

1 **High-quality chromosome-level genomes of two tilapia species reveal their evolution of**
2 **repeat sequences and sex chromosomes**

3 Wenjing Tao^{1,†}, Luohao Xu^{2,3,†,*}, Lin Zhao¹, Zexian Zhu², Xin Wu¹, Qianwen Min¹, Deshou
4 Wang^{1,*}, Qi Zhou^{2,3,4,*}

5

6 1. Key Laboratory of Freshwater Fish Reproduction and Development (Ministry of Education),
7 Key Laboratory of Aquatic Science of Chongqing, School of Life Sciences, Southwest
8 University, Chongqing, 400715, China.

9 2. MOE Laboratory of Biosystems Homeostasis & Protection, Life Sciences Institute, Zhejiang
10 University, Hangzhou, 310058, China

11 3. Department of Molecular Evolution and Development, University of Vienna, Vienna, 1090,
12 Austria

13 4. Center for Reproductive Medicine, The 2nd Affiliated Hospital, School of Medicine, Zhejiang
14 University, Hangzhou, 310058, China

15

16 [†]Contributed equally

17 *Corresponding authors: Q.Z.: zhouqi1982@zju.edu.cn, or D.W.: wdeshou@swu.edu.cn, or
18 L.X.: luohaox@gmail.com

19 **Abstract**

20 **Background**

21 Tilapias are one of the most farmed fishes that are coined as ‘aquatic chicken’ by the
22 food industry. Like many other teleosts, Nile tilapia and blue tilapia exhibit very recent
23 transition of sex chromosome systems since their divergence about 5 million years ago,
24 making them a great model for elucidating the molecular and evolutionary mechanisms
25 of sex chromosome turnovers. Studies into their sex-determining pathways are also
26 critical for developing genetic sex control in aquaculture.

27 **Results**

28 We report here the newly produced genomes of Nile tilapia and blue tilapia that
29 integrate long-read sequencing and chromatin conformation data. The two nearly
30 complete genomes have anchored over 97% of the sequences into linkage groups
31 (LGs), and assembled majorities of complex repetitive regions including telomeres,
32 centromeres and rDNA clusters. In particular, we inferred two episodes of repeat
33 expansion at LG3 respectively in the ancestor of cichlids and that of tilapias. The
34 consequential large heterochromatic region concentrated at one end of LG3 comprises
35 tandem arrays of mRNA and small RNA genes, among which we have identified a
36 candidate female determining gene *Paics* in blue tilapia. *Paics* show female-specific
37 patterns of single-nucleotide variants, copy numbers and expression patterns in gonads
38 during early gonadogenesis.

39 **Conclusions**

40 Our work provide a very important genomic resource for functional studies of cichlids,
41 and suggested that unequal distribution of repeat content that impacts the local
42 recombination rate might make some chromosomes more likely to become sex
43 chromosomes.

44 **Introduction**

45 Tilapias belong to the largest vertebrate family of African cichlids (about 3000 species,
46 order Perciformes) that underwent explosive speciation within the last 10 million years
47 (MY) [1-3]. While two thirds of the cichlids are mainly endemic in lakes of East Africa,
48 and are used as the textbook model for studying mechanisms of sympatric speciation;
49 various tilapia species successfully colonized a much wider range of habitats and have
50 become some of the most important aquaculture species. In particular, the earliest
51 record of raising Nile tilapia (*Oreochromis niloticus*, *ON*) can be dated back to Ancient
52 Egypt. Now it is projected to soon overtake carp and salmon as the most important
53 farmed fish. A second popular tilapia species, blue tilapia (*Oreochromis aureus*, *OA*)
54 diverged from *ON* less than 5 MY ago [4], and has a better cold and saline tolerance,
55 thus is frequently used to produce hybrids with *ON*. Tilapia species from the
56 *Oreochromis* and *Sarotherodon* genera, and many East African cichlids are
57 mouthbrooders [5]. That is, females undergo periods of fasting when brooding the eggs,
58 and sometimes even caring for the fry for extended time. Such a tremendous energy
59 cost of females is one of the major causes that render the larger-sized males the
60 favored sex in tilapia aquaculture. The current predominant practice of sex control in
61 tilapia production is to use the cost-effective hormones rather than adjusting the
62 temperature or population density to induce sex reversal, despite the potential risks to
63 the consumers and the environment [6, 7]. This is mainly due to the lack of detailed
64 knowledge about the genetic sex determining (GSD) pathways of tilapia species.

