Assembly of a young vertebrate Y chromosome reveals convergent signatures of sex chromosome evolution

Catherine L. Peichel[1,2]
Shaugnessy R. McCann[1]
Joseph A. Ross[1,3]
Alice F. S. Naftaly[4]
James R. Urton[1,3]
Jennifer N. Cech[1,3]
Jane Grimwood[5]
Jeremy Schmutz[5]
Richard M. Myers[5]
David M. Kingsley[6]
Michael A. White[1,4]

[1]Divisions of Human Biology and Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA
[2]Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
[3]Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle, Washington 98195, USA
[4]Department of Genetics, University of Georgia, Athens 30602, Georgia, USA
[5]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA
[6]Department of Developmental Biology and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA

Corresponding authors:

Catherine L. Peichel
Institute of Ecology and Evolution
University of Bern
Baltzerstrasse 6
3012 Bern, Switzerland
catherine.peichel@iee.unibe.ch

Michael A. White
Department of Genetics
University of Georgia
120 Green St.
Athens, GA 30601
USA
whitem@uga.edu

Running title: Threespine stickleback Y chromosome assembly

Keywords: threespine stickleback, Y chromosome, gene duplication, sex determination

Abstract

Heteromorphic sex chromosomes have evolved repeatedly across diverse species. Suppression of recombination between X and Y chromosomes leads to rapid degeneration of the Y chromosome. However, these early stages of degeneration are not well understood, as complete Y chromosome sequence assemblies have only been generated across a handful of taxa with ancient sex chromosomes. Here we describe the assembly of the threespine stickleback (*Gasterosteus aculeatus*) Y chromosome, which is less than 26 million years old. Our previous work identified that the non-recombining region between the X and the Y spans ~17.5 Mb on the X chromosome. Here, we combined long-read PacBio sequencing with a Hi-C-based proximity guided assembly to generate a 15.87 Mb assembly of the Y chromosome. Our assembly is concordant with cytogenetic maps and Sanger sequences of over 90 Y chromosome clones from a bacterial artificial chromosome (BAC) library. We found three evolutionary strata on the Y chromosome, consistent with the three inversions identified by our previous cytogenetic analyses. The young threespine stickleback Y shows convergence with older sex chromosomes in the retention of haploinsufficient genes and the accumulation of genes with testis-biased expression, many of which are recent duplicates. However, we found no evidence for large amplicons found in other sex chromosome systems. We also report an excellent candidate for the master sex-determination gene: a translocated copy of *Amh* (*Amhy*). Together, our work shows that the same evolutionary forces shaping older sex chromosomes can cause remarkably rapid changes in the overall genetic architecture on young Y chromosomes.

Introduction

Sex chromosomes evolve from autosomal ancestors when recombination is suppressed between the homologous pairs (reviewed in Bachtrog 2013). Thus, sex chromosomes are an intriguing region of the genome to understand how mutations and repetitive DNA accumulate in the absence of recombination and how gene content evolves once a chromosome becomes sex-

76 limited. Y chromosomes were once thought to be an evolutionary dead end, inevitably losing
77 functional gene copies across the entire chromosome as deleterious mutations quickly
78 accumulate (Griffin 2012). Contrary to this expectation, assembly of multiple mammalian Y
79 chromosome sequences (Skaletsky et al. 2003; Hughes et al. 2010; 2012; Bellott et al. 2014; Soh
80 et al. 2014), the chicken W chromosome (Bellott et al. 2017), and an invertebrate Y chromosome
81 (Mahajan et al. 2018) has revealed that the sequence of the sex-limited chromosome is much
82 more dynamic, punctuated by gene gains and losses, rather than becoming entirely degenerated.
83 Although short-read sequencing of sex chromosomes has yielded insight into how
84 ancestral single-copy genes have evolved between X and Y chromosomes (e.g. Zhou et al. 2014;
85 Papadopulos et al. 2015; White et al. 2015), these approaches cannot be used to study how Y
86 chromosomes have structurally evolved. Short-reads cannot span many of the lengthy repeat
87 units characteristic of Y chromosomes, leading to a collapse of these regions during the assembly
88 process. Because of the inherent difficulty in assembling these highly repetitive regions of the
89 genome, Y chromosomes have been omitted from many reference genome assemblies. The few
90 existing reference Y chromosome assemblies were constructed through labor intensive, iterative
91 Sanger sequencing of large inserts from bacterial artificial chromosome (BAC) libraries (Skaletsky
92 et al. 2003; Hughes et al. 2010; 2012; Soh et al. 2014), recently supplemented by a combination
93 of Pacific Biosciences long-read sequencing, chromatin interaction maps, and optical mapping
94 (Mahajan et al. 2018).
95 Through these assemblies, two classes of genes have been identified on ancient sex
96 chromosomes. The first are dosage-sensitive genes that were present in the common ancestor
97 of both chromosomes and have been maintained as single copies on the Y chromosome across
98 multiple mammalian lineages (Bellott et al. 2014; Cortez et al. 2014) as well as on the
99 degenerating W chromosome of birds (Bellott et al. 2017). The second are genes that exist in
100 high copy number families on Y chromosomes and generally have gene expression patterns
101 restricted to the testes, suggesting roles in spermatogenesis (Skaletsky et al. 2003; Murphy et al.

102 2006; Hughes et al. 2010; Paria et al. 2011; Soh et al. 2014; Janečka et al. 2018). It is clear that
103 the genetic architecture of sex chromosomes can be shaped by multiple processes over long
104 evolutionary time scales. However, in many species sex chromosomes can be much younger,
105 due to the frequent turnover of the linkage groups controlling sex determination (e.g. Ross et al.
106 2009; Kitano and Peichel 2012; Bachtrog et al. 2014; Blackmon and Demuth 2014; Myosho et al.
107 2015; Jeffries et al. 2018). Reference assemblies of young sex chromosomes are largely absent,
108 with the exception of the young neo-Y chromosome assembly of *Drosophila miranda* (Mahajan et
109 al. 2018; Ellison and Bachtrog 2019), making it unclear whether the genetic architecture of newly
110 evolving sex chromosomes is rapidly shaped by these evolutionary forces or if it is a phenomenon
111 unique to ancient sex chromosomes.

112 The threespine stickleback fish (*Gasterosteus aculeatus*) is an excellent model system to
113 explore the early structural evolution of sex chromosomes. Although the threespine stickleback
114 has a high-quality reference genome assembly (Jones et al. 2012) that has gone through multiple
115 iterations of refinement (Roesti et al. 2013; Glazer et al. 2015; Peichel et al. 2017), the assembly
116 was derived from a female fish, precluding the Y chromosome from assembly. The threespine
117 stickleback has a relatively young X/Y sex chromosome system that is shared across the
118 *Gasterosteus* genus but not with other species in the Gasterosteidae family and therefore evolved
119 less than 26 million years ago (Bell et al. 2009; Kitano et al. 2009; Ross et al. 2009; Varadharajan
120 et al. 2019) (compared to the Y chromosome of mammals that evolved ~180 million years ago
121 (Bellott et al. 2014; Cortez et al. 2014)). Crossing over is suppressed between the X and Y
122 chromosomes over a majority of their length, resulting in an approximately 2.5 Mb
123 pseudoautosomal region of the 20.6 Mb X chromosome (Roesti et al. 2013). The region of
124 suppressed crossing over is coincident with three pericentric inversions that differentiate the X
125 and Y chromosomes (Ross and Peichel 2008). Illumina-based sequencing suggested the non-
126 crossover region on the Y chromosome was composed of two differently aged evolutionary strata,
127 the oldest of which retained genes that were predicted to be haploinsufficient (White et al. 2015),

128 similar to mammals. However, all studies in threespine stickleback have relied on mapping short-
129 reads to the reference X chromosome, limiting our understanding to regions conserved between
130 the X and Y. It has not yet been possible to explore how unique structure and sequence is evolving
131 across this young Y chromosome.

132 Here, we report the first high-quality reference assembly of a young vertebrate Y
133 chromosome. We combined high-coverage, long-read sequencing with chromatin conformation
134 capture sequencing (Hi-C) to assemble a full scaffold of the threespine stickleback Y
135 chromosome. Our assembly is completely concordant with more than 90 Sanger sequenced
136 inserts from a bacterial artificial chromosome (BAC) library and with a known cytogenetic map
137 (Ross and Peichel 2008). Throughout the male-specific region we have identified several novel
138 sequence and structural characteristics that parallel patterns observed on more ancient sex
139 chromosome systems. The young sex chromosome of threespine stickleback is a useful model
140 system to understand how the genetic architecture of sex-limited chromosomes initially evolves.
141

142 **Results**

143 **De novo assembly of the threespine stickleback Y chromosome**

144 We used high-coverage PacBio long-read sequencing to assemble a threespine
145 stickleback genome from a male fish of the Paxton Lake Benthic population (British Columbia,
146 Canada). Raw read coverage was approximately 75.25x across the genome (34.84 Gb total
147 sequence) (Supplemental Table 1). The longest raw PacBio reads were assembled using the
148 Canu pipeline, refined by Arrow, resulting in a primary contig assembly of 622.30 Mb across 3,593
149 contigs (Supplemental Table 1). This assembly size was considerably larger than the Hi-C revised
150 threespine stickleback female genome assembly (463.04 Mb including autosomes and X
151 chromosome) (Jones et al. 2012; Glazer et al. 2015; Peichel et al. 2017). The increased assembly
152 length was largely due to heterozygous loci being separated into individual alleles (haplotigs).
153 3,134 contigs (574.67 Mb) of the total Canu assembly aligned to 442.41 Mb of autosomes in the

154 reference assembly. Only 129 contigs partially aligned to the genome (less than 25% of the contig
155 length aligned; 10.15 Mb) and 148 contigs did not align at all to the genome (3.58 Mb). We
156 collapsed 118.89 Mb of haplotigs, reducing the 574.67 Mb alignment to 455.78 Mb of non-
157 redundant sequence across the autosomes, an estimate closer to the 442.41 Mb of autosomes
158 in the female reference genome assembly.

159 We targeted Y-linked contigs in the Canu assembly by identifying contigs that shared
160 sequence homology with the reference X chromosome or did not align to the autosomes. In the
161 youngest region of the threespine stickleback sex chromosomes (the previously identified stratum
162 two), the X and Y chromosomes still share considerable sequence homology. However, within
163 this stratum, heterozygosity is even higher than what is observed across the autosomes (White
164 et al. 2015). Based on this divergence, Canu should separate X- and Y-linked contigs during the
165 initial assembly process. Contigs aligned to the X chromosome formed a distribution of sequence
166 identity that was not unimodal, reflecting the presence of both X- and Y-linked contigs
167 (Supplemental Figure 1). Setting a sequence identity threshold of 96% resulted in a set of 114 X-
168 linked contigs that totaled 21.26 Mb, compared to the previous 20.62 Mb X chromosome
169 reference assembly. There were 68 putative Y-linked contigs that had a sequence identity less
170 than or equal to 96%, totaling 12.64 Mb. The oldest region of the Y chromosome (stratum 1)
171 contains many regions that have either been deleted or have diverged to such an extent that
172 sequencing reads cannot be mapped to this region (White et al. 2015). Consequently, there may
173 be contigs unique to the Y chromosome that cannot be captured through alignments to the
174 reference X chromosome. To account for these loci, we also included the contigs that only partially
175 aligned to the genome (less than 25% of the contig length aligned; 129 contigs; 10.15 Mb) or did
176 not align at all to the genome (148 contigs; 3.58 Mb) in the set of putative Y-linked contigs (345
177 total contigs).

178

179 **Hi-C proximity-guided assembly yielded contiguous scaffolds of the sex chromosomes**

180 We used chromosome conformation capture (Hi-C) sequencing and a proximity-guided
181 method to assemble the set of putative X- and Y-linked contigs into scaffolds. Using the 3D-DNA
182 assembler (Dudchenko et al. 2017), 105 of the 114 X-linked contigs were combined into three
183 main scaffolds that totaled 20.78 Mb. The scaffolds were largely colinear with the reference X
184 chromosome, with scaffolds one and two aligning to the pseudoautosomal region and scaffold
185 three mostly aligning to the remainder of the X chromosome that does not recombine with the Y
186 (Supplemental Figure 2).

187 We assembled the putative Y-linked contigs using the same process. Of the 345 total
188 contigs, 115 were initially combined into a single primary scaffold that totaled 17.15 Mb. We
189 visually inspected the Hi-C interaction map for any sign of misassembled contigs. There was a
190 clear mis-joining of contigs near one end of the primary scaffold, where there were fewer short-
191 range Hi-C interactions at the diagonal combined with an overall absence of long-range Hi-C
192 interactions between all of the contigs in this region and the remainder of the Y scaffold
193 (Supplemental Figure 3). We manually removed this cluster of contigs from the primary scaffold
194 (45 contigs; 1.86 Mb), resulting in an initial Y chromosome scaffold totaling 15.28 Mb across 70
195 contigs (Figure 1A).

