12 from the
273 Quare *de novo* genome assembly (Almeida et al. 2021), and (C) *P. reticulata* chromosome 12 from the reference
274 genome assembly and chromosome 12 from the Quare *de novo* genome assembly. Forward alignments are shown
275 in blue and reverse alignments in red.

276

Table 2. Percentage of concordant, properly paired* read alignments

Data	Sequencing data	<i>P. reticulata</i>		<i>P. wingei</i>	
	Genome assembly	<i>P. reticulata</i>	<i>X. maculatus</i>	<i>P. wingei</i>	<i>X. maculatus</i>
Method	bwa -aln/-sampe <i>Wright et al. 2017</i>	59	5	75	17
	bowtie2 -local <i>Kirkpatrick et al. 2020</i>	83	56	87	71
	bwa -mem -M <i>Bergero et al. 2019</i>	83	72	84	79

*Both mates of a read pair map to the same chromosome or scaffold, with the expected insert size and read orientation.

277

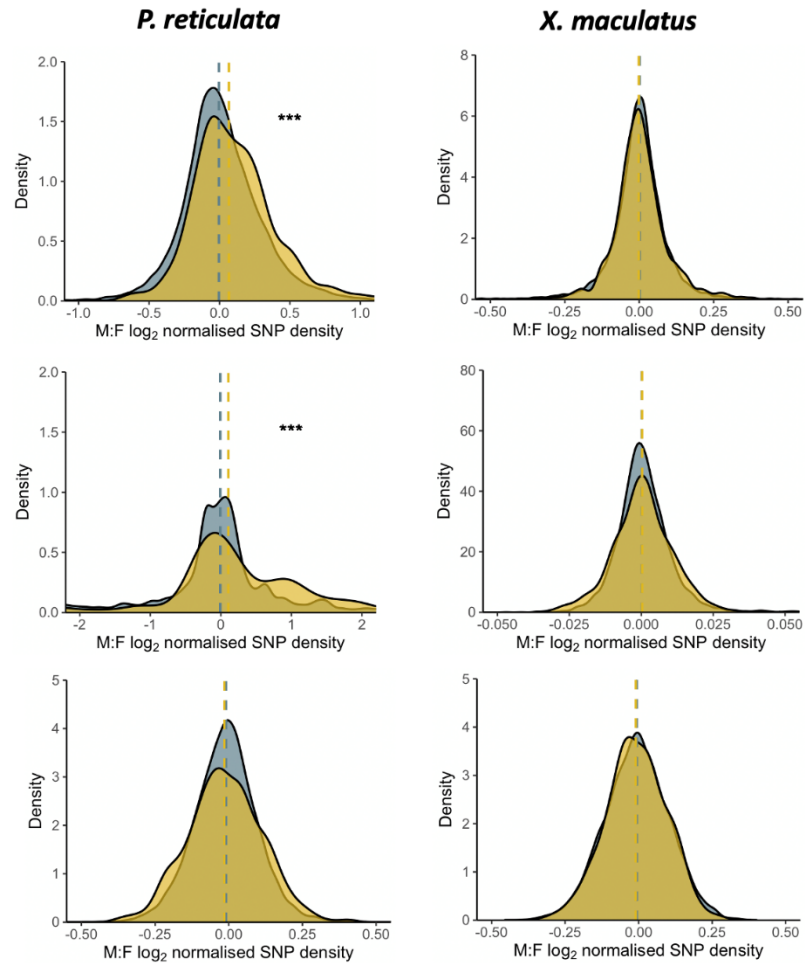
278 Regions of the sex chromosomes where recombination has recently been halted or greatly suppressed
279 still retain a high degree of similarity between X and Y sequences. They are also expected to show an
280 elevated SNP density in males compared to females, as Y-linked reads carrying Y-specific polymorphisms
281 will still align to the homologous X region of the female reference genome (Fig. 1; Vicoso et al. 2013). We
282 observed this pattern in *P. wingei* (Darolti et al. 2019) and in replicate upstream populations of *P.*
283 *reticulata* (Wright et al. 2017; Almeida et al. 2020), and we designated this as Stratum II. It is important to
284 note that in contrast to Stratum I, Stratum II appears to have formed independently several times.
285 Therefore, to further quantify divergence between the sex chromosomes we investigated SNP density
286 differences between the sexes using several methodological approaches. In *P. reticulata*, we observe a
287 significantly elevated male SNP density on the sex chromosomes in both of the analyses that aligned reads
288 to the *de novo P. reticulata* genome (Wilcoxon rank sum test $p < 0.001$, Fig. 5A, B). By contrast, the SNP
289 density profiles of the autosomes and the sex chromosomes were indistinguishable in all the analyses that
290 used *X. maculatus* as the reference genome (Fig. 5D, E, F), due to the accumulation of numerous fixed
291 differences between *P. reticulata* and *X. maculatus* which conceal the subtle polymorphisms differences
292 between *P. reticulata* males and females. The *P. wingei* X and Y chromosomes have previously been
293 suggested to be more diverged than those of *P. reticulata*, as shown through more pronounced coverage
294 and SNP density differences between the sexes (Darolti et al. 2019) and a greater accumulation of
295 repetitive sequences on the sex chromosomes in *P. wingei* compared to *P. reticulata* (Morris et al. 2018;
296 Almeida et al. 2021). Our results here confirm this, as we find a significantly higher male:female SNP
297 density for the sex chromosomes compared to the autosomes across all methodological analyses, as well
298 as when using either one of the *P. wingei de novo* or the *X. maculatus* genomes (Wilcoxon rank sum test
299 $p < 0.001$, Fig. 6).