65 There is a strong and persistent interest in studying the tilapia SD mechanisms
66 and sex chromosomes, in order to produce all-male fingerlings, and also to use tilapias
67 as a model to unravel the molecular and evolutionary mechanisms of vertebrate sex
68 chromosome turnovers [8-10]. In contrast to the conserved and stable sex
69 chromosomes within mammals, birds or *Drosophila*, teleost fish harbor a remarkable
70 diversity of male heterogametic (XY, like that of mammals), female heterogametic (ZW,
71 like that of birds), and environmental SD (ESD) mechanisms frequently between sister
72 species [11-13]. Fish sex chromosomes also do not usually exhibit a high degree of
73 differentiation [13-15], which hampers the identification of the sex chromosomes or the

74 exact SD region cytologically. Some species like *ON* combine both GSD and ESD,
75 suggesting sex in these species is a threshold trait that can be determined by genetic
76 and environmental factors [9]. Despite the complexity of SD systems, and a lack of
77 abundant genomic resources and functional genetic tools until very recently, there have
78 been great efforts of mapping the SD regions among the tilapia species. Early
79 inspection of synaptonemal complex speculated that a large pair of chromosomes
80 corresponding to linkage group 3 (LG3) with incomplete pairing at its terminals maybe
81 the XY chromosome pair of *ON* [16-19]. However, genetic mapping using various types
82 of markers (e.g., microsatellites) indicated that another chromosome LG1 carries an
83 unknown male SD gene as an XY system [20, 21]. The SD region was recently
84 narrowed down into a 9Mb region, through mapping the Illumina reads of both sexes
85 against a high-quality LG1 sequence generated by PacBio reads from a female
86 Egyptian strain of *ON* (*ONEg*). Similarly, by mapping the reads of *OA*, the SD region
87 was inferred to span 50Mb of LG3, as a ZW system [22]. The rapid transition of sex
88 chromosome system between the two species *OA* and *ON* occurred within only 5 MY.
89 More strikingly, another study has mapped the male SD gene in a Japanese strain of
90 *ON* (*ONJp*) onto LG23 rather than LG1 [23-26]. The Y-linked male SD gene is a
91 duplicated copy of *anti-Mullerian hormone* (*Amhy*), and its disruption by CRISPR/Cas9
92 causes male-to-female sex reversal [25]. This is the first functionally validated SD gene
93 of cichlids, and has demonstrated a probably even more recent turnover of SD genes
94 between tilapias.

95 We present here the chromosome-level genome assemblies and comparative
96 analyses of *ONJp* and *OA* vs. the Lake Malawi cichlid species *Metriaclima zebra* (*MZ*).
97 Besides their aquaculture significance, *ON* has been used as the outgroup for studying
98 genomic mechanisms of cichlid species radiation [3]. Moreover, *ONJp* is the first cichlid
99 stock on which transgenics and gene-editing have been successfully conducted [25, 27,
100 28], with the demonstrated potentials for future functional studies of cichlids. We
101 harnessed the single-molecule real-time sequencing technology and produced highly
102 accurate and continuous assemblies of the homogametic sexes of *ONJp* and *OA*,
103 covering their rDNA clusters and centromeric regions. By incorporating chromatin

104 conformation (Hi-C) data, we further anchored over 97% genome sequences of each
105 species into linkage groups, in particular a large heterochromatic region of small RNA
106 gene clusters located at the terminal of LG3. Finally, we narrowed down the SD regions
107 of both species and provided insights into the history of sex chromosome turnover of
108 both species.

109

110 **Results**

111 **Genome assembly and annotation of tilapia genomes**

112 We produced 96 \times and 85 \times genomic coverage of Nanopore long-read sequences, with
113 a read N50 length of 26kb and 39kb for a female *ONJp* individual (with XX genotype)
114 and a male *OA* individual (ZZ genotype) respectively. Such a high sequencing coverage
115 has overcome the higher error rate of Nanopore reads than that of PacBio reads, and
116 produced similar numbers of genome size, contig N50 length and genome
117 completeness measurement (BUSCO score), compared to those of *ONEg* and *MZ*
118 previously derived from PacBio reads (**Figure 1a-b, Table 1**) [29, 30] or other
119 chromosome-level fish genomes (**Supplementary Fig. S1**). With the linkage
120 information provided by the Hi-C technology, we anchored 97.4% and 97.8% of the
121 genome of *ONJp* and *OA* into chromosomes (**Supplementary Fig. S2**), followed by
122 genome polishing with high coverage of Illumina reads and manual curation of scaffold
123 orders within the chromosomes. The percentages of anchored sequences of the two
124 genomes are higher than that (90.2%) of *ONEg* by genetic map [30]. And notably, we
125 found no interchromosomal and very few intrachromosomal rearrangements by the
126 genome-wide comparison between *ONJp* and *ONEg*, confirming the correct orientation
127 of scaffolds within our chromosome assemblies. The unanchored sequences are
128 enriched for repetitive elements that have alignments with multiple anchored
129 chromosomal sequences. The most significant improvement of *ONJp* over *ONEg* is
130 concentrated at the highly repetitive end of LG3, which made LG3 the largest
131 assembled chromosome (over 130 Mb) in the genome (**Figure 1c, Supplementary**
132 **Table S1**). Its overall repeat content (63%) is estimated to be around 2 fold higher than
133 any other chromosomes in the genome (**Figure 1d**), and such a large chromosome-