196

197 **Bacterial artificial chromosome library sequences are concordant with the assembled Y**
198 **chromosome**

199 To assess the overall accuracy of our assembly, we compared our assembly to Sanger
200 sequenced inserts from a bacterial artificial chromosome (BAC) library constructed from males
201 from the same population. Mean insert size among the 101 sequenced BAC clones was 168.13
202 kb, similar in size to the average contig length within the Y chromosome scaffolds (217.85 kb).
203 Using the BAC sequences, we were able to identify whether any of the contigs within the scaffold
204 contain collapsed haplotigs between the X and Y chromosome (mosaic contigs should contain

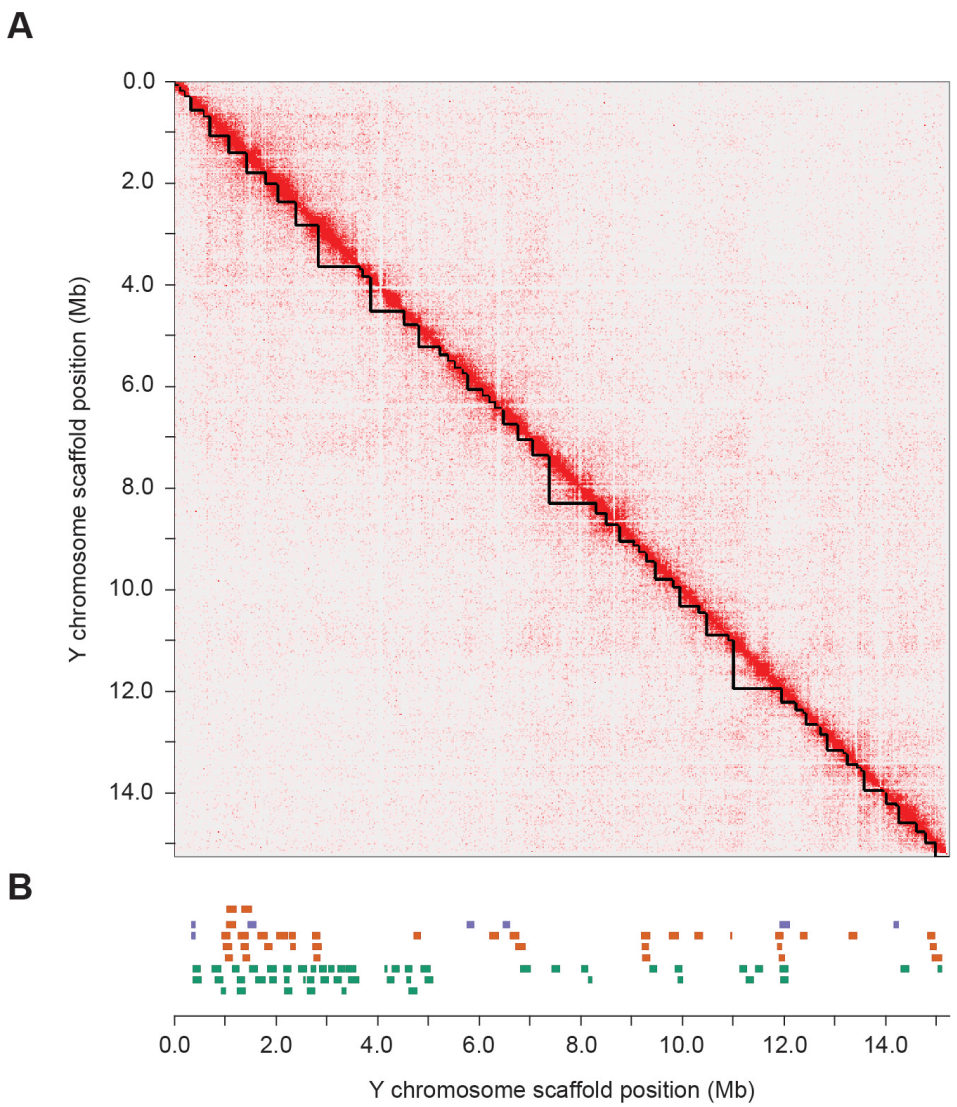

Figure 1. Hi-C chromosome conformation capture sequencing generated a single Y chromosome scaffold. (A) The contact matrix shows an enrichment of interactions between contigs in close proximity along the diagonal. Contig boundaries in the assembly are denoted by the black triangles along the diagonal. (B) Sanger sequenced BAC inserts that align concordantly throughout the scaffold are shown, with BACs that spanned gaps between contigs in orange, BACs that extended into, but did not span gaps in purple, and BACs that were contained completely within an individual contig in green.

205 reduced sequence identity when aligned to known Y chromosome BAC contigs). In addition, the
206 contig ordering across the scaffold was verified by BAC contig sequences that spanned gaps in
207 the assembly. We aligned all 101 sequenced BAC contigs to the Y chromosome scaffold and
208 found 92 of the BAC contigs aligned concordantly with the assembly (Figure 1B). These BACs
209 aligned to 40 of the 70 contigs in the assembly with a high sequence identity (7.72 Mb of non-

210 overlapping sequence in the 15.28 Mb assembly aligned concordantly to the BAC contigs). The
211 remaining 9 BAC contigs that did not align concordantly indicate there are small-scale structural
212 differences between the Canu Y chromosome assembly and the BAC clones derived from a
213 separate Paxton Lake male threespine stickleback, either reflecting errors in the Y chromosome
214 assembly, rearrangements in the BAC clone sequences, or true polymorphisms segregating in
215 the Paxton Lake benthic population. Four of the discordant BACs aligned to regions of the
216 reference Y that were greater than the Sanger sequenced length of the BAC insert, suggesting
217 possible indels. The remaining five discordant BACs contained sub alignments with mixed
218 orientations, suggesting possible small-scale inversions not present in our assembly.

219 Among the aligned BAC contigs, many provided additional sequence information, either
220 spanning gaps between contigs in the Y chromosome assembly or extending from contigs into
221 gaps in the assembly. Of the 92 BAC contigs that aligned concordantly, seven BAC contigs
222 extended into five different gaps in the assembly and 35 BAC contigs spanned 18 different gaps
223 in the assembly (26% of the total gaps in the assembly) (Figure 1B). The remainder of the aligned
224 BAC contigs aligned completely within an individual contig in the Y assembly. We merged this
225 additional sequence into the initial Y chromosome assembly, resulting in a merged Y chromosome
226 scaffold that contained 52 contigs, totaling 15.78 Mb.

227

228 **The Y chromosome assembly is concordant with known cytogenetic maps**

229 The threespine stickleback Y chromosome has undergone at least three pericentric
230 inversions relative to the X chromosome, forming a non-crossover region that spans a majority of
231 the chromosome in males (Ross and Peichel 2008). These inversions were mapped by ordering
232 a series of cytogenetic markers along both the X and Y chromosomes (Figure 2A). To determine
233 whether our Y chromosome assembly was consistent with the known cytogenetic marker
234 ordering, we used BLAST to locate the position of each marker within the assembly. We were
235 able to locate four of the five markers used from the male-specific region in our assembly. The

236 position of these cytogenetic markers was concordant with our assembly (Figure 2B). The missing
237 marker in the non-crossover region (*STN235*) likely reflects a region of our Y reference that is not
238 fully assembled or it is a true deletion within the Paxton Lake benthic population, relative to the
239 Pacific Ocean marine population used for the cytogenetic map (Ross and Peichel 2008).
240 The location of the oldest region within the Y chromosome (the previously identified
241 stratum one) had been ambiguous. Cytogenetic markers from this region could not be hybridized

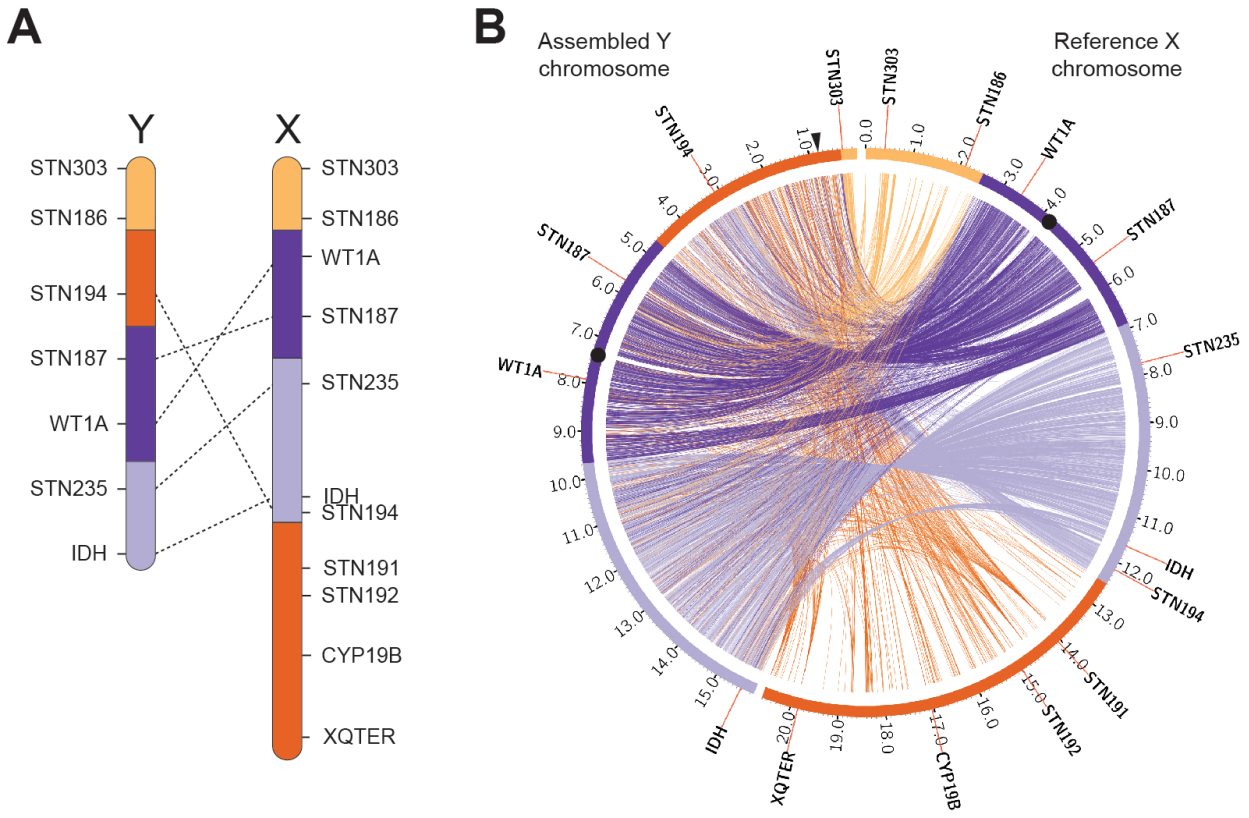

Figure 2. The threespine stickleback Y chromosome assembly is concordant with cytogenetic maps. (A) The Y chromosome has diverged from the X chromosome through a series of inversions determined through ordering of cytogenetic markers (dashed lines indicate rearrangements of the linear order of markers (Ross and Peichel 2008)). (B) Alignments of the assembled Y chromosome (left) to the X chromosome (right) reveal the same inversions in the *de novo* assembly. Stratum one is indicated by orange, stratum two is indicated by dark purple, stratum three is indicated by light purple, and the pseudoautosomal region is indicated by yellow. A majority of the pseudoautosomal region is not included in the reference Y chromosome assembly because this region was not targeted (see methods). The location of the candidate sex determination gene *(Amhy)* is indicated by the black arrow. Centromeres are shown by black circles. Positions are shown in megabases.

242 to the Y chromosome (Ross and Peichel 2008), suggesting this region may be largely deleted or
243 highly degenerated. Subsequent work using Illumina short-read sequencing revealed that some
244 genes from this region were still present on the Y chromosome under strong purifying selection,
245 but the location of these genes within the Y could not be determined by mapping reads to the X
246 chromosome (White et al. 2015). The cytogenetic marker, *Idh,* is located at the distal end of our
247 Y chromosome assembly, remarkably consistent with the placement of *Idh* in the cytogenetic map
248 (Ross and Peichel 2008), indicating stratum one is no longer located at the distal end of the Y
249 chromosome as it is on the X chromosome. Instead, we found a high density of stratum one
250 alignments near the boundary of the pseudoautosomal region at the opposite end of the
251 chromosome (Figure 2B). Within this stratum, there was an overall lower density of alignments
252 between the X and Y chromosome, consistent with previous patterns mapping Illumina short
253 reads to the reference X chromosome (White et al. 2015). The placement of stratum one in the
254 assembly was consistent with the hybridization of fluorescent *in situ* hybridization probes,
255 designed from stratum one BAC inserts. These probes clearly hybridized to the chromosome end
256 opposite of *Idh* (Supplemental Figure 4).