Method

SNP calling: bow2pro
Filtering: min. coverage 10, MAF > 0.3
SNP density: per gene
Wright et al. 2017

SNP calling: bcftools
Filtering: QUAL 20, min. read depth 3
SNP density: per 50kb windows
Kirkpatrick et al. 2020

SNP calling: GATK
Filtering: QUAL 30, min. read depth 20
SNP density: per 50kb windows
Bergero et al. 2019



300

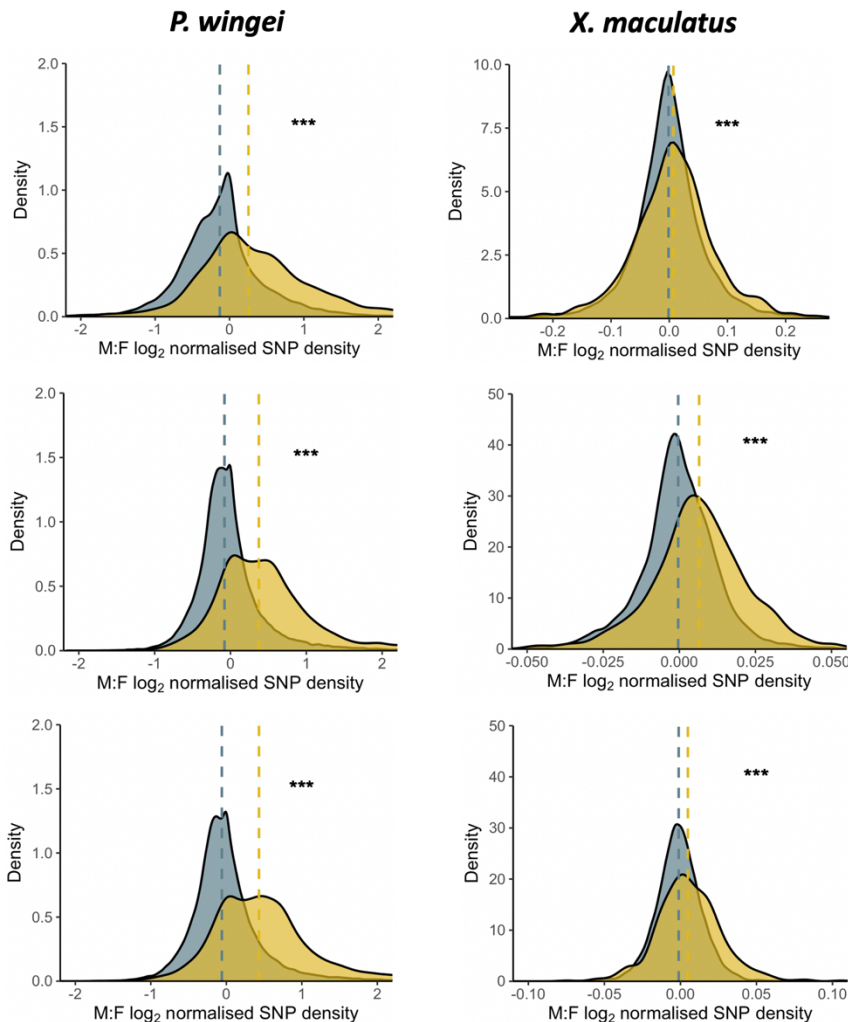
301 **Figure 5. Distribution of *P. reticulata* male:female SNP density for the autosomes (gray) and the sex**
302 **chromosomes (yellow).** Dashed vertical lines indicate median SNP densities and significant differences between
303 the autosomes and the sex chromosomes are shown (***) p -value < 0.001).