134 specific heterochromatic region has been found in both *ONJp* and *OA* (**Supplementary**
135 **Fig. S2**). Particularly, the last 70 Mb sequence of LG3 exhibits an extremely high repeat
136 content of about 75%. This is consistent with previous cytogenetic studies that identified
137 LG3 as the largest characteristic subtelocentric chromosome shared by all the
138 examined Tilapiine species [16, 31].

139 The total number of annotated genes is comparable between the *MZ* cichlid
140 versus the two tilapia species (**Table 1**). We found certain gene ontology (GO)
141 categories of genes are enriched (FDR<0.05, **Supplementary Table S2**) for significant
142 family expansion or contraction specifically at the ancestor of tilapias after their
143 divergence from the other cichlids (**Supplementary Fig. S3**). For example, besides the
144 reported olfactory receptor gene families [3, 32], we found immune-response related
145 genes (e.g. *CTLA4*, **Figure 1e**), and 'G-protein coupled receptor protein signaling
146 pathway' genes (**Figure 1f**) related to environmental sensing have specifically
147 increased their copy numbers in the two tilapias. These genes may have contributed to
148 tilapias' adaptation to more varieties of ecological niches compared to the lake cichlids,
149 and explained why they have been introduced as aquaculture species to over 150
150 countries.

151

152 **Characterization of complex repetitive genomic regions of tilapias**

153 Non-coding repetitive sequences can play critical structural and regulatory roles in the
154 genome, and there have been great efforts in mapping and characterizing such
155 elements (e.g., satellites, short interspersed nuclear elements (SINE), rDNA) in the
156 cichlids [33]; [34-37]. The two highly-continuous tilapia genomes allow us to scrutinize
157 these highly repetitive genomic regions that are mostly absent or unanchored in the
158 previous version of genome assemblies. There are two characteristic satellite
159 sequences SATA and SATB present in large tandem arrays with up to hundreds of
160 thousands of copies in the tilapia genomes. In particular, variants of SATA were
161 previously identified to be concentrated at the centromeric regions of *ON* chromosomes
162 [36, 38], and used as a phylogenetic marker to separate different tilapia tribes [37]. We
163 used high (between 30 to 620) copy numbers of SATA as a marker and annotated the

164 putative centromeric regions of over half of the chromosomes in both *ONJp* and *OA*
165 genomes (**Figure 1g**). As expected, we found the locations of putative centromeres are
166 colocalized with the junctions between the two arms of large intrachromosomal
167 interaction domains (**Supplementary Fig. S4, Supplementary Table S3**), similar to
168 what has been reported in other vertebrates [39]. The monomer sequence of SATA
169 satellites shows high degrees of variations (indels or SNPs) at certain monomer
170 positions between copies of the same or different chromosomes, but there are
171 intriguingly no variations at all in the last 58bp region across all the mapped loci of the
172 two species (**Supplementary Fig. S5**). This suggests concerted evolution and potential
173 functional constraints within this region. Most assembled putative centromeres are close
174 to or at the tip of the chromosomes. Their genomic locations are conserved between the
175 *ON* and *OA* genomes, without obvious centromere repositioning events that may play a
176 role in speciation [40] (**Supplementary Fig. S6**). This is in accordance with the reported
177 highly conserved karyotype between blue and Nile tilapia species that consists of
178 almost exclusively acrocentric or subtelocentric chromosomes [33, 41, 42]. The other
179 satellite SATB of longer monomer length (1.9kb) [36] has been previously shown to be
180 concentrated on the short arm of one chromosome, or on those of up to 14 pairs of
181 chromosomes, depending on the experimental conditions of fluorescence *in situ*
182 hybridization (FISH) [38]. We confirmed here that SATB satellite is frequently co-
183 localized with SATA and enriched in pericentromeric regions of at least 8 chromosomes
184 in both tilapias (**Supplementary Fig. S7**).