257 Because we were primarily focused on sequences that were highly divergent from the X
258 chromosome or absent from the female reference genome entirely, our strategy did not target the
259 pseudoautosomal region for assembly into the Y chromosome. Nevertheless, our assembly did
260 place a small fraction of the ~2.5 Mb pseudoautosomal region on the distal end of the male-
261 specific Y chromosome, adjacent to stratum one. The cytogenetic marker *STN303* was included
262 in this region, which is located on the opposite end of the pseudoautosomal region on the X
263 chromosome (Figure 2). This discordance in marker placement within the pseudoautosomal
264 region likely indicates a mis-assembly of the region. The pseudoautosomal region contains
265 repetitive sequence, complicating overall assembly of the region (see transposable elements
266 section). Indeed, the contigs spanning this region and *STN303* have a smaller size (five contigs;

267 median: 88,098 bp) than the remaining contigs within the Y chromosome or X chromosome,
268 consistent with highly heterozygous, repetitive sequence.
269
270 **The location of centromeric repeats are concordant with a metacentric chromosome**
271 A 186 bp centromeric AT-rich alpha satellite repeat was previously identified in female fish
272 by chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Cech and Peichel 2015).
273 Although this repeat hybridized strongly to autosomes and the X chromosome, there was only
274 weak hybridization of the probe to the Y chromosome, suggesting the Y chromosome might have
275 a divergent centromeric repeat and/or contain substantially less satellite DNA than the autosomes
276 (Cech and Peichel 2015). We used ChIP-seq with the same antibody against centromere protein
277 A (CENP-A) in males to identify any Y chromosome repeats. Relative to the input DNA, we found
278 strong enrichment of reads from the immunoprecipitation mapping to the center of the Y
279 chromosome assembly, indicative of CENP-A binding (Figure 3A; Supplemental Figure 5). The
280 enrichment was located between cytogenetic markers *STN187* and *WT1A*, consistent with the
281 predicted location of the centromere in the cytogenetic map and the metacentric chromosome
282 morphology in karyotypes (Ross and Peichel 2008). These results further confirm the ordering of
283 contigs within our Y chromosome scaffold.
284 Underlying the CENP-A peak, we found a core centromere AT-rich repeat. We identified
285 14 copies of the repeat in our Y chromosome assembly, which shared an average pairwise
286 sequence identity of 84.6% with the core repeat that hybridized to the remainder of the genome
287 (Cech and Peichel 2015) (Supplemental Figure 6). The repeats fell at the edges of a gap,
288 indicating that a majority of the repeats were not assembled into our primary scaffold. Uneven
289 coverage signal in Hi-C libraries from repetitive DNA can trigger the 3D-DNA assembler to remove
290 these regions from contigs during the editing step (Durand et al. 2016; Dudchenko et al. 2017).
291 Consistent with this, both contigs that flanked the centromere gap in the Y chromosome assembly
292 had additional sequence that was removed by the 3D-DNA pipeline as “debris.” The first contig

293 that was adjacent to the gap (contig 11894) contained six copies of the repeat and had an

294 additional 57,692 bp that was removed as “debris.” The second contig on the opposite side of the

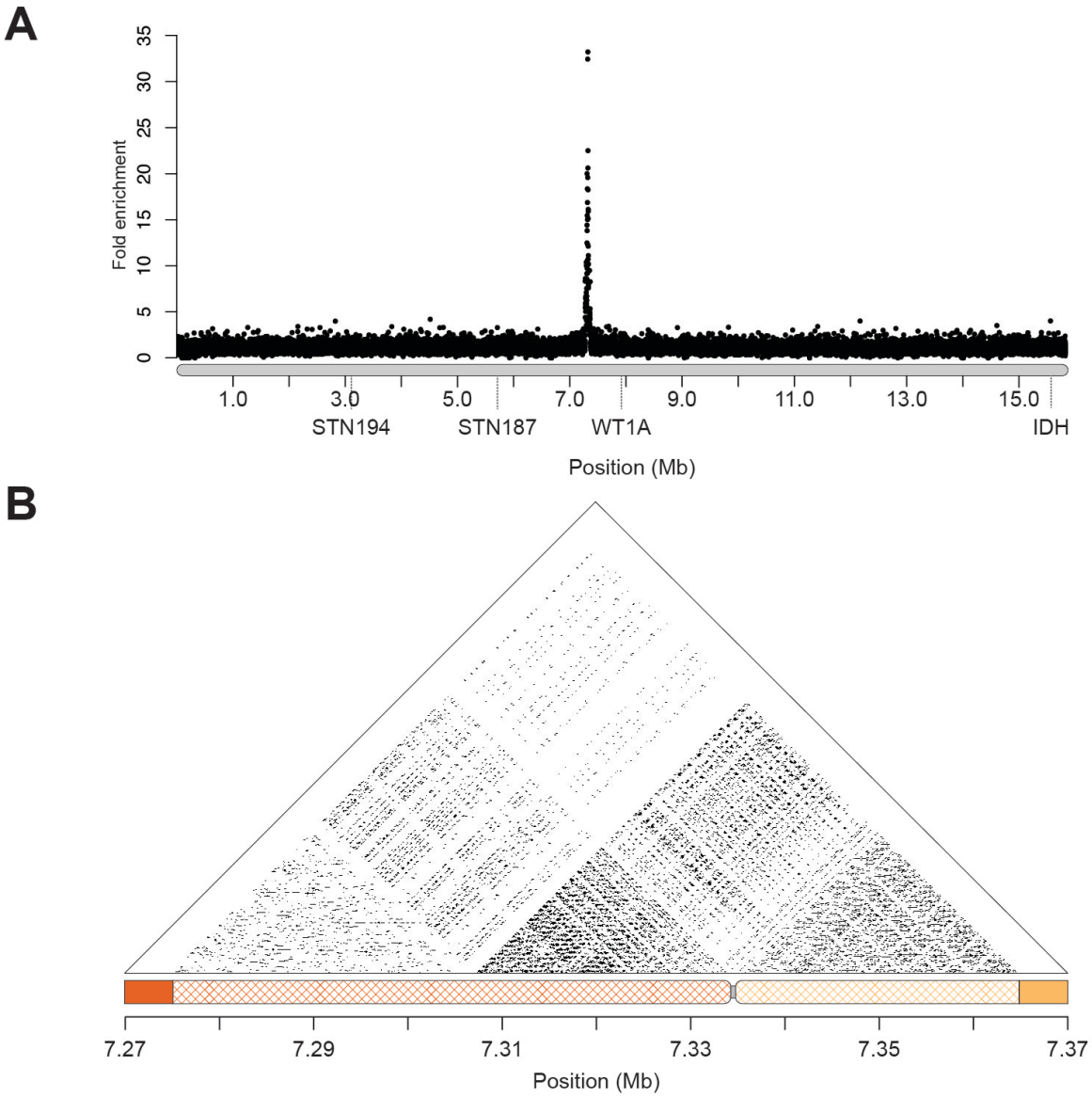

Figure 3. Sequences immunoprecipitated with CENP-A are enriched at the center of the chromosome. (A) Short-read sequences from a chromatin-immunoprecipitation (ChIP-seq) with CENP-A were aligned to the reference Y chromosome assembly. There is a prominent peak between markers STN187 and WT1A where the centromere is located in cytogenetic maps (Ross and Peichel 2008). CENP-A enrichment from a second male fish is shown in Supplemental Figure 5. (B) Alpha satellite monomeric repeats are organized into higher order repeats (HORs). Sequence identity is shown in 100 bp windows across the centromere sequence of the Y chromosome. 87 kb of sequence containing the monomeric repeat was rejoined (crosshatched) to contigs that were previously fragmented in the scaffolding process (orange contig: 11894; yellow contig: 11839). The gap between the two contigs is shown in grey.

295 gap (contig 11839) had eight copies of the repeat and an additional 29,308 bp of sequence that
296 was removed as “debris.” We used BLAST to search for additional repeats in the debris using the
297 majority consensus sequence of the 14 previously identified centromere repeats in the Y
298 assembly. There were an additional 304 repeats in the debris sequence from contig 11894, and
299 163 repeats in the debris sequence from contig 11839. We added the debris sequence back into
300 the total Y chromosome assembly, increasing the assembled centromere size by 87 kb (total Y
301 chromosome length: 15.87 Mb) (Figure 3B). Average pairwise percent sequence identity among
302 all monomeric repeats in the Y chromosome assembly was 89.5%. Compared to the core
303 threespine stickleback centromere repeat previously identified, the Y chromosome centromere
304 repeat was more divergent. Average pairwise percent sequence identity between all the motifs in
305 the Y chromosome assembly and the centromere repeat identified from female fish was only
306 86.8%.

307 Centromeres are often composed of highly similar blocks of monomeric repeats,
308 organized into higher order repeats (HORs) (Alexandrov et al. 1993; McNulty and Sullivan 2018;
309 Hartley and O’Neill 2019). Previous characterization of the monomeric centromere repeat in
310 threespine stickleback did not reveal a HOR organization; however, this analysis was limited by
311 the identification of only a few short stretches of the monomeric repeat on each autosome (Cech
312 and Peichel 2015). The ~87 kb of assembled centromere on the Y shows a clear higher order
313 patterning around the centromeric region, consistent with complex HORs (Figure 3B).

314

315 **The Y chromosome has three evolutionary strata**

316 Previous estimates of synonymous site divergence (d_S) in coding regions have indicated
317 there are two evolutionary strata on the threespine stickleback sex chromosomes (White et al.
318 2015), despite the presence of at least three major inversion events in the cytogenetic map of the
319 sex chromosomes (Ross and Peichel 2008). Because these estimates relied on aligning short-
320 read Illumina sequences to the reference X chromosomes, overall divergence could have been

321 biased by mapping artifacts, especially in the oldest region of the Y chromosome. We investigated
322 whether our Y chromosome assembly supported the earlier model of two evolutionary strata or
323 whether there could be additional strata uncovered in the current *de novo* assembly. We aligned
324 all ENSEMBL predicted X chromosome coding regions outside of the pseudoautosomal region to
325 the Y chromosome reference assembly to estimate divergence. Of the 1187 annotated coding
326 sequences, we were able to align 504 (42.5%) to the male-specific region of the Y chromosome.
327 We found a clear signature of three evolutionary strata, consistent with inversion breakpoints
328 within the cytogenetic map as well as within our *de novo* reference assembly. The oldest stratum

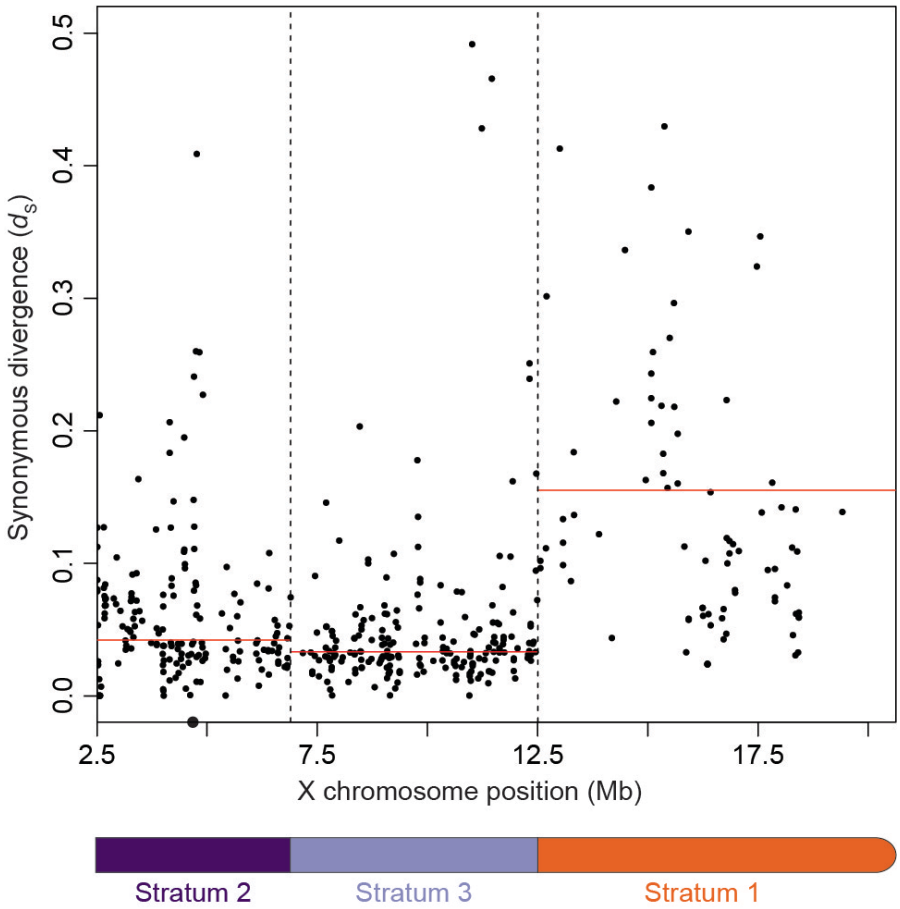

Figure 4. The sex chromosomes have three distinct evolutionary strata. Synonymous divergence (d_S) between the X and Y chromosome was estimated for every annotated transcript on the X chromosome. Genes are ordered by position on the X chromosome (Mb). Median divergence across each region is shown by the red line; values are given in Table 2. Strata breakpoints are indicated by the vertical dashed lines. The centromere is indicated by a black circle.