Method

SNP calling: bow2pro
 Filtering: min. coverage 10, MAF > 0.3
 SNP density: per gene
Wright et al. 2017

SNP calling: bcftools
 Filtering: QUAL 20, min. read depth 3
 SNP density: per 50kb windows
Kirkpatrick et al. 2020

SNP calling: GATK
 Filtering: QUAL 30, min. read depth 20
 SNP density: per 50kb windows
Bergero et al. 2019



304

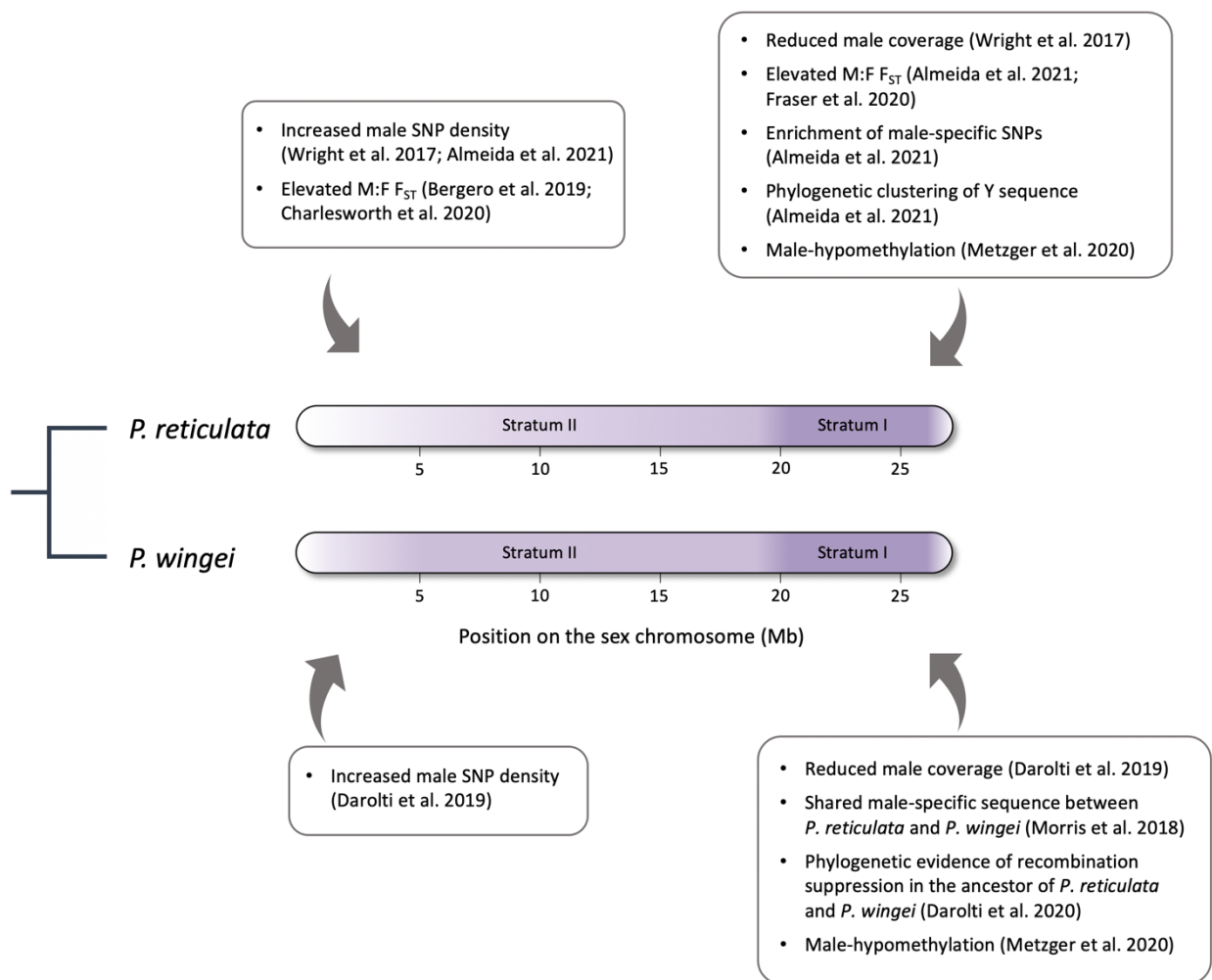
305 **Figure 6. Distribution of *P. wingei* male:female SNP density for the autosomes (gray) and the sex chromosomes**
 306 **(yellow).** Dashed vertical lines indicate median SNP densities and significant differences between the autosomes
 307 and the sex chromosomes are shown (***) p -value < 0.001.