185 The other classic tandem array sequences that are of great interest but
186 extremely difficult to assemble are the ribosomal DNA (rDNA) clusters with hundreds of
187 thousands of copies in the genome. This is evidenced by the fact that rDNAs have been
188 mapped for their locations in over 500 fish species, but are only studied for their partial
189 genomic sequences in three species [43]. Eukaryotic rRNA genes are divided into two
190 classes of 45S (corresponding to the nucleolar organizer regions, NORs) and 5S rRNA
191 genes. They are transcribed by different RNA polymerases, and often located on
192 different chromosomes in teleost species. Here we dissected the complex sequence
193 structures and mapped the rRNA gene clusters in *ONJp* and *OA* genomes. Both

194 species show similar numbers and chromosomal locations of mapped loci
195 (**Supplementary Fig. S8**), but the *OA* genome (**Figure 2b**) probably captures a more
196 complete sequence composition with its better assembly quality thus is used for
197 demonstration here. We mapped the major 45S and 5S rRNA clusters respectively on
198 LG14/6/4 and LG23/22, which is consistent with previous cytogenetic results [33]. The
199 total copy numbers of 45S and 5S rDNA were estimated to be 123 and 171 throughout
200 the *OA* genome. The 45S rDNA cluster on LG14 is located at the end of acrocentric
201 chromosomes (**Figure 2a**), and consists of 11 transcriptional units coding for the 18S,
202 5.8S and 28S rRNAs. Each unit containing internal transcribed spacers (ITS) is
203 separated by intergenic non-transcribed spacers (IGS). Three tandem units are
204 organized in inverted orientation to the other eight units, suggesting recombination may
205 happen between these units by forming a hairpin structure (**Figure 2c**). The IGSs of the
206 two groups of tandem units are only partially homologous to each other in sequence,
207 and are themselves tandem arrays of multimers. Remarkably, there are nine copies of
208 3.7 kb repetitive sequences (>97% sequence similarities between copies) in one IGS
209 (**Figure 2c**), and each copy consists of 25 copies of 102bp sequences (about 95%
210 sequence similarity between copies) that are separated into two clusters. Such nested
211 tandem arrays of repetitive sequences resemble the higher order repeats (HORs) of
212 human centromere [44] and remain to be studied for their functions.

213 There are two classes of 5S rDNA (**Figure 2e**), which respectively consist of
214 1.4kb (type I) and 0.5kb (type II) repeat units. The type I 5S rDNA cluster residing on the
215 LG23 consists of 26 tandem duplications of repeat units. Each repeat unit contains a 5S
216 RNA, an inverted 5S RNA and a SINE repeat. This SINE repeat seems to be derived
217 from 5S RNA, with more than half of its sequence homologous to 5S RNA.

218
219 **LG3 heterochromatic regions encompass tandem arrays of protein-coding and**
220 **small RNA genes**

221 The heterochromatin concentrated at the end of LG3 forms a large unpaired region
222 during male meiosis, which was presumed to be the sex chromosome pair of *ON* [18].
223 The repetitive nature of this part of LG3, together with its extremely low recombination,

224 may have contributed to the difficulty of finding genetic markers to anchor a large
225 portion of LG3 in the *ONEg* assembly [22]. Later QTL mapping and genomic analyses,
226 however confirmed that LG1 or LG23 is the sex chromosome of *ON* [22, 26], while LG3
227 has been frequently adopted as the sex chromosome in other tilapia species [22, 45].
228 To trace the origin of such unusual autosome-specific heterochromatin, we compared
229 the assembled sequences of LG3 of *ONJp* and *OA* vs. that of *MZ*. Based on their
230 syntenic alignment (**Figure 3a**) and the distribution of repeat content along the LG3
231 (**Figure 3b**), we inferred that there were probably two episodes of repeat expansion
232 (RE): the first one is shared by both *MZ* and *ONJp* (**Supplementary Fig. S9**), thus may
233 have occurred at the ancestor of all cichlids. It is manifested as a turning point at around
234 the 30Mb position, where the repeat content starts to increase, while the GC content
235 and recombination rate start to decrease [30], forming a heterochromatic region (Het-1)
236 with over 50% of the sequences as repetitive elements. The clear negative correlation
237 between these genomic features can be explained by the scarcity of GC-biased gene
238 conversion caused by the low recombination rate [46]. The second RE impacts the
239 region beyond the 70Mb position (Het-2), and probably occurred more recently at the
240 ancestor of tilapia species. It accounts for the dramatic size increase of LG3 from 45 Mb
241 in *MZ* to more than 130 Mb in both tilapias, and the increase of repeat content from
242 around 50% to over 70% (**Figure 3c**). This massive repeat expansion involves all
243 repeat families but DNA transposons and simple repeats (**Figure 3b**). Most repeats
244 seem to have started the expansion from the euchromatin region toward the other
245 chromosome end (**Figure 3c**), the latter of which is enriched for the younger repeat
246 elements that show a low level of sequence divergence from their consensus
247 sequences (**Supplementary Fig S9**). Of particular interest are three previously
248 uncharacterized repeat families: although they only account for 1.9% of all LG3 repeat
249 sequences, they are almost exclusively concentrated on LG3, with two of them only at
250 the LG3 Het-2 region (**Figure 3d, Supplementary Fig. S10a**). They are shared by all
251 sequenced cichlids studied here (thus we named them as CLD repeats), and include
252 one DNA transposon (DNA_CLD1), and two uncategorized repeats UNCLD1 and
253 UNCLD2. But their copy numbers have specifically increased in tilapias: for other tilapia