329 (stratum one) encompassed the same region of the X chromosome as previously described in
330 the Illumina-based study and had highly elevated d_S (stratum one median d_S: 0.155). In contrast
331 to the Illumina-based estimates, our new assembly revealed that the remainder of coding regions
332 across the X chromosome formed two distinct strata, with different estimates of d_S (Figure 4;
333 Table 1). We also investigated whether the older strata had increased non-synonymous
334 divergence (d_N) consistent with inefficient selection from the lack of crossing over between the
335 chromosomes (Charlesworth 1978; Rice 1987). As predicted, stratum one had a significantly
336 higher d_N than strata two and three (Table 1). Stratum two had a significantly lower d_N than the
337 other strata. This was also reflected by a significantly lower d_N/d_S ratio (Table 1), suggesting genes
338 in stratum two are under stronger purifying selection.
339

340 **The Y chromosome is evolving a unique genetic architecture**

341 Haploinsufficient genes have been repeatedly retained on degenerating sex
342 chromosomes of mammals and birds (Bellott et al. 2014; 2017), and may be enriched in stratum
343 one of the stickleback Y chromosome (White et al. 2015). We explored whether our expanded set
344 of annotated genes exhibited signatures of haploinsufficiency by identifying orthologs between
345 the X-annotated genes and human genes ranked for haploinsufficiency (Decipher
346 Haploinsufficiency Predictions (DHP) v. 3) (Firth et al. 2009; Huang et al. 2010). Within strata one
347 and two, we found orthologs with a retained Y-linked allele had lower DHP scores than genes
348 without a Y ortholog, indicating that retained genes were more likely to exhibit haploinsufficiency

Table 1. Nucleotide divergence between X and Y chromosome homologs.

	X genes	Y genes	Percent remaining on Y	d_S	d_N	d_N/d_S
Stratum 1	612	110	18.0%	0.155[a]	0.030[a]	0.287[a]
Stratum 2	243	179	73.7%	0.042[b]	0.008[b]	0.198[b]
Stratum 3	332	215	64.8%	0.033[c]	0.011[c]	0.338[a]

[a,b,c] Groups significantly different by a pairwise Mann-Whitney U Test; $P < 0.05$

349 (Figure 5; Mann-Whitney U Test; stratum one $P < 0.001$; stratum two $P = 0.035$). We found a
350 similar trend for genes retained on the Y chromosome in stratum three, but this result was not
351 significant (Figure 5; Mann-Whitney U Test; $P = 0.085$). Nevertheless, this lower score suggests
352 enrichment for haploinsufficient genes may already be underway within the youngest region of
353 the Y chromosome.

354 Genes can be acquired on the Y chromosome through duplications from autosomes
355 (reviewed in Gvozdev et al. 2005), a process that has had a prominent impact on the overall gene
356 content of ancient sex chromosomes (Saxena et al. 1996; Lahn and Page 1999; Carvalho et al.
357 2001; Skaletsky et al. 2003; Mahajan and Bachtrog 2017; Tobler et al. 2017; Chang and
358 Larracuente 2018), but the overall influence of this process on the genetic architecture of newly
359 evolving sex chromosomes has not been documented. To identify whether the young stickleback

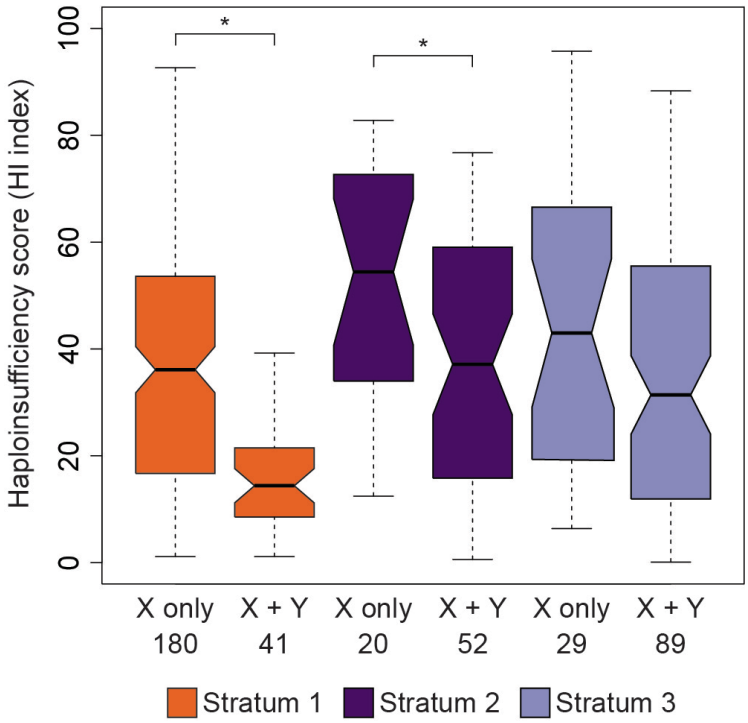

Figure 5. Genes retained on the Y chromosome in strata one and two are more likely to exhibit haploinsufficiency. Human proteins with predicted haploinsufficiency indexes, in which a lower value indicates that a gene is more likely to be haploinsufficient, were matched to one-to-one human-threespine stickleback fish orthologs from the X chromosome. Haploinsufficiency indexes were significantly lower for genes retained on both the X and Y chromosomes than for genes present only on the X chromosome (i.e. lost from the Y chromosome) in both strata one and two. Asterisks indicate $P < 0.05$ (Mann-Whitney U test).

360 Y chromosome also contained genes shared with autosomes, but not the X chromosome, we first
361 used the MAKER gene annotation pipeline (Cantarel et al. 2008; Holt and Yandell 2011) to
362 assemble a complete set of coding regions across the Y chromosome reference sequence. We
363 identified a total of 626 genes across the male-specific region of the Y chromosome, 33 of which
364 had paralogs on autosomes, but not on the X chromosome (5.3%) (Table 2). A majority of these
365 genes (25 of 33; 75.8%) appeared to have undergone duplications within the Y chromosome
366 following translocation from the autosomes (genes had copy numbers ranging from two to six).
367 Gene translocation onto sex chromosomes can occur through RNA-mediated mechanisms
368 (retrogenes) or through DNA-based translocations (reviewed in Long et al. 2013). Of the
369 stickleback genes that had multiple introns within the autosomal homolog (31 of 33 genes), we
370 did not detect a single paralog on the Y chromosome that had a complete loss of introns.

371 Genes that accumulate on Y chromosomes are predicted to have male beneficial
372 functions. On many ancient sex chromosomes, genes that have translocated to the Y
373 chromosome from autosomes exhibit testis-biased expression (Saxena et al. 1996; Lahn and
374 Page 1999; Skaletsky et al. 2003). suggesting important roles in spermatogenesis. To determine
375 whether the translocated genes on the threespine stickleback Y chromosome are enriched for
376 male functions, we looked for testis-biased gene expression between testis tissue and three other
377 tissues (liver, brain, and larvae). Compared with all tissues, we found stronger testis-biased
378 expression among the genes that translocated to the Y chromosome, compared to the single-
379 copy genes with a homolog on the X chromosome (Figure 6; Mann-Whitney U Test; $P < 0.05$).
380 Because DNA-based translocations of genes often contain their native regulatory elements, we

Table 2. Origin of genes in each stratum on the Y chromosome.

	X ancestral, Y single copy	X ancestral, Y duplicated	Autosomal, Y single copy	Autosomal, Y duplicated	Unknown origin, Y single copy	Unknown origin, Y duplicated	Total
Stratum 1	114 (72.6%)	14 (9.0%)	3 (1.9%)	6 (3.8%)	19 (12.1%)	1 (0.6%)	157
Stratum 2	154 (80.6%)	11 (5.8%)	2 (1.0%)	11 (5.8%)	11 (5.8%)	2 (1.0%)	191
Stratum 3	233 (83.8%)	22 (7.9%)	3 (1.1%)	8 (2.9%)	11 (4.0%)	1 (0.3%)	278

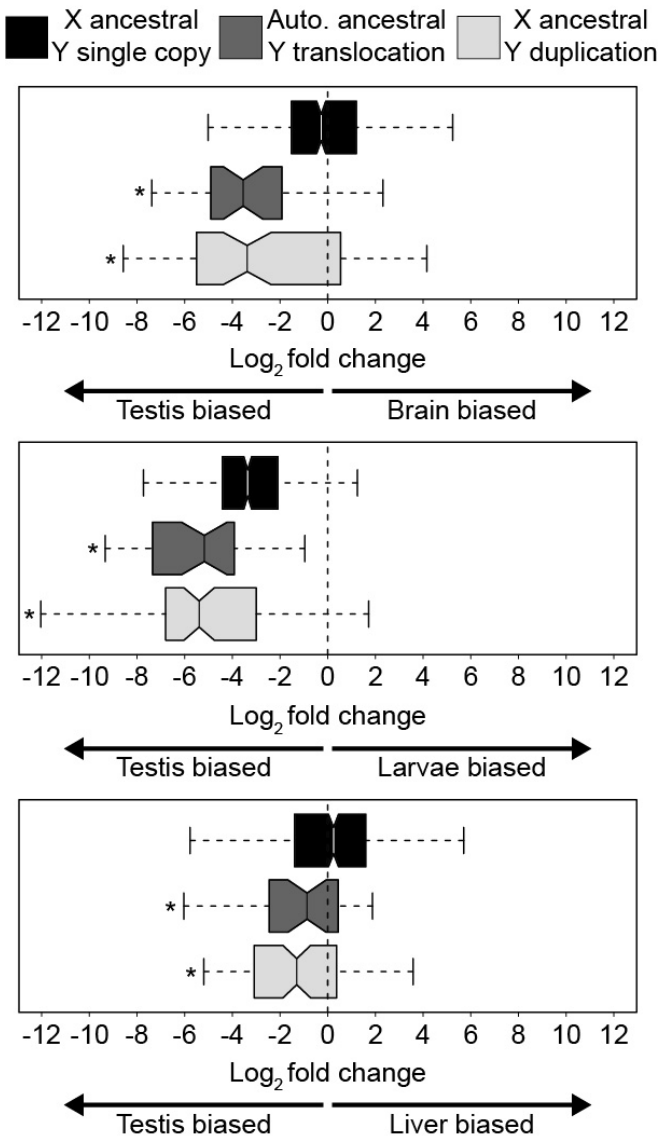

Figure 6. Genes present on the Y that have been translocated from the autosomes or genes that have been duplicated on the Y and are derived from ancestral X-linked homologs show testis-biased gene expression. Log_2 fold change between testis tissue and three other tissues (brain, larvae, and liver) is shown. For each tissue comparison, asterisks denote groups with significantly different expression from single-copy genes present on the Y that are derived from X-linked homologs (Mann-Whitney U Test; $P < 0.05$).

examined whether the autosomal paralogs also exhibited testis-biased expression to a similar degree as the Y-linked paralogs. Consistent with this pattern, we observed a similar degree of testis-biased expression between testis and liver tissue among the ancestral paralogs on the autosomes (median translocated genes Log_2 fold change: -0.867; median ancestral autosomal paralog Log_2 fold change: -1.558; Mann-Whitney U test, $P = 0.818$). This pattern did not hold for

386 comparisons between testis and larvae (Median translocated genes Log_2 fold change: -5.178;
387 Median ancestral autosomal paralog Log_2 fold change: -1.371; Mann-Whitney U Test $P < 0.001$)
388 and testis and brain (Median translocated genes Log_2 fold change: -3.548; Median ancestral
389 autosomal paralog Log_2 fold change: -1.601; Mann-Whitney U Test $P = 0.036$). Combined, our
390 results indicate that the genes which translocated to the Y chromosome and were retained often
391 had testis-biased expression ancestrally.

392 Duplicated genes on the Y chromosome can also be derived from ancestral genes shared
393 between the X and Y. Of the 626 genes annotated across the male-specific region of the Y
394 chromosome, 47 (7.5%) had greater than one copy on the Y chromosome and also had an X-
395 linked allele (Table 2). None of these genes were structured within large amplicons, which are
396 characteristic of many mammalian Y chromosomes (Skaletsky et al. 2003; Hughes et al. 2010;
397 2012; Li et al. 2013; Soh et al. 2014; Skinner et al. 2016; Brashear et al. 2018; Janečka et al.
398 2018). Instead, copy number ranged from two to seven copies total. We explored if this duplicated
399 class of genes also exhibited testis-based expression similar to what we observed with the
400 autosome translocated genes. Consistent with the previous patterns, we found strong testis-
401 biased expression between testis and all other tissues among duplicated genes that have an X-
402 linked homolog (Figure 6; Mann-Whitney U Test; $P < 0.001$, all comparisons). Similar to the
403 ancestral autosomal paralogs, we found that genes often exhibited testis-biased expression
404 ancestrally on the X chromosome before duplicating on the Y (Supplemental Table 2). However,
405 this pattern did not hold in all tissue comparisons. In some cases, genes exhibited stronger testis-
406 biased expression after duplicating on the Y chromosome.