308

309 Discussion

310 Replication is fundamental to scientific pursuits, and confirmation is necessary to build a robust
 311 understanding of the natural world. The expansion of public data efforts has greatly aided transparency
 312 and replication efforts, and this remarkable and rapid shift in the scientific culture is exemplified by
 313 genomics research, where most of the major journals require deposition of sequencing data as a condition
 314 of publication. Failures to replicate results are concerning, and necessitate further work to validate or
 315 nullify. However, it is important to understand that different replication approaches will have different
 316 risks of Type II errors, or erroneous negative results. This is especially problematic for the detection of
 317 subtle, small effect patterns, such as with initial divergence between X and Y chromosomes.

318 Here we used the same dataset across various methodologies and genome assemblies to test the
 319 sensitivity and accuracy of different approaches. Our results show how small changes in the precision of
 320 methods can lead to the failure to detect patterns of sex chromosome differentiation in the guppy. The
 321 low overall divergence between the X and Y can make detection difficult, but it has nonetheless been
 322 observed across multiple datasets, spanning DNA, RNA and methylation data, as well as multiple methods,
 323 including comparisons of male and female coverage and SNP density (Wright et al. 2017; Darolti et al.
 324 2019; Almeida et al. 2021), identification of male-specific sequence (Morris et al. 2018; Almeida et al.
 325 2021), phylogenetic analysis of recombination suppression (Darolti et al. 2020; Almeida 2021), and
 326 comparative epigenomics (Metzger et al. 2020) (Fig. 7).