254 species without a genome generated by third-generation sequencing, we estimated
255 their relative repeat copy numbers by kmer frequency scaled against the genome
256 coverage, and found minor expansion of UNCLD1 and UNCLD2, but a 12.6 fold
257 expansion of DNA_CLD1 across all sequenced tilapias relative to *MZ* (**Supplementary**
258 **Fig S10b**). The bombardment of various repeats has clearly demarcated the entire
259 chromosome of *OA* and *ONJp* into one large active (A) and one large repressive (B)
260 chromatin compartment (**Figure 3e-f**), revealed by our chromatin interaction analyses.
261 We artificially marked the boundary between the A/B compartment also as the one
262 between Het-1/-2 regions. As expected, genes located at the Het-1 or -2 regions of LG3
263 are expressed at a significantly ($P < 2.2\text{e-}16$, Wilcoxon rank sum test) lower level than
264 those on the other LGs across all examined tissues (**Figure 3g**).

265 The heterochromatic regions of LG3, however, are not gene deserts, but instead
266 more frequently harbor gene duplications than other LGs (**Figure 4a**), probably due to
267 the non-homologous recombination or replication slippage mediated by the excessive
268 repeats [16]. This is exemplified by independently formed gene clusters of tandem
269 duplication among the *MZ* and the two tilapia species within their Het-2 regions (**Figure**
270 **4b**). Some species-specific gene duplicates have probably evolved novel functions: for
271 example, gene copies of *Zina33* are mainly expressed in the heart and kidney of *MZ*,
272 but have acquired new expression patterns in brain and liver in *ONJp* in some copies
273 (**Supplementary Fig. S11**). Besides facilitating the generation of these new gene
274 duplicates that may contribute to the species-specific adaptation, we also found a
275 disproportionately large number of predicted PIWI-interacting RNA (piRNA) or small-
276 interfering RNA (siRNAs) encoding loci on LG3, which account for about 30% of the
277 small RNA loci throughout the genome of *ONJp* (**Figure 4c**). Particularly, the predicted
278 small RNA loci form a gradient along the LG3 of their density (number of loci per 100kb)
279 and are mostly concentrated on the more recently formed Het-2 region (**Figure 4d-e**).
280 Similar to the reported expression patterns of piRNAs or siRNAs in other model species,
281 these small RNAs are predominantly expressed in the gonads relative to the liver tissue
282 (**Figure 4f**), suggesting they play a similar role of suppressing transposon activities and
283 guard the germline genome integrity as they do in other species [47]. Interestingly, the

284 piRNA-encoding repeat elements show a bimodal distribution of ages reflected by their
285 sequence divergence level from the respective consensus sequences (**Supplementary**
286 **Fig. S12**), with the peak of younger repeats largely overlapped with those of LG3. This
287 together with the more concentrated distribution of small RNA loci at Het-2 provide
288 evidence that the more recent RE of LG3 probably has selected for the emergence of
289 novel small loci as a response to tame the new transposons acquired on LG3.