407

408 **Transposable elements have accumulated throughout the Y chromosome**

409 Transposable elements also rapidly accumulate on sex chromosomes once
410 recombination is suppressed (reviewed in Bachtrog 2013). The threespine stickleback Y
411 chromosome has a higher density of transposable elements throughout the male-specific region

412 of the Y chromosome, compared to the X chromosome (Supplemental Figure 7). We found the
413 highest densities within stratum one, consistent with recombination being suppressed in this
414 region for the greatest amount of time. We also found a high density of transposable elements
415 within the recombining pseudoautosomal region (Supplemental Figure 7).
416
417 **Stratum one contains a candidate sex determination gene**
418 The master sex determination gene has not been identified in the threespine stickleback.
419 Although master sex determination genes can be highly variable among species (Bachtrog et al.
420 2014; Capel 2017), many species of fish share some common genes that have been co-opted
421 into this role during the independent evolution of Y chromosomes. For instance, orthologs of both
422 the anti-Müllerian hormone (*Amhy)* (Hattori et al. 2012; Li et al. 2015; Pan et al. 2019) as well as
423 the anti-Müllerian hormone receptor (*Amhr2)* (Kamiya et al. 2012) have been used as the master
424 sex determination gene in multiple species of fish. We searched for evidence of these genes
425 among the annotated transcripts on the Y chromosome. We found the complete coding sequence
426 of the anti-Müllerian hormone on the Y chromosome (hereafter referred to as *Amhy*), located
427 within the oldest stratum adjacent to the pseudoautosomal region boundary (positions 817,433 -
428 821,230) We did not locate an allele on the X chromosome, suggesting *Amhy* is an ancient
429 duplication and translocation from autosome eight.
430 We explored whether *Amhy* had divergence patterns and expression patterns consistent
431 with a functional role in sex determination. We aligned the protein coding sequence of AMHY to
432 the threespine stickleback AMH paralog on autosome eight as well as to other vertebrate AMH
433 proteins. We observed conservation of amino acids in the AMH and TGF-β domains of the protein
434 sequence on the Y chromosome paralog that are conserved across vertebrates (Supplemental
435 Figure 8), suggesting the Y chromosome paralog is under selection in these regions to preserve
436 function. We surveyed expression patterns of *Amhy* across the six tissues used in the gene
437 annotations, including a larval tissue collected around the time sex determination is believed to

438 occur (stages 22-26 (Swarup 1958; Lewis et al. 2008)). *Amhy* expression was significantly higher
439 in larval tissue compared to brain (Log_2 fold change: -2.031; FDR = 0.012), but expression was
440 statistically indistinguishable when compared to testis (Log_2 fold change: -0.284; FDR = 0.918) or
441 liver (Log_2 fold change: -2.054; FDR = 0.052). Additional functional genetics work is currently
442 underway to test if this gene is necessary and sufficient for initiating male development.
443

444 **Discussion**

445 **Evolution of the threespine stickleback Y chromosome**

446 Using a combination of long-read sequencing and chromosome conformation capture (Hi-
447 C) sequencing for scaffolding, we were able to assemble a highly accurate Y chromosome
448 reference assembly for the threespine stickleback, concordant with sequenced BAC inserts and
449 known cytogenetic markers (Ross and Peichel 2008). Our new reference assembly revealed
450 several patterns of sequence evolution that were not accurately resolved using short-read
451 sequencing (White et al. 2015). First, synonymous divergence was underestimated throughout
452 the Y chromosome by relying on single nucleotide polymorphisms ascertained through short-read
453 sequencing. This effect was greatest in the oldest region of the Y chromosome (stratum one).
454 Median d_S was approximately 8.7-fold greater within stratum one when long-read sequences were
455 used. Synonymous divergence was approximately 2.8-fold greater across the younger strata in
456 the new reference assembly compared to the d_S estimates from short-read sequencing. The short-
457 read sequencing was also unable to distinguish two independent strata within this region, likely
458 from an insufficient signal caused by underestimating the true number of SNPs in the region. Our
459 results argue for caution in using short-read sequencing technologies to characterize sex-specific
460 regions of Y or W chromosomes.

461 Divergence times for each of the strata can be approximated based on divergence rates
462 between the threespine stickleback fish and the ninespine stickleback fish (*Pungitius pungitius*),
463 which last shared a common ancestor as many as 26 million years ago (Bell et al. 2009; Ross et

464 al. 2009; Varadharajan et al. 2019). Combined with a mean genome-wide estimate of
465 synonymous divergence between the two species (0.184; Guo et al. 2013), we determined
466 stratum one likely arose less than 21.9 million years ago, close to when the two species diverged.
467 Using the same calibration, stratum two formed less than 5.9 million years ago and stratum three
468 formed less than 4.7 million years ago.

469 Our complete scaffold of the Y chromosome allowed us to refine the previous evolutionary
470 model that suggested the Y chromosome diverged from the X chromosome through three
471 pericentric inversions (Ross and Peichel 2008). Our new model requires four independent
472 inversions to incorporate the placement of stratum one that was unaccounted for in the original
473 cytogenetic map (Supplemental Figure 9). The first inversion in our model involved only stratum
474 one at the end of the sex chromosomes, whereas the final inversion in our model occurred after
475 all the strata had formed, inverting the entire male-specific region of the Y chromosome. This final
476 inversion moved the candidate sex determination gene, *Amhy*, to its current position, adjacent to
477 the pseudoautosomal region.

478

479 **Y chromosome centromere evolution**

480 Due to their highly repetitive nature, centromeric arrays have been challenging to
481 sequence and assemble using traditional approaches. However, long-read technologies have
482 shown recent promise in traversing these inaccessible regions (Jain et al. 2018; Mahajan et al.
483 2018; Bracewell et al. 2019). Using long-read sequencing, we were also able to recover two
484 contigs in our assembly that contained arrays of an alpha satellite monomeric repeat that had
485 sequence similarity to a monomeric repeat isolated from the remainder of the genome (Cech and
486 Peichel 2015). Centromeres across species are highly variable both at the level of the individual
487 monomer and how monomers are organized at a higher level (Henikoff et al. 2001; Malik and
488 Henikoff 2002; Alkan et al. 2011; Melters et al. 2013; McNulty and Sullivan 2018; Hartley and
489 O'Neill 2019). This incredible variability can even occur within species. For example, in humans

490 centromeric HORs are not identical between nonhomologous chromosomes (Manuelidis 1978;
491 Willard 1985), and the Y chromosomes of mouse and humans contain divergent or novel
492 centromeric repeats relative to the autosomes (Wolfe et al. 1985; Pertile et al. 2009; Miga et al.
493 2014). Consistent with these patterns, we observed a decrease in sequence similarity between
494 the Y chromosome monomeric repeat and the consensus repeat identified from the remainder of
495 the threespine stickleback genome (Cech and Peichel 2015). We found the Y chromosome was
496 also ordered into a complex HOR; however, we cannot determine if the structure of the Y
497 chromosome HOR is similar or dissimilar from other threespine stickleback chromosomes. The
498 centromere sequence from other chromosomes is currently limited to short tracts of monomeric
499 repeats (Cech and Peichel 2015).

500 Cytogenetic work has shown the threespine stickleback Y chromosome centromere may
501 contain a divergent satellite repeat relative to the X chromosome and autosomes (Cech and
502 Peichel 2015; 2016). This hypothesis was based on a weak fluorescent *in situ* hybridization signal
503 on the Y chromosome from DNA probes designed from the consensus repeat. Our Y chromosome
504 assembly indicates a mechanism driving this pattern may be the reduced sequence identity
505 shared between the Y chromosome monomeric repeat and the consensus monomeric repeat. An
506 alternative explanation is that the weak hybridization signal is not due to the differences in
507 monomeric repeat sequence, but it is actually caused by a reduction in overall size of the Y
508 chromosome centromere. Although we isolated ~87 kb of centromere sequence, we did not
509 identify a contig that spans the complete centromere, leaving the actual size of the centromere
510 unknown. Additional sequencing work is necessary to test this alternative model.

511

512 **The genetic architecture of the threespine stickleback Y chromosome is rapidly evolving**

513 Despite the young age of the threespine stickleback Y chromosome relative to mammals,
514 we found acquisition of novel genes throughout all strata of the Y chromosome. We did not detect
515 massive amplification of gene families as observed on mammalian sex chromosomes (Skaletsky

516 et al. 2003; Murphy et al. 2006; Hughes et al. 2010; Paria et al. 2011; Soh et al. 2014; Janečka
517 et al. 2018), but many genes that had translocated from the autosomes or were present in the
518 common ancestor of the sex chromosomes had multiple copies on the Y chromosome. The copy
519 numbers we observed are on the same order as the duplicated genes on the sex chromosomes
520 of multiple species of *Drosophila* (Chang and Larracuente 2018; Ellison and Bachtrog 2019). The
521 gene duplications on the threespine stickleback sex chromosomes may reflect selection on the
522 early amplification of genes important for male fertility (Gvozdev et al. 2005) or to prevent
523 degradation by providing a repair template through gene conversion (Rozen et al. 2003; Skaletsky
524 et al. 2003; Backström et al. 2005; Bhowmick et al. 2007; Connallon and Clark 2010; Davis et al.
525 2010; Marais et al. 2010; Hallast et al. 2013; Soh et al. 2014; Skinner et al. 2016; Peneder et al.
526 2017; Trombetta and Cruciani 2017; Chang and Larracuente 2018). Alternatively, the duplications
527 we observe on the threespine stickleback Y chromosome may simply reflect recent translocations
528 and duplications that have yet to degenerate and pseudogenize.

529 Gene expression patterns of duplicated and translocated genes suggest this process is
530 not entirely neutral. We observed strong testis-biased expression among genes that had
531 duplicated and translocated to the Y chromosome, similar to patterns observed on other Y
532 chromosomes (Carvalho et al. 2000; 2001; Skaletsky et al. 2003; Murphy et al. 2006; Hughes et
533 al. 2010; Paria et al. 2011; Soh et al. 2014; Mahajan and Bachtrog 2017; Janečka et al. 2018).
534 Interestingly, we observed multiple ways that testis-biased genes can accumulate on the Y
535 chromosome. For one, many genes exhibit ancestral testis-biased expression. Genes that have
536 translocated from the autosomes to the Y chromosome had a similar degree of testis-biased
537 expression as the ancestral autosomal gene. The X-linked homologs of genes that are duplicating
538 on the Y chromosome also had testis-biased expression ancestrally. This suggests genes can be
539 selected to be retained on the Y chromosome because of existing male-biased expression
540 patterns. Our observations mirror translocations on the ancient human Y chromosome; the
541 amplified *DAZ* genes arose from an autosomal paralog that was expressed in the testis (Saxena

542 et al. 1996). Examples of autosome-derived translocations to the Y chromosome also exist in
543 *Drosophila* and can have ancestral testis-biased functions (Carvalho et al. 2001). On the other
544 hand, we also found that autosome-derived translocated genes evolved stronger testis-biased
545 expression in a tissue specific-context compared to ancestral expression. The variation in testis-
546 biased expression observed among tissue comparisons suggests the acquisition of testis
547 functions for many genes is incomplete. This makes the threespine stickleback Y chromosome a
548 useful system to understand the regulatory changes required for genes to evolve novel functions
549 in the testis.

550 Genes that translocate to the Y chromosome either arise through RNA-mediated
551 mechanisms or through DNA-based translocations (reviewed in Long et al. 2013). Of the
552 translocations we observed, we only detected DNA-based translocations. Work in other species
553 has shown that DNA-based duplications occur more frequently than RNA-mediated mechanisms
554 (Zhang et al. 2010a; 2010b; Long et al. 2013; Chang and Larracuente 2018). Our results support
555 this bias on young sex chromosomes. It is possible that the frequency of DNA-based duplications
556 is even higher on young sex chromosomes compared to ancient sex chromosomes. DNA-based
557 duplications are driven by erroneous double strand break repair. On the ancient sex
558 chromosomes of rodents, double strand break initiation is suppressed on the sex chromosomes
559 of males (Moens et al. 1997; Lange et al. 2016). This would limit the opportunity for DNA-based
560 translocations to occur due to aberrant double strand break repair during meiosis. However, on
561 young sex chromosomes double strand break frequencies may still be occurring at an appreciable
562 frequency. Coupling a diverging Y chromosome with accumulating repetitive DNA would create
563 additional opportunities for double strand break repair through non-allelic processes, increasing
564 the number of duplications and translocations (Sasaki et al. 2010).