327

328 **Figure 7. Structure of the *P. reticulata* and *P. wingei* sex chromosomes as predicted by multiple methods, including**
 329 **comparisons of male and female coverage, SNP density, phylogenetic and methylation analyses.**

330 By using the same sample data across multiple methods and genomes, our results illustrate how
331 important methodological differences can alter perceived results, and highlight the need for replication
332 studies to at minimum replicate the analysis using identical methods on the original or equivalent dataset.
333 When possible, replication efforts should go beyond minimum, and expand the analysis by employing
334 more sophisticated methods on existing or expanded datasets. Despite this, some replication efforts use
335 less sophisticated approaches, and in these cases, there is a real concern that a perceived failure of
336 replication is instead the result of a lack of precision or statistical power. This is particularly problematic
337 in the field of genomics, as there is little consensus about the gold standard in methodologies, particularly
338 with regard to data processing and filtering procedures. The lack of standardized practices, coupled with
339 the rich nature of genomic data and the complexity of genomes can make it difficult to discern subtle but
340 important patterns.

341 Our approach of evaluating the same underlying data with multiple methods and genomes does not
342 account for natural variation across samples and populations, which is substantial (Wright et al. 2017;
343 Almeida et al. 2021). For our *P. reticulata* samples, we chose individuals from an upstream low predation
344 Quare population which we have previously shown to have an intermediate signal of sex chromosome
345 divergence (Wright et al. 2017; Almeida et al. 2021). Samples from populations with greater or lesser
346 signal, or sampling variation due to differences in inversions, duplications and divergence among
347 individuals may also contribute to observed differences.

348 ***Stratum I***

349 We have previously observed evidence for a small region of ancestral recombination suppression in *P.*
350 *wingei* and *P. reticulata* (Wright et al. 2017; Darolti et al. 2019; Almeida et al. 2021). This has been
351 replicated in some studies, for example Fraser et al. (2020) also found evidence of small regions of Y
352 divergence, and Kirkpatrick et al. (2020) confirmed the phylogenetic clustering of Y sequence in this
353 stratum. However, other studies (Bergero et al. 2019; Charlesworth et al. 2020; Kirkpatrick et al. 2021) did
354 not fully replicate these findings.

355 It is worth noting that Stratum I region of the guppy Y chromosome is enriched for repetitive elements
356 (Almeida et al. 2021), and reads from this region may, depending on the parameters used, map to
357 repetitive elements across the genome, obscuring real read depth differences between males and females
358 if non-coding sequence is included in the analysis. Focusing on uniquely mapping reads when comparing
359 coverage differences between males and females can minimize issues associated with Y repetitive regions.

360 However, our comparative analysis revealed that a pattern of X-Y differentiation can still be recovered
361 without restricting the analysis to uniquely mapping reads (Fig. 2, Fig. 3). More stringent SNP filtering
362 parameters can also help eliminate noise in genomic comparisons, and this is particularly important when
363 studying young sex chromosomes as they are expected to exhibit subtle divergence signatures. We were,
364 however, able to identify a signal of elevated male SNP density on the sex chromosomes relative to the
365 autosomes, indicative of Y divergence, across several methodological approaches using different degrees
366 of filtering stringency (Fig. 5, Fig. 6).

367 Beyond mapping parameters, by far the most substantial source of variation in the results of the different
368 pipelines we compared lies in the reference genome used. This is in part due to the extensive structural
369 variation across populations and species (Fig. 4), but also due to sequence evolution. These two factors
370 combined mean that error compounds over phylogenetic distances, and as the distance between the
371 samples and the genome they are mapped to increases, the ability to detect reduced male:female read
372 depth decreases. This is most evidenced in the strategy by Kirkpatrick et al. (2020), who mapped reads
373 from *Poecilia* species to the *Xiphophorus* genome. They argued that changes over the 40 my phylogenetic
374 distance separating these genera was outweighed by the fact that the *Xiphophorus* genome is more
375 complete. However, the read mapping rate in Table 2 reveals instead that this strategy is less accurate
376 than using less-complete species- or population-specific genome assemblies, as a significantly smaller
377 proportion of *Poecilia* reads map to the *Xiphophorus* genome across all methods, thereby reducing usable
378 data. This problem is exacerbated by the substantial structural differences between *Xiphophorus* and
379 *Poecilia* on the sex chromosome (Fig. 4), further complicating the comparison. Interestingly, their mapping
380 and filtering methods would have detected Stratum I if they had mapped to a con-specific genome (Fig.
381 2B, Fig. 3B). To a lesser extent, this is also a problem when mapping data to genomes assembled on
382 different *P. reticulata* populations. The genome used can also greatly affect the perceived patterns of SNP
383 diversity, and relying on the distantly related *Xiphophorus* genome can obscure a signal of elevated male
384 SNP density on the sex chromosomes due to fixed differences between the target species reads and the
385 *Xiphophorus* sequence (Fig. 5).