290

291 **Turnover of sex chromosomes and sex determination pathways**

292 The complete X (LG23) chromosome sequence of *ONJp* and the Z chromosome (LG3)
293 of *OA* provide us a great opportunity to gain insights into the evolution process and the
294 consequences of rapid turnover of SD systems. Previous work has demonstrated the Y-
295 linked duplicated copy of *Amh* (*Amhy*) on LG23 as the SD gene of *ONJp* [25]. This has
296 been confirmed by our analyses of Illumina reads generated from male *ONJp*
297 individuals with an XY karyotype. As expected, the XY reads show excessive numbers
298 of SNPs (i.e., differences between the X- and Y-linked alleles or gametologs) along the
299 X chromosome (**Figure 5a**), and also a nearly doubled read coverage of a YY male
300 (derived from crossing the wild-type male with the sexually reversed XY female)
301 indicative of duplication at the region encompassing *Amh* (**Figure 5b**), compared to the
302 surrounding regions, or the patterns derived from female (XX) reads (**Supplementary**
303 **Fig. S13**). With the same rationale, we used the ZW reads of *OA* and identified LG3 as
304 its sex chromosome pair (**Figure 5c**) with excessive ZW-derived SNPs. We managed to
305 exclude the segregating polymorphic sites and further narrowed down the previously
306 identified SD region (SDR) on the Z chromosome from about 40 Mb long into 0.6Mb, by
307 inspecting the newly produced resequencing data, as well as other published data [22,
308 48] of different *OA* populations. We first identified the fixed female-specific SNPs or
309 indels that are shared among all the populations which are only concentrated at a 10Mb
310 long region (**Figure 5d**). We then focused on an enclosed region that shows the highest
311 density of female-specific heterozygotes, i.e., the largest differences between the Z and
312 W chromosomes. We genotyped randomly selected candidate sex-linked markers
313 within the region (**Supplementary Table S4**), and found one deletion (**Supplementary**

314 **Fig. S14)** and one SNP site that are specific to the W chromosomes of all the inspected
315 female *OA* individuals. This candidate SDR spans 620kb and harbors three candidate
316 SD genes, *Banf2*, *Paics-1*, and *Paics-2*. Intriguingly, these genes show an elevated
317 female vs. male read coverage, suggesting that they are duplicated on the W
318 chromosome (**Figure 5e**).

319 We hypothesize that a master SD gene, might be expected to show transient
320 sex-specific gene expression during early gonadogenesis, similar to the *Sry* of eutherian
321 mammals[49]. To inspect the candidate SD genes of *OA* for their expression, and also
322 to elucidate the impact of sex chromosome turnovers between *ONJp* and *OA* on their
323 downstream SD pathway genes, we collected the gonad transcriptomes of both sexes
324 from these two species' corresponding stages. The collected stages span the onset
325 (from 5 days after hatching, or 5-dah), an early (30-dah) and a late (180-dah) stages of
326 gonad differentiation [50], during which the histological differences between gonads of
327 the two sexes become more apparent (**Figure 5f**). Consistently, we found that the
328 numbers of sex-biased genes dramatically increase from 5-dah to the later stages in
329 both sexes of both species (**Supplementary Fig. S15**). In particular, *Paics* have
330 multiple tandem copies at the SDR of both Z (**Figure 4b**) and W (**Figure 5e**)
331 chromosomes of *OA*, and show an increasing ovary-specific expression pattern (**Figure**
332 **5g**) through early gonadogenesis in *OA* but not in *ONJp*, suggesting it is likely a
333 candidate female SD gene. The orthologous genes of *Paics* are specifically expressed
334 in gonads of both sexes in *MZ* (**Supplementary Fig. S11**), suggesting that evolution of
335 the potential female-determining function of *Paics* in *OA* might involve suppressing its
336 expression in males. The validation and detailed dissection of *Paics* function require a
337 complete sequence of W chromosomes and more experimental work in future.

338 The recent transition between the XY chromosomes of *ONJp* and the ZW
339 chromosomes of *OA* is expected to rewire the downstream SD pathways of the two
340 species. To test that, we compared the two species' gonad expression trajectories of
341 orthologous genes of known vertebrate SD genes: majority of them show a conserved
342 sex-biased temporal expression pattern, but to a different degree, between the two
343 species across the sampled stages (**Supplementary Fig. S16-17**). This suggests that

344 these genes participate in the SD process of both species, but may play a different role
345 because of the turnover of their upstream SD genes. Among them, knockouts in *ONJp*
346 of conserved teleost male-determining genes *Amhy*, *Gsdf* or *Dmrt1*, or those of female-
347 determining genes *Foxl2* or *Cyp19a1a* all leads to sex reversal [27, 51, 52]. The
348 temporal expression patterns of these genes are consistent with their known
349 hierarchical positions in the SD pathway of *ONJp*: for example, *Dmrt1* has robust male-
350 specific expression and steady upregulation since 5dah, before its validated
351 downstream target genes *Gsdf* [52], *Sox9b* [53] and *Sox30* [54] reaching their peak
352 expression levels specifically in males. This is similar in the female determining pathway
353 between the upstream gene *Foxl2* vs. its downstream target *Cyp19a1a* [27]. Of
354 particular interest is the much higher expression level of *Dmrt1* in *ONJp* than in *OA* in
355 their gonads of 5dah when sex is determined. This is probably because of the
356 origination of new master male SD gene *Amhy* in *ONJp*, whose disruption has been
357 demonstrated to suppress the expression of *Dmrt1* in the male gonads [25]. The
358 increased expression of *Dmrt1* may also account for those of its downstream genes
359 *Sox9b* and *Sox30* in *ONJp* than in *OA*. Interestingly, the upstream female SD gene
360 *Foxl2* is also upregulated in *ONJp* than in *OA*. Since *Dmrt1* and *Foxl2* have a conserved
361 antagonistic relationship during vertebrate SD process that disruption of one would
362 cause the upregulation of the other in the respective sex [55]; an increased expression
363 level of *Foxl2* in *ONJp* female could result from the co-evolution in response to *Dmrt1* in
364 male, or replacement of early female SD role of *Foxl2* in *OA* by the newly evolved
365 candidate SD gene *Paics*.