565

566 ***Amhy* is a candidate sex determination gene**

567 We identified the *Amhy* gene as a candidate for male sex determination in the threespine
568 stickleback. *Amh* has been co-opted as a male sex determination gene in multiple species of fish
569 (Hattori et al. 2012; Li et al. 2015; Pan et al. 2019). The master sex determination gene is one of
570 the primary genes that initiates evolution of a proto-Y chromosome (reviewed in Bachtrog 2013).
571 Consistent with this, *Amhy* is located in the oldest region of the stickleback Y chromosome
572 (stratum one), adjacent to the pseudoautosomal region. The gene is expressed in developing
573 stickleback larvae, consistent with a role in early sex determination. Finally, amino acids that are
574 highly conserved across vertebrates in the functional domains of the protein are also conserved
575 on the Y chromosome paralog in stickleback fish, suggesting *Amhy* is functional. Based on the
576 known role of AMH signaling in sex determination in other fish, and the location, expression, and
577 sequence of the Y chromosome paralog in stickleback fish, we propose that *Amhy* is the
578 threespine stickleback master sex determination gene. Additional functional genetics work is
579 underway to test this hypothesis.
580
581 **Conclusions**
582 Our threespine stickleback Y chromosome assembly highlights the feasibility of combining
583 PacBio long-read sequencing with Hi-C chromatin conformation scaffolding to generate a high-
584 quality reference Y chromosome assembly. With the reduction in per base pair cost associated
585 with the newest generation of sequencers, the comparative genomics of sex chromosomes will
586 be more accessible. This will be especially useful for taxa like stickleback fish that have multiple
587 independently derived sex chromosome systems among closely related species (Ross et al.
588 2009). This provides a unique opportunity to understand the convergent evolution of sex
589 chromosome structure as well as the diversity of sex determination mechanisms.
590
591 **Materials and Methods**
592 **Ethics statement**

593 All procedures using animals were approved by the Fred Hutchinson Cancer Research
594 Center Institutional Animal Care and Use Committee (protocol 1575), the Veterinary Service of
595 the Department of Agriculture and Nature of the Canton of Bern (protocol BE17/17), and the
596 University of Georgia Animal Care and Use Committee (protocol A2018 10-003-R1).
597
598 **DNA isolation and PacBio Sequencing**
599 Total DNA was isolated from a single adult male threespine stickleback that was the
600 laboratory-reared offspring of wild-caught fish collected from the Paxton Lake benthic population
601 (Texada Island, British Columbia). Nucleated erythrocytes were isolated from blood (extracted by
602 repeated pipetting in bisected tissue with 0.85x SSC buffer). High molecular weight DNA was
603 isolated by centrifuging blood for 2 minutes at 2000 xg, followed by resuspension of cells in 5 ml
604 of 0.85x SSC and 27 µl of 20 µg/ml proteinase K. Cells were lysed by adding 5 mL of 2x SDS
605 buffer (80 mM EDTA, 100mM Tris pH 8.0, 1% SDS), followed by incubation at 55ºC for 2 minutes.
606 DNA was isolated from the lysate by adding 10 mL of buffered phenol/chloroform/isoamyl-alcohol,
607 rotating slowly at room temperature for 30 minutes, followed by centrifuging at 4ºC for 1 minute
608 at 2000 xg. Two further extractions were performed by adding 10 mL of chloroform, rotating slowly
609 at room temperature for 1 hour, followed by centrifuging at 4ºC for 1 minute at 2000 xg. DNA was
610 precipitated using 1 mL of 3M sodium-acetate (pH 6.0) and 10 mL of cold 100% ethanol. The
611 pellet was washed with cold 70% ethanol and resuspended in 100 µl of 10 mM Tris (pH 8.0). DNA
612 quality was assessed on a FEMTO Pulse (Agilent, Santa Clara, CA, USA); the peak size was
613 132,945 bp. Size selection, library preparation and sequencing on a PacBio Sequel platform was
614 conducted at the Next Generation Sequencing Platform at the University of Bern (Bern,
615 Switzerland). 37.69 Gb was sequenced across seven SMRT cells, resulting in approximately
616 75.25x coverage across the genome.
617
618 **PacBio assembly**

619 Canu (v 1.7.1) (Koren et al. 2017) was used to error correct, trim, and assemble the raw
620 PacBio reads into contigs. Default parameters were used except corOutCoverage was increased
621 to 50 (from the default of 40) to target a larger number of reads for assembly of the sex
622 chromosomes (the X and Y chromosomes in males have only half the available read coverage,
623 relative to the autosomes). Increasing corOutCoverage did not decrease the N50 read size for
624 the assembly (default 40x coverage N50: 31,494 bp; 50x coverage N50: 22,133 bp). The Canu
625 assembly was polished using Arrow (v. 2.2.2). Raw PacBio reads were first aligned to the
626 assembled Canu contigs using pbalign (v. 0.3.1) with default parameters. Arrow was run on the
627 subsequent alignment also using default parameters. We identified redundancy between
628 haplotigs of the autosomal contigs by aligning all the contigs to each other using nucmer (Kurtz
629 et al. 2004) and filtering for alignments between contigs that were at least 1 kb in length and had
630 at least 98% sequence identity (to account for the elevated heterozygosity).
631
632 **Hi-C proximity guided assembly**
633 The X and Y chromosomes of threespine stickleback share a considerable amount of
634 sequence homology (White et al. 2015). In order to differentiate X-linked and Y-linked Canu
635 contigs for scaffolding, we aligned the contigs to the revised reference X chromosome sequence
636 (Peichel et al. 2017), using nucmer in the MUMmer package (v. 4.0) (Kurtz et al. 2004). Putative
637 X- and Y-linked contigs were separated by overall sequence identity. Putative X-linked contigs
638 were defined as having more than 25% of the contig length aligned to the reference X
639 chromosome with a sequence identity greater than 96%, whereas putative Y-linked contigs were
640 defined as having a sequence identity with the reference X chromosome of less than 96%. Contigs
641 which had less than 25% of the length aligning to the reference genome or did not align at all
642 were retained as putative Y-linked unique sequence. Selection of the sequence identity threshold
643 was guided by our overall ability to re-assemble the known X chromosome sequence from the set
644 of putative X-linked PacBio Canu contigs. We tested thresholds from 92% sequence identity to

645 98% sequence identity and chose the threshold that resulted in the smallest size difference
646 between the PacBio assembly and the X chromosome sequence from the reference assembly
647 (Peichel et al. 2017) (Supplemental Table 3). We wrote custom Perl scripts to separate the X-
648 and Y-linked contigs.

649 To scaffold the contigs, we used chromosome conformation capture (Hi-C) sequencing
650 and proximity-guided assembly. Hi-C sequencing was previously conducted from a lab-reared
651 adult male also from the Paxton Lake benthic population (Texada Island, British Columbia) (NCBI
652 SRA database: SRP081031) (Peichel et al. 2017). Hi-C reads were aligned to the complete set
653 of contigs from the Canu assembly using Juicer (v. 1.5.6) (Durand et al. 2016a; Dudchenko et al.
654 2017). 3D-DNA (v. 180114) was used to scaffold the putative X- and Y-linked contigs separately
655 (Durand et al. 2016a; Dudchenko et al. 2017). Default parameters were used except for --editor-
656 repeat-coverage, which controls the threshold for repeat coverage during the misjoin detector
657 step. Because Y chromosomes often have more repetitive sequence than the remainder of the
658 genome, we scaffolded the X- and Y-linked contigs using --editor-repeat-coverage thresholds that
659 ranged from 8 to 18. We chose the minimum threshold that resulted in a Y chromosome scaffold
660 that maximized the total number of Y chromosome Sanger sequenced BACs that either aligned
661 concordantly within contigs included in the scaffold or correctly spanned gaps between contigs in
662 the scaffold (--editor-repeat-coverage 11; Supplemental Table 4) (see Alignment of BAC
663 sequences and merging assemblies).

664

665 **BAC isolation and Sanger sequencing**

666 Y-chromosome specific BACs were isolated from the CHORI-215 library (Kingsley et al.
667 2004), which was made from two wild-caught males from the same Paxton Lake benthic
668 population (Texada Island, British Columbia, Canada) used for the PacBio and Hi-C sequencing.
669 The Y-chromosome specific BACs were identified using a variety of approaches. Initially,
670 sequences surrounding known polymorphic markers (*Idh*, *Stn188*, *Stn194*) on linkage group 19

671 were used as probes to screen the CHORI-215 BAC library filters, and putative Y-specific BACs
672 were identified by the presence of a Y-specific allele at that marker (Peichel et al. 2004; Ross and
673 Peichel 2008). In addition, all CHORI-215 BAC end sequences (Kingsley and Peichel 2006) were
674 used in a BLAST (blastn) search of the threespine stickleback genome assembly, which was
675 generated from an XX female (Jones et al. 2012). All BACs for which neither end mapped to the
676 genome or had elevated sequence divergence from the X chromosome were considered as
677 candidate Y-chromosome BACs. These candidate BACs were verified to be Y-specific using
678 fluorescent *in situ* hybridization (FISH) on male metaphase spreads, following previously
679 described protocols (Ross and Peichel 2008; Urton et al. 2011). The hybridizations were
680 performed with CHORI-213 BAC 101E08 (*Idh*), which clearly distinguishes the X and Y
681 chromosomes (Ross and Peichel 2008) labeled with ChromaTide Alexa Fluor 488-5-dUTP, and
682 the putative Y-specific BAC labeled with ChromaTide Alexa Fluor 568-5-dUTP (Invitrogen,
683 Carlsbad, CA, USA). Starting with these verified Y-specific BACs, we then used the CHORI-215
684 BAC end sequences to iteratively perform an *in silico* chromosome walk. At each stage of the
685 walk, BACs were verified as Y-specific using FISH. In total 102 BACs were sequenced.

686 BAC DNA was isolated from a single bacterial colony and purified on a Qiagen MaxiPrep
687 column. DNA was sheared to 3-4kb using Adaptive Focused Acoustics technology (Covaris,
688 Woburn, MA, USA) and cloned into the plasmid vector pIK96 as previously described (Ferris et
689 al. 2010). Universal primers and BigDye Terminator Chemistry (Applied Biosystems) were used
690 for Sanger sequencing randomly selected plasmid subclones to a depth of 10x. The
691 Phred/Phrap/Consed suite of programs were then used for assembling and editing the sequence
692 (Ewing et al. 1998; Ewing and Green 1998; Gordon et al. 1998). After manual inspection of the
693 assembled sequences, finishing was performed both by resequencing plasmid subclones and by
694 walking on plasmid subclones or the BAC clone using custom primers. All finishing reactions
695 were performed using dGTP BigDye Terminator Chemistry (Applied Biosystems, USA). Finished
696 clones contain no gaps and are estimated to contain less than one error per 10,000 bp.

697

698 **Alignment of BAC sequences and merging assemblies**

699 Sequenced BAC inserts were aligned to the scaffolded Y chromosome using nucmer (v.
700 4.0) (Kurtz et al. 2004). A BAC was considered fully concordant with the PacBio Y chromosome
701 scaffold if the following conditions were met: both ends of the alignment were within 1 kb of the
702 actual end of the Sanger sequenced BAC, the full length of the alignment was within 10 kb of the
703 actual length of the Sanger sequenced BAC, and the total alignment shared a sequence identity
704 with the PacBio Y chromosome scaffold of at least 99%. BAC alignments were also identified that
705 spanned gaps between contigs in the scaffold. An alignment that spanned gaps was considered
706 valid if the following conditions were met: both ends of the alignment were within 1 kb of the actual
707 end of the Sanger sequenced BAC, the total alignment length was not greater than the actual
708 length of the Sanger sequenced BAC, and the total alignment shared a sequence identity with
709 the PacBio Y chromosome scaffold of at least 99%. Finally, BACs were identified that extended
710 from contigs into gaps within the scaffold but did not completely bridge the gaps. BACs that
711 extended into gaps were identified if one end of the alignment was within 1 kb of the actual end
712 of the Sanger sequenced BAC, the alignment extended completely to the end of a contig in the
713 scaffold, and the total alignment shared a sequence identity with the PacBio Y chromosome
714 scaffold of at least 99%. We wrote custom Perl scripts to identify concordant BACs, BACs that
715 spanned gaps in the scaffold, and BACs that extended into gaps within the scaffold.

716 Sanger sequenced BACs that spanned gaps and extended into gaps provided additional
717 sequence that was not originally present in the PacBio scaffolded Y chromosome. We merged
718 this additional sequence into the PacBio scaffold using custom Perl scripts. If multiple Sanger
719 sequenced BACs spanned a gap or extended into a gap, the BAC with the highest percent
720 sequence identity was used.

721

722 **Repetitive element annotation**

723 Repetitive elements were first modeled together on the PacBio scaffolded X and Y
724 chromosomes using RepeatModeler (v. 1.0.11) (Smit et al. 2013) with default parameters.
725 Repeats were masked across both scaffolds using RepeatMasker (v. 4.0.7) (Smit et al. 2013) with
726 default parameters and the custom database created by RepeatModeler. Sequence identity
727 among family members within each class of transposable elements was estimated with BLAST.
728 Repeat families were first summarized from the RepeatMasker output using the Perl tool, "one
729 code to find them all" (Bailly-Bechet et al. 2014). Pairwise alignments were then conducted within
730 each chromosome across the genome. For each chromosome, pairwise blastn alignments were
731 conducted for each repeat family. If multiple alignments were found between family members only
732 the alignment with the highest score was retained (the score was calculated as alignment length
733 – number of mismatches – number of gap openings). Percent identity as reported by blastn was
734 used for each retained alignment.
735
736 **Identification of the Y centromere**
737 The Y chromosome centromere was localized using chromatin immunoprecipitation
738 targeting centromere protein A (CENP-A) as previously described (Cech and Peichel 2015).
739 Immunoprecipitated and input DNA from two males from the Japanese Pacific Ocean population
740 (Akkeshi, Japan) were 150-bp paired-end sequenced using an Illumina HiSeq 2500. Reads were
741 quality trimmed with Trimmomatic (v. 0.36) (Bolger et al. 2014) using a sliding window of 4 bases,
742 trimming the remainder of the read when the average quality within a window dropped below 15.
743 Trimmed paired-end reads were aligned to the scaffolded Y chromosome assembly with Bowtie2
744 (v. 2.3.4.1) (Langmead and Salzberg 2012), using default parameters. This resulted in an overall
745 alignment rate of 83.9% (chromatin only input) and 81.0% (immunoprecipitation) for the first male
746 and 82.3% (chromatin only input) and 79.8% (immunoprecipitation) for the second male. We
747 quantified the read depth of aligned reads at every position across the Y chromosome using the
748 genomecov package of BEDTools (v. 2.28.0) (Quinlan and Hall 2010). We calculated fold-

749 enrichment of reads in the immunoprecipitation versus the input DNA at every position across the
750 Y chromosome. Each position was normalized by the total number of reads in the respective
751 sample before calculating the immunoprecipitation to input DNA ratio. The mean fold-enrichment
752 was calculated every 1 kb across the Y chromosome. Fold-enrichment was quantified using a
753 custom Perl script.