386 **Stratum II**

387 As recombination is increasingly suppressed in nascent regions of a sex chromosome, we expect the
388 accumulation of Y-specific SNPs, and we observed this in replicate upstream populations of *P. reticulata*
389 (Wright et al 2019; Almeida et al. 2021) and in *P. wingei* (Darolti et al. 2019), consistent with convergent
390 evolution across populations and species (Darolti et al. 2020). Whether this is due to the important

391 environmental effects on recombination rate (Plough 1917; Grell 1971; Stevison et al. 2019), sexual
392 conflict (Wright et al. 2017), neutral shifts in male recombination hotspots (Wright et al 2016; Bergero et
393 al. 2019) or selection against recombinants in the wild remains an important area of further work.

394 Additionally, given that many mechanisms of recombination suppression only accumulate over time
395 (Furman et al. 2020 and references cited), it also remains unclear how complete recombination
396 suppression is in this region, and whether rare recombination events observed in this region in lab-reared
397 males (Bergero et al. 2019) occur in wild populations. Regardless, it is important to note that suppressed
398 recombination does not necessarily mean that recombination never occurs between the X and Y
399 chromosomes, but rather that it is at least exceedingly rare or recombinant individuals are selected
400 against.

401 Because of the expected heterogeneity observed in the initial stages of the divergence process (Bergero
402 et al. 2013; Natri et al. 2013; Reichwald et al 2015), sliding window approaches may be insufficient to
403 reveal overall patterns of elevated male SNP density expected in these regions. Density distributions or
404 direct statistical comparisons between species may be required. This is evidenced by our observation of
405 elevated male:female SNP density across nearly all methods (Fig. 5 and 6), with the exception of *P.*
406 *reticulata* data mapped to the *Xiphophorus* genome, again illustrating the problems with mapping over
407 vast evolutionary distances.

408

409 **Concluding remarks**

410 Here we have used the same data to compare methods and genomes in the discovery of nascent sex
411 chromosomes. We hope that our results provide a gold standard for future work in other study systems,
412 and resolve some of the recent controversy over the sex chromosomes in *Poecilia*.

413

414 **Acknowledgements**

415 We thank Jacelyn Shu of jacelyndesigns.com for drafting Fig. 1. We gratefully acknowledge funding from
416 the ERC (grant agreement 680951), NSERC and a Canada 150 Research Chair to JEM.

417

418 **References**

419 Almeida, P., Sandkam, B.A., Morris, J., Darolti, I., Breden, F., Mank, J.E. (2021) Divergence and
420 remarkable diversity of the Y chromosome in guppies. *Molecular Biology & Evolution* 38, 619-633