366

367 **Discussion**

368 The great diversity of phenotypes and sex chromosomes generated in a relatively short
369 evolutionary time range makes African cichlids a classic model for studying the
370 mechanisms of species radiation and sex chromosome transitions. Since the release of
371 five representative cichlid genomes over five years ago [3], analyses of more cichlids'
372 (e.g., those from Lake Malawi [56, 57] and Lake Mweru [58]) and higher qualities of
373 genomes (e.g., that of *ONEg* [22]) have been published, demonstrating the lasting

374 interest in these species. Here we focused on two important aquaculture species from
375 the much less species-rich but much more widely distributed *Oreochromis* genera that
376 have undergone very recent transition between XY and ZW sex chromosome systems.
377 Their high-quality genomes demonstrated by our in-depth analyses of the complex
378 repetitive regions provided novel insights into the genome architecture and sex
379 chromosome evolution of teleosts.

380 **Chromosome-specific heterochromatin made LG3 a ‘sexy’ chromosome for
381 becoming tilapia sex chromosomes**

382 A few known genes (so-called ‘usual suspects’[59]), for example, *Dmrt1*, *Amh*, *Sox3*
383 and *Gsdf* etc. have a conserved role in the vertebrate SD pathway, and frequently
384 evolved to become a master SD gene through point mutations or duplication of one
385 allele in the proto-sex chromosome pair. Their residing ancestral chromosomes or
386 chromosomal fragments (the ‘sexy’ chromosome) are therefore recruited as sex
387 chromosomes more often than other chromosomes [60, 61]. For example, the chicken Z
388 chromosome harboring *Dmrt1* has been independently recruited as sex chromosomes
389 in monotremes and in a gecko species [60]. Substantial proportions of bullfrog sex
390 chromosomes harboring *Sox3* are homologous to the human X chromosome [62].
391 Among tilapias, LG1 and LG3 are the sexy chromosomes that have been most
392 frequently found as sex chromosomes in all the investigated species [9, 21], although
393 some species have recently evolved SD genes on other LGs (e.g., LG14 of *O.*
394 *mossambicus* and LG23 of *ONJp*) [25, 45]. In contrast, LG1 and LG3 have not been
395 found as sex chromosomes in other non-tilapia cichlids, except for a sex-linked QTL on
396 LG3 in two Lake Malawi cichlid species[63]. LG23 of *ONJp* became sex chromosomes
397 because of the Y-linked duplication of *Amh* [25]. However, no other ‘usual suspects’ or
398 known master SD genes have been detected within the SDR of LG1 and LG3 [20, 22,
399 30, 45], suggesting an alternative scenario for recruiting them as sex chromosomes.

400 In this study, we suggest that tilapia- and chromosome-specific expansion of
401 heterochromatin may have contributed to the more frequent recruitment of LG3 as sex
402 chromosomes. By assembling the nearly complete heterochromatic region and
403 comparison to the Lake Malawi cichlid *MZ*, we inferred that there were two waves of RE

404 (Figure 3c, Supplementary Fig. S9). One is probably shared by tilapias and Lake
405 cichlids, and the other is only shared by tilapias on LG3. Both REs dramatically
406 increased the TE content (Figure 1d), and decreased the recombination rate across
407 over two thirds of the LG3 region[30], compared to other LGs. The classic model of sex
408 chromosome evolution, as indicated by mammals and birds, hypothesizes that the
409 origination of SD genes would select for suppression of recombination and lead to
410 accumulation of repetitive elements on the Y or W chromosomes[64]. Given the
411 divergence level and the SDR length between sex chromosomes of *OA* are much
412 smaller than those of mammals and birds, it seems more likely a reversed process
413 occurred in which RE and reduction of recombination rate predated and facilitated the
414 emergence of new SD genes on LG3. The abundant repetitive sequences can promote
415 gene duplications or other types of mutations that endow the new SD function to the
416 pre-existing alleles. And the substantial linkage disequilibrium created by the large
417 heterochromatic region probably will further fix the combination of the newly invaded SD
418 locus with other sexually antagonistic loci on the same chromosomes. A similar
419 scenario has been suggested for other cichlids [65, 66] and guppies [67].