754 The autosomal core centromere repeat (GenBank accession KT321856) (Cech and
755 Peichel 2015) was aligned to the Y centromere region using BLAST (blastn) (Altschul et al. 1990).
756 Only hits that had an alignment length ±10 bp of the core 187 bp repeat were retained. Average
757 percent identity was calculated among the remaining BLAST hits. We determined a majority
758 consensus sequence from the core 14 centromere repeat units from the initial Y chromosome
759 assembly. The majority consensus was used to identify additional repeats in the "debris"
760 fragments that flanked the gap in the scaffold where the Y centromere was originally identified.
761 The majority consensus was aligned to the debris fragments using BLAST (blastn), retaining
762 alignments that had an alignment length ±10 bp of the core 187 bp repeat. Pairwise alignments
763 between all repeats within the Y chromosome was conducted using BLAST (blastn). Average
764 percent identity among all pairwise alignments was calculated using a custom Perl script.

765

766 **Molecular evolution of genes on the Y chromosome**

767 To characterize divergence between ancestral genes shared by the X and Y
768 chromosomes, we aligned the coding sequence of each ENSEMBL predicted gene to the Y
769 chromosome using Exonerate (v. 2.4.0) (Slater and Birney 2005) using the parameters --model
770 est2genome --bestn 15. Only coding sequences for which at least 95% of its sequence length
771 aligned to the Y chromosome were retained for further analysis. d_S and d_N were quantified for
772 each pairwise alignment using the codeml module of PAML (phylogenetic analysis by maximum
773 likelihood) (runmode = 2) (Yang 2007). If an X coding sequence aligned to multiple locations on
774 the Y chromosome, only the alignment with the lowest d_S was retained. In addition, all alignments

775 with d_S greater than two were removed. These stringent filtering steps aimed to limit alignments
776 to the true homolog, rather than to distantly related paralogs of genes that are present in greater
777 than one copy on the sex chromosomes. For estimating d_N/d_S, transcripts with a value of 99 were
778 omitted. Strata breakpoints were broadly based upon the inversion breakpoints in the cytogenetic
779 map (Ross and Peichel 2008), adjusted at a fine-scale by the inversion breakpoints in the
780 alignments between the assembled Y chromosome and the reference X chromosome
781 (breakpoints on the Y chromosome: PAR/stratum one: 0.34 Mb; stratum one/stratum two: 4.67
782 Mb; stratum two/stratum three: 9.67 Mb).

783 To characterize whether there were any novel genes acquired by the Y chromosome as
784 well as any duplicated genes, we aligned each MAKER annotated gene on the PacBio assembled
785 Y chromosome to the whole genome as well as back to the Y chromosome, following the same
786 exonerate procedure. If a homolog was identified on an autosome, we only retained the homolog
787 if d_S was lower than the median d_S across the oldest region of the Y chromosome (stratum one:
788 0.101). Using this stringent filter avoids incorrectly assigning ancient paralogs on the autosomes
789 as the homolog. If multiple alignments were identified on the X chromosome, only the alignment
790 with the lowest d_S was retained. If multiple overlapping paralogs from a single gene were identified
791 on the Y chromosome, only the paralog with the lowest d_S was retained. Alignments to the
792 unassigned contigs (ChrUn) were ignored because these contigs cannot be unambiguously
793 assigned to the X chromosome or to the autosomes.

794

795 **Gene annotation across the PacBio assembled Y chromosome**

796 Genes were annotated on the repeat masked Y chromosome scaffold using the MAKER
797 genome annotation pipeline (v. 3.01.02) (Cantarel et al. 2008; Holt and Yandell 2011) using
798 evidence from multiple RNA-seq transcriptomes, all predicted protein sequences from ENSEMBL
799 (release 95), and *ab initio* gene predictions from SNAP (Korf 2004) and Augustus (Stanke et al.
800 2006). RNA-seq from was conducted on multiple tissue samples. RNA from adult male whole

801 brains was previously extracted and sequenced from wild-caught fish from the Japanese Pacific
802 Ocean population, Akkeshi, Japan (NCBI BioProject accession: PRJNA277770) (White et al.
803 2015). Male larval tissue was collected from stages 22-26 (Swarup 1958) when sex determination
804 is believed to occur (Lewis et al. 2008). Larvae were collected from laboratory-reared progeny of
805 wild-caught fish from the Lake Washington population (Seattle, Washington). Larvae were pooled
806 into two samples, one consisting of five males and the other consisting of six males. Total RNA
807 was extracted using TRIzol reagent (Invitrogen, USA) following standard protocols. Library
808 preparation and sequencing was conducted by the Fred Hutchinson Cancer Research Center
809 Genomics Shared Resource. Single-end sequencing was carried out on a Genome Analyzer II
810 for 72 cycles. Liver and testis tissues were also collected from adult and juvenile fish from
811 laboratory-reared progeny of wild-caught fish from the Japanese Pacific Ocean population
812 (Akkeshi, Japan). Livers and testes were collected from two males and pooled. Three juvenile
813 samples and three adult samples were collected. Total RNA was extracted using TRIzol reagent
814 (Invitrogen, USA) following standard protocols. Library preparation and sequencing was
815 conducted by the Georgia Genomics and Bioinformatics Core at the University of Georgia. Paired-
816 end sequencing was carried out on a NextSeq 500 for 150 cycles. All reads were quality trimmed
817 with Trimmomatic (v. 0.36) (Bolger et al. 2014) using a sliding window of 4 bases, trimming the
818 remainder of the read when the average quality within a window dropped below 15.

819 We aligned sequences to the masked revised whole-genome reference assembly (Peichel
820 et al. 2017) using Tophat (v. 2.3.4.1) (Kim et al. 2013). Default parameters were used except for
821 the liver and testis tissues. For these tissues, we used --read-mismatches 4 and --read-edit-dist
822 4 to account for the greater number of SNPs in the 150 bp reads. These alignment parameters
823 produced an overall alignment rate to the masked genome of 80.4% for the brain tissue, 68.0%
824 in the adult liver tissue, 64.5% in the juvenile liver tissue, 64.7% for adult testis tissue, 65.5% for
825 the juvenile testis tissue, and 68.9% for the larval tissue. Aligned reads from all samples within a
826 tissue were pooled to construct a single tissue-specific set of transcripts using Cufflinks (v. 2.2.1)

827 (Roberts et al. 2011) with default parameters. Exons from the GTF file were converted to FASTA
828 sequences with gffread.
829 MAKER was run over three rounds. For the first round of MAKER, we only used evidence
830 from the RNA-seq transcripts and all annotated protein sequences from ENSEMBL (release 95)
831 using default parameters and est2genome=1, protein2genome=1 to infer gene predictions directly
832 from the transcripts and protein sequences. We used these gene models to train SNAP. In
833 addition, Augustus was trained using gene models from BUSCO conserved orthologs found on
834 the PacBio scaffolded Y chromosome and the revised reference assembly (Glazer et al. 2015;
835 Peichel et al. 2017) with the Actinopterygii dataset and default BUSCO (v. 3.0.2) parameters
836 (Simão et al. 2015; Waterhouse et al. 2017). MAKER was run using the new SNAP and Augustus
837 models with est2genome=0 and protein2genome=0. For the third round of MAKER, SNAP was
838 retrained with the updated gene models and MAKER was run again with the updated SNAP
839 model, the previous Augustus model, and est2genome=0 and protein2genome=0. The threespine
840 stickleback repeat library derived using RepeatModeler was used during the annotation pipeline
841 using the rmlib option.
842
843 **Identification of haploinsufficient genes**
844 One-to-one human-threespine stickleback fish orthologs were identified from the
845 ENSEMBL species comparison database. Orthologs were restricted to those with a human
846 orthology confidence of 1. The high confidence orthologs were matched to the human
847 haploinsufficiency predictions from the DECIPHER database (v. 3) (Firth et al. 2009; Huang et al.
848 2010).
849
850 **Differential expression of genes on the Y chromosome**
851 For each tissue used in the gene annotations, the total number of RNA reads that mapped
852 to the reference Y chromosome was counted using htseq-count (HTSeq software package; v.

853 0.9.1) (Anders et al. 2015). Read counts were obtained across all 626 MAKER identified genes
854 across the male-specific region of the Y chromosome plus all additional paralogs (132 paralogs).
855 Default parameters were used with the addition of --stranded=no and --nonunique all. Ambiguous
856 reads were included in the counts because of the large number of paralogs on the Y chromosome
857 with high sequence identity. In the case a read could not be unambiguously mapped it was
858 assigned to all features to which it matched. Genes were removed from the analysis if they had a
859 read count of zero in all samples. Scaling factors for normalization were calculated using the
860 trimmed mean of M-values (TMM) method in the Bioconductor package, edgeR (Robinson et al.
861 2010). The TMM method minimizes the log-fold changes between samples for most genes. This
862 approach may not be appropriate for a Y chromosome, which is expected to be enriched for male-
863 biased gene expression. Therefore, we calculated scaling factors for all autosomal transcripts and
864 normalized the Y chromosome transcripts using these scaling factors. ENSEMBL annotated
865 transcripts were used for the autosomes. Replicates were grouped based on tissue (testis: six
866 samples; liver: six samples; brain: three samples; larvae: two samples). Log_2 fold-change was
867 calculated for each gene in each tissue comparison using edgeR.
868
869 **Characterization of *Amhy***
870 The protein sequence of AMHY was aligned to AMH sequences from human (GenBank
871 AAH49194.1), mouse (GenBank NP_031471.2), chicken (GenBank NP_990361.1), zebrafish
872 (GenBank AAX81416.1), and the paralog of AMH on threespine stickleback chromosome eight
873 (ENSGACP00000016697) using CLUSTALW with default parameters in Geneious Prime (v.
874 2019.1.1) (https://www.geneious.com). Gene expression level was quantified for *Amhy* in the six
875 different tissues used for gene annotation. Read counts per million (CPM) for each tissue was
876 calculated from the TMM-scaled samples from the differential expression analysis.
877
878 **Data Access**

879 All raw sequencing data and the Y chromosome reference sequence generated in this
880 study have been submitted to the NCBI BioProject database
881 (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA591630.
882

883 **Acknowledgements**

884 The authors acknowledge the following funding sources: NIH R01 GM071854 (CLP), NIH
885 P50 HG002568 (RMM, DMK, CLP), NIH R01 GM116853 (CLP), Fred Hutchinson Cancer
886 Research Center (CLP), Swiss National Science Foundation 31003A_176130 (CLP),
887 Evolutionary, Ecological, or Conservation Genomics Research Award from the American Genetic
888 Association (MAW), Howard Hughes Medical Institute and the Life Sciences Research
889 Foundation (MAW), NIH Cell and Molecular Biology Training Grant T32 GM07270 (JAR), NIH
890 Genome Training Grant T32 HG00035 (JRU), NIH Chromosome Metabolism and Cancer Training
891 Grant T32 CA009657 (JNC), National Science Foundation Graduate Research Fellowship DGE
892 1256082 (JNC), Office of the Vice President of Research at the University of Georgia (MAW), and
893 National Science Foundation IOS 1645170 (MAW). The authors also acknowledge Ivan Liachko
894 and Shawn Sullivan (Phase Genomics) for assistance with Hi-C scaffolding, and the following
895 core facilities for sequencing: Fred Hutchinson Cancer Research Center Genomics Shared
896 Resource, Next Generation Sequencing Platform of the University of Bern, The Georgia
897 Genomics and Bioinformatics Core at the University of Georgia. Arne Nolte, Joana Meier, and
898 Marcel Häsler provided advice on the isolation of high molecular weight DNA. Amanda Bruner
899 provided guidance in staging the threespine stickleback embryos.
900

901 **Disclosure Declaration**

902 The authors have no conflicts of interest to disclose.

References

Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB. 1993. Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization. *Nucleic Acids Res* **21**: 2209–2215.

Alkan CC, Cardone MM, Catacchio CC, Antonacci FF, O'Brien SJ, Ryder OA, Purgato SS, Zoli MM, Valle GG, Eichler EE, et al. 2011. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Res* **21**: 137-145.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Anders S, Pyl P, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics* **14**: 113-124.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J, Mayrose I, Ming R, et al. 2014. Sex determination: why so many ways of doing it? *PLoS Biology* **12**: e1001899.

Backström N, Ceplitis H, Berlin S, Ellegren H. 2005. Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. *Mol Biol Evol* **22**: 1992-1999.

Bailly-Bechet M, Haudry A, Lerat E. 2014. "One code to find them all": a Perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**: 13.

Bell M, Stewart J, Park P. 2009. The world's oldest fossil threespine stickleback fish. *Copeia* **2**: 256-265.