- 421 Bergero, R., Qui, S., Forrest, A., Borthwick, H., Charlesworth D. (2013) Expansions of the pseudo-
422 autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes.
423 *Genetics* 194, 673-686
- 424 Bergero, R., Gardner, J., Bader, B., Yong, L., Charlesworth, D. (2019) Exaggerated heterochiasmy in a fish
425 with sex-linked male coloration polymorphisms. *Proceedings of the National Academy of Sciences USA*
426 116, 6924-6931
- 427 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence
428 data. *Bioinformatics*, 30, 2114-2120.
- 429 Carvalho, A.B., Clark, A.G. (2013). Efficient identification of Y chromosome sequences in the human and
430 *Drosophila* genomes. *Genome Research* 23, 1894–1907
- 431 Charlesworth, D., Bergero, R., Graham, C., Gardner, J., Yong, L. 2020. Locating the sex determining
432 region of linkage group 12 of guppy (*Poecilia reticulata*). *G3* 10, 3639-3649
- 433 Darolti, I., Wright, A.E., Mank, J.E. (2020) Guppy Y chromosome integrity maintained by incomplete
434 recombination suppression. *Genome Biology & Evolution* 12, 965-977
- 435 Darolti, I., Wright, A.E., Sandkam, B.A., Morris, J., Bloch, N.I., Farré, M., Fuller, R.C., Bourne, G.R. Larkin,
436 D.M., Breden, F., Mank, J.E. (2019) Extreme heterogeneity in sex chromosome differentiation and dosage
437 compensation in livebearers. *Proceedings of the National Academy of Sciences, USA* 116, 19031-19036
- 438 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G.,
439 Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group. (2011). The
440 variant call format and VCFtools. *Bioinformatics* 27, 2156-2158
- 441 Fraser, B.A., Whiting, J.R., Paris, J.R., Weadick, C.J., Parsons, P.J., Charlesworth, D., Bergero, R., Bemm, F.,
442 Hoffmann, M., Kottler, V.A., Liu, C., Dreyer, C., Weigel, D. (2020) Improved reference genome uncovers
443 novel sex-linked regions in the guppy (*Poecilia reticulata*). *Genome Biology & Evolution* 12, 1789–1805
- 444 Furman, B.L.S., Metzger, D.C.H., Darolti, I., Wright, A.E., Sandkam, B.A., Almeida, P., Shu, J.J., Mank, J.E.
445 (2020) Sex chromosome evolution: So many exceptions to the rules. *Genome Biology & Evolution* 12,
446 750-763
- 447 Grell, R.F., 1971. Heat induced exchange in fourth chromosome of diploid females of *Drosophila*
448 *melanogaster*. *Genetics* 69, 523-527
- 449 Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic
450 sequence comparison. *Genome research* 21, 487-493
- 451 Kirkpatrick, M., Dixon, G., Sardell, J.M., Schartl, M., Peichel, C.L. (2020). Evolution of the canonical sex
452 chromosomes of the guppy and relatives. *Biorxiv* doi: <https://doi.org/10.1101/2020.09.25.314112>
- 453 Kumar, S., Stecher, G., Suleski, M., Hedges, S.B. (2017) TimeTree: A Resource for Timelines, Timetrees,
454 and Divergence Times. *Molecular Biology and Evolution* 34, 1812–1819
- 455 Kunstner, A., Hoffmann, M., Fraser, B.A., Kottler, V.A., Sharma, E., Weigel, D., Dreyer, C. (2016) The
456 genome of the Trinidadian guppy, *Poecilia reticulata*, and variation in the Guanapo population. *PLOS*
457 *ONE* 11, e0169087
- 458 Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 357-
459 359

- 460 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and
461 population genetical parameter estimation from sequencing data. *Bioinformatics*, 27, 2987-2993
- 462 Li, H., Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler
463 transform. *Bioinformatics* 25, 1754–1760
- 464 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.,
465 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and
466 SAMtools. *Bioinformatics* 25, 2078-2079
- 467 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler,
468 D., Gabriel, S., Daly, M., DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework
469 for analyzing next-generation DNA sequencing data. *Genome research* 20, 1297-1303
- 470 Metzger, D.C.H., Mank, J.E. (2020) Conserved sex-biased DNA methylation patterns target key
471 developmental genes and non-recombining region of the guppy sex chromosome. *Biorxiv* doi:
472 <https://doi.org/10.1101/2020.08.21.261792>
- 473 Morris, J., Darolti, I., Bloch, N.I., Wright, A.E., Mank, J.E. (2018). Shared and species-specific patterns of
474 nascent Y chromosome evolution in two guppy species. *Genes* 9, 238
- 475 Nanda, I., Schartl, M., Epplen, J.T., Feichtinger, W., Schmid, M (1993) Primitive sex chromosomes in
476 poeciliid fishes harbour simple repetitive DNA sequences. *Journal of Experimental Biology* 265, 301-
477 308
- 478 Natri, H.M., Shikano, T., Merila, J. (2013) Progressive recombination suppression and
479 differentiation in recently evolved neo-sex chromosomes. *Molecular Biology & Evolution* 30, 1131-
480 1144
- 481 Palmer, D.H., Rogers, T.F., Dean, R., Wright, A.E. (2019) How to identify sex chromosomes and their
482 turnover. *Molecular Ecology*. 28, 4709– 4724
- 483 Plough, H.H. 1917. The effect of temperature on crossing over in *Drosophila*. *Journal of Experimental*
484 *Zoology* 24, 147–209
- 485 Postlethwait, J.H., Warren, W.C. (2013) The genome of the platyfish, *Xiphophorus maculatus*, provides
486 insights into evolutionary adaptation and several complex traits. *Nature Genetics* 45, 567-572
- 487 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
488 features. *Bioinformatics* 26, 841-842
- 489 R Core Team (2019) R: A language and environment for statistical computing. R Foundation for
490 Statistical Computing, Vienna.
- 491 Reichwald, K., et al. (2015) Insights into sex chromosome evolution and agene from the genome of a
492 short-lived fish. *Cell* 163, 1527-1538
- 493 Sandkam, B.A., Almeida, P., Darolti, I., Furman, B.L.S., van der Bijl W., Morris, J., Bourne, G.R., Breden, F.,
494 Mank, J.E. (2021) Extreme Y chromosome polymorphism corresponds to five male reproductive morphs
495 in a freshwater fish. *Nature Ecology & Evolution* 5, 939-948
- 496 Schartl, M., Walter, R.B., Shen, Y.J., Garcia, T., Catchen, J., Amores, A., Braasch, I., Chalopin, D., Volf, J.,
497 J.N., Lesch, K.P., Bisazza, A., Minx, P., Hillier, L., Wilson, R.K., Fuerstenberg, S., Boore, J., Searle, S.,

- 498 Postlethwait, J.H., Warren, W.C. (2013). The genome of the platyfish, *Xiphophorus maculatus*, provides
499 insights into evolutionary adaptation and several complex traits. *Nature genetics* 45, 567-572
- 500 Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., Ellegren, H. (2015). Evolutionary
501 analysis of the female-specific avian W chromosome. *Nature Communications* 6,7330
- 502 Stevison, L.S., Sefich, S., Chase, R., Graze, R.M. (2017). Recombination rate plasticity: revealing
503 mechanisms by design. *Philosophical Transactions of the Linnean Society, B.* 372, 20160459
- 504 Vicoso, B., Bachtrog, D. (2015). Numerous transitions of sex chromosomes in Diptera. *PLOS Biology*, 13,
505 e1002078
- 506 Vicoso, B., Bachtrog, D. (2013). Reversal of an ancient sex chromosome to an autosome in *Drosophila*.
507 *Nature*, 499, 332–335
- 508 Vicoso, B., Emerson, J. J., Zektser, Y., Mahajan, S., & Bachtrog, D. (2013). Comparative sex chromosome
509 genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS*
510 *biology*, 11, e1001643
- 511 Wright, A.E., Dean, R., Zimmer, F., Mank, J.E. (2016) How to make a sex chromosome. *Nature*
512 *Communications* 7, 12087
- 513 Wright, A.E., Darolti, I., Bloch, N.I., Oostra, V., Sandkam, B., Buechel, S.D., Kolm, N., Breden, F., Vicoso,
514 B., Mank, J.E. (2017) Convergent recombination suppression suggests a role of sexual selection in guppy
515 sex chromosome formation. *Nature Communications* 8: 14251