Bellott DW, Bellott DW, Hughes JF, Hughes JF, Skaletsky H, Skaletsky H, Brown LG, Brown LG, Pyntikova T, Pyntikova T, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**: 494-499.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nature Genetics* **49**: 387-394.

Bhowmick B, Satta Y, Takahata N. 2007. The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Res* **17**: 441-450.

Blackmon H, Demuth JP. 2014. Estimating Tempo and Mode of Y Chromosome Turnover: Explaining Y Chromosome Loss With the Fragile Y Hypothesis. *Genetics* **197**: 561-572.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bracewell R, Chatla K, Nalley MJ, Bachtrog D. 2019. Dynamic turnover of centromeres drives karyotype evolution in Drosophila. *Elife* **8**: e49002.

Brashear WA, Raudsepp T, Murphy WJ. 2018. Evolutionary conservation of Y Chromosome ampliconic gene families despite extensive structural variation. *Genome Res* **28**: 1841-1851.

Cantarel B, Cantarel BL, Korf I, Korf I, Robb S, Robb SM, Parra G, Parra G, Ross E, Ross E, et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**: 188-196.

Capel B. 2017. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nature Reviews Genetics* **18**: 675-689.

Carvalho A, Dobo BA, Vibranovski MD, Clark AG. 2001. Identification of five new genes on the Y chromosome of Drosophila melanogaster. *Proc National Acad Sci* **98**: 13225–13230.

Carvalho A, Lazzaro BP, Clark AG. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. *Proc National Acad Sci* **97**: 13239–13244.

Cech JN, Peichel CL. 2016. Centromere inactivation on a neo-Y fusion chromosome in threespine stickleback fish. *Chromosome Res* **24**: 437–450.

Cech JN, Peichel CL. 2015. Identification of the centromeric repeat in the threespine stickleback fish (Gasterosteus aculeatus). *Chromosome Research* **23**: 767-779.

Chang C-H, Larracuente AM. 2018. Heterochromatin-enriched assemblies reveal the Sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* **211**: 333-348.

Charlesworth B. 1978. Model for evolution of Y chromosomes and dosage compensation. *Proc National Acad Sci* **75**: 5618-5622.

Connallon T, Clark AG. 2010. Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186**: 277-286.

Cortez D, Cortez D, Marin R, Marin R, Toledo-Flores D, Toledo-Flores D, Froidevaux L, Froidevaux L, Liechti A, Liechti A, et al. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488-493.

Davis JK, Thomas PJ, Program N, Thomas JW. 2010. A W-linked palindrome and gene conversion in New World sparrows and blackbirds. *Chromosome Research* **18**: 543-553.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden A, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**: 92–95.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden E. 2016a. Juicebox provides a visualization system for Hi-C contact Maps with unlimited zoom. *Cell Syst* **3**: 99-101.

Durand NC, Shamim MS, Machol I, Rao S, Huntley MH, Lander ES, Aiden E. 2016b. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**: 95–98.

Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y chromosomes. *PLOS Genetics* **15**: e1008251.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8**: 186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* **8**: 175–185.

Ferris P, Olson BJ, Hoff PL, Douglass S, Casero D, Prochnik S, Geng S, Rai R, Grimwood J, Schmutz J, et al. 2010. Evolution of an expanded sex-determining locus in *Volvox*. *Science* **328**: 351–354.

Firth HV, Richards SM, Bevan PA, Clayton S, Corpas M, Rajan D, Vooren S, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genetics* **84**: 524–533.

Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT. 2015. Genome assembly improvement and mapping convergently evolved skeletal traits in sticklebacks with genotyping-by-sequencing. *G3* **5**: 1463-1472.

Gordon D, Abajian C, Green P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res* **8**: 195–202.

Griffin DK. 2012. Is the Y chromosome disappearing?—Both sides of the argument. *Chromosome Res* **20**: 35–45.

Guo B, Chain FJ, Bornberg-Bauer E, Leder EH, Merilä J. 2013. Genomic divergence between nine- and three-spined sticklebacks. *BMC Genomics* **14**: 756.

Gvozdev VA, Kogan GL, Usakin LA. 2005. The Y chromosome as a target for acquired and amplified genetic material in evolution. *Bioessays* **27**: 1256–1262.

Hallast P, Balaresque P, Bowden GR, Ballereau S, Jobling MA. 2013. Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. *PLoS Genetics* **9**: e1003666.

Hartley G, O'Neill RJ. 2019. Centromere repeats: hidden gems of the genome. *Genes* **10**: 223.

Hattori RS, Murai Y, Oura M, Masuda S, Majhi SK, Sakamoto T, Fernandino JI, moza G, Yokota M, Strüssmann CA. 2012. A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination. *Proc National Acad Sci* **109**: 2955-2959.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**: 1098-1102.

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

Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics* **6**: e1001154.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82-86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536-539.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**: 321.

Janečka JE, Davis BW, Ghosh S, Paria N, Das PJ, Orlando L, Schubert M, Nielsen MK, Stout TA, Brashear W, et al. 2018. Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. *Nat Communications* **9**: 2945.

Jeffries DL, Lavanchy G, Sermier R, dl MJ, Miura I, Borzée A, Barrow LN, Canestrelli D, Crochet P-A, Dufresnes C, et al. 2018. A rapid rate of sex-chromosome turnover and non-random transitions in true frogs. *Nature Communications* **9**: 4088.

Jones FC, Grabherr MG, Chan Y, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55-61.

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

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**: R36.

Kingsley D, Zhu B, Osoegawa K, Jong PJ, Schein J, Marra M, Peichel CL, Amemiya C, Schluter D, Balabhadra S, et al. 2004. New genomic tools for molecular studies of evolutionary change in threespine sticklebacks. *Behaviour* **141**: 1331-1344.

Kingsley DM, Peichel CL. 2006. The molecular genetics of evolutionary change in sticklebacks. *Biology of the Three-Spined Stickleback* 41–81.

Kitano J, Peichel CL. 2012. Turnover of sex chromosomes and speciation in fishes. *Environ Biol Fish* **94**: 549–558.

Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, Absher DM, Grimwood J, Schmutz J, Myers RM, et al. 2009. A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**: 1079-1083.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Korf I. 2004. Gene finding in novel genomes. *Bmc Bioinformatics* **5**: 59.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Lahn BT, Page DC. 1999. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat Genet* **21**: 429-433.

Lange J, Yamada S, Tischfield SE, Pan J, Kim S, Zhu X, Socci ND, Jasin M, Keeney S. 2016. The landscape of mouse meiotic double-strand break formation, processing, and repair. *Cell* **167**: 695-708.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357-359.

Lewis ZR, McClellan MC, Postlethwait JH, Cresko WA, Kaplan RH. 2008. Female-specific increase in primordial germ cells marks sex differentiation in threespine stickleback (*Gasterosteus aculeatus*). *Journal of morphology* **269**: 909-921.

Li G, Davis BW, Raudsepp T, Wilkerson AJ, Mason VC, Ferguson-Smith M, O'Brien PC, Waters PD, Murphy WJ. 2013. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res* **23**: 1486-1495.

Li M, Sun Y, Zhao J, Shi H, Zeng S, Ye K, Jiang D, Zhou L, Sun L, Tao W, et al. 2015. A Tandem duplicate of anti-müllerian hormone with a missense SNP on the Y chromosome Is essential for male sex determination in Nile tilapia, *Oreochromis niloticus*. *Plos Genet* **11**: e1005678.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annual review of genetics* **47**: 307–33.

Mahajan S, Bachtrog D. 2017. Convergent evolution of Y chromosome gene content in flies. *Nature communications* **8**: 785.

Mahajan S, Wei K, Nalley MJ, Gibilisco L, Bachtrog D. 2018. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLOS Biology* **16**: e2006348.

Malik HS, Henikoff S. 2002. Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev* **12**: 711–718.

Manuelidis L. 1978. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66**: 23–32.

Marais GA, Marais G, Campos P, Campos PR, Gordo I, Gordo I. 2010. Can intra-Y gene conversion oppose the degeneration of the human Y chromosome? A simulation study. *Genome biology and evolution* **2**: 347-357.

McNulty SM, Sullivan BA. 2018. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res* **26**: 115–138.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby GJ, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* **14**: R10.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent JW. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res* **24**: 697–707.

Moens P, Chen D, Shen Z, Kolas N, Tarsounas M, Heng H, Spyropoulos B. 1997. Rad51 immunocytology in rat and mouse spermatocytes and oocytes. *Chromosoma* **106**: 207-215.

Murphy WJ, Wilkerson PA, Raudsepp T, Agarwala R, Schäffer AA, Stanyon R, Chowdhary BP. 2006. Novel gene acquisition on carnivore Y chromosomes. *Plos Genet* **2**: e43.

Myosho T, Takehana Y, Hamaguchi S, Sakaizumi M. 2015. Turnover of sex chromosomes in celebensis group medaka fishes. *G3* **5**: 2685-2691.

Pan Q, Feron R, Yano A, Guyomard R, Jouanno E, Vigouroux E, Wen M, Busnel J-M, Bobe J, Concordet J-P, et al. 2019. Identification of the master sex determining gene in Northern pike (Esox lucius) reveals restricted sex chromosome differentiation. *Plos Genet* **15**: e1008013.

Papadopulos AS, Chester M, Ridout K, Filatov DA. 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*. **112**: 13021-13026.

Paria N, Raudsepp T, Wilkerson AJ, O'Brien PC, Ferguson-Smith MA, Love CC, Arnold C, Rakestraw P, Murphy WJ, Chowdhary BP. 2011. A gene catalogue of the euchromatic male-specific region of the horse Y chromosome: comparison with human and other mammals. *Plos One* **6**: e21374.

Peichel CL, Ross JA, Matson CK, Dickson M, Grimwood J, Schmutz J, Myers RM, Mori S, Schluter D, Kingsley DM. 2004. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. **14**: 1416 -1424.

Peichel CL, Sullivan ST, Liachko I, White MA. 2017. Improvement of the threespine stickleback genome using a Hi-C-based proximity-guided assembly. *Journal of Heredity* **108**: 693-700.

Peneder P, Wallner B, Vogl C. 2017. Exchange of genetic information between therian X and Y chromosome gametologs in old evolutionary strata. *Ecol Evol* **7**: 8478-8487.

Pertile MD, Graham AN, Choo KH, Kalitsis P. 2009. Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res* **19**: 2202–2213.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
Rice W. 1987. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116**: 161-167.

Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325–2329.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome-patterns and consequences. *Molecular Ecology* **22**: 3014-3027.

Ross JA, Peichel CL. 2008. Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. *Genetics* **179**: 2173-2182.

Ross JA, Urton JR, Boland J, Shapiro MD, Peichel CL. 2009. Turnover of sex chromosomes in the stickleback fishes (gasterosteidae). *PLoS Genetics* **5**: e1000391.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**: 873-876.

Sasaki M, Lange J, Keeney S. 2010. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Bio* **11**: 182-195.

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve M, Reijo R, Rozen S, Dinulos M, Disteche CM, et al. 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* **14**: 292.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825-837.

Skinner BM, Sargent CA, Churcher C, Hunt T, Herrero J, Loveland JE, Dunn M, Louzada S, Fu B, Chow W, et al. 2016. The pig X and Y Chromosomes: structure, sequence, and evolution. *Genome Res* **26**: 130–139.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Soh SY, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**: 800-813.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7 Suppl 1**: S11.1 8.

Swarup H. 1958. Stages in the development of the stickleback *Gasterosteus aculeatus* (L.). *Journal Of Embryology And Experimental Morphology* **6**: 373-383.

Tobler R, Nolte V, Schlötterer C. 2017. High rate of translocation-based gene birth on the Drosophila Y chromosome. *Proc National Acad Sci* **114**: 11721–11726.

Trombetta B, Cruciani F. 2017. Y chromosome palindromes and gene conversion. *Human genetics* **136**: 605-619.

Urton J, McCann S, Peichel C. 2011. Karyotype differentiation between two stickleback species (Gasterosteidae). *Cytogenet Genome Res* **135**: 150-159.

Varadharajan, Rastas P, Löytynoja A, Matschiner M, Calboli FC, Guo B, Nederbragt AJ, Jakobsen KS, Merilä J. 2019. A high-quality assembly of the nine-spined stickleback (*Pungitius pungitius*) genome. *Genome Biol Evol* **11**: 3291–3308.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**: 543-548.

White MA, Kitano J, Peichel CL. 2015. Purifying selection maintains dosage-sensitive genes during degeneration of the threespine stickleback Y chromosome. *Mol Biol Evol* **32**: 1981-1995.

Willard H. 1985. Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet* **37**: 524–32.

Wolfe J, Darling SM, Erickson RP, Craig IW, Buckle VJ, Rigby PWJ, Willard HF, Goodfellow PN. 1985. Isolation and characterization of an alphoid centromeric repeat family from the human Y chromosome. *J Mol Biol* **182**: 477–485.

Zhang Y, Vibranovski, Krinsky B, Long M. 2010a. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res* **20**: 1526-1533.

Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010b. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biology* **8**: e1000494.

Zhou Q, Zhou Q, Zhang J, Zhang J, Bachtrog D, Bachtrog D, An N, An N, Huang Q, Huang Q, et al. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**: 1246338-1246338.