420 **Newcomers of tilapia SD genes**

421 The low sex chromosome divergence level (Figure 5) suggested that LG3 as a sex
422 chromosome pair of *OA* evolved very recently, although it requires further confirmation
423 of SDR and candidate SD genes on LG3 of other tilapias (e.g., *O. karongae* and *O.*
424 *tanganicae*)[10, 45]. We identified the candidate SD genes *Paics* on LG3 of *OA*, which
425 have two Z-linked copies within the SDR. *Paics* genes have an increased copy number
426 on the W chromosome and an ovary-specific expression pattern during gonadogenesis
427 of *OA* but not in *ONJp*, but their human ortholog is ubiquitously expressed across all
428 tissue types without an obvious sex-biased pattern (Supplementary Fig. S18). The
429 human *Paics* encodes the phosphoribosylaminoimidazole carboxylase that participates
430 in the purine biosynthesis without sex-related functions. How did *Paics* evolve their
431 ovary-specific expression from the ancestral non-biased expression pattern; and what
432 are the role of the extra W-linked *Paics* copies, if any, during the female SD process of

433 *OA* remain intriguing questions for the future experimental studies. They also require a
434 complete sequence of the W chromosome of *OA*.

435 Many ‘usual suspects’ or their duplications were identified as the master SD
436 genes in teleosts, e.g., *Amhy* of *ONJp*, *Dmrt1bY* (*DMY*) or *Sox3Y* in different medaka
437 species[68-70] etc. Nevertheless, more ‘newcomers’ of master SD genes like *Paics*, i.e.,
438 genes that have no previously known SD functions, have now been discovered. For
439 example, the candidate male SD gene of the channel catfish seems to be the male-
440 specific isoform of breast cancer anti-resistance 1 (*BCAR1*) gene [71]. And the male SD
441 gene of rainbow trout *sdY* (*Oncorhynchus mykiss*) is derived from duplication and
442 truncation of an immunity-related gene *irf9* [72]. It has been recently shown that *sdY*
443 functions by hijacking the female SD regulatory loop between *Foxl2* and *Cyp19a1* to
444 promote testis differentiation[73]. Thus, it is possible that *Paics*, if demonstrated to be a
445 true female SD gene, might similarly interfere with the interaction loop of male SD
446 pathway (e.g., between *Dmrt1* and *Sox9b*) or promote that of the female SD pathway
447 involving *Foxl2* and *Cyp19a1a*. All these key SD genes indeed have a different
448 expression level between *OA* and *ONJp* in the early gonads (**Figure 5g**). Tests of these
449 hypotheses will require transgenic expression of *Paics* genes of *OA* in the females of
450 technically more accessible *ONJp* to evaluate their impact on the known SD genes.

451 **An important resource for cichlid functional genomic studies and aquaculture**

452 Previous comparison between *ONEg* and *MZ* [30] indicated that the genomic
453 differences between the two species are dominated by intra- rather than inter-
454 chromosomal rearrangements, consistent with a largely conserved karyotype among
455 African cichlids shown by cytogenetic studies [42, 74]. Therefore, the much improved
456 genomes of *ONJp* and *OA* generated by this study in chromosome shape are to provide
457 a high-quality reference for future studies into the patterns and mechanisms of
458 speciation of Lake cichlids. They will be very useful for anchoring the genomes of other
459 cichlid species into chromosomes, and annotating their genes and conserved functional
460 non-coding regulatory elements. More importantly, so far a total of 6 SD genes have
461 been successfully knocked out in *ONJp* [25, 27, 51, 52, 75], with other well-established
462 genetic resources and techniques like antibodies and transgenics, as well as the high-

463 quality genome available now, *ONJp* becomes a promising model for testing the
464 functions of identified genes responsible for the focal phenotypes (e.g., sex
465 determination, body colors) in the future.

466 The completeness of our new genomes is exemplified by our assembly of
467 complex repetitive regions like the LG3 heterochromatin and rDNA clusters (**Figure 2-3**). The genomic sequences, particularly the fixed sex-specific markers identified here in
468 *ONJp* and *OA* (**Figure 5b,e**) can assist hybridization schemes aiming for producing
469 monosex tilapia fry. For example, these markers can be used to discriminate between
470 sexually reversed ZZ female *OA* vs. the wild-type females, so that the ZZ female can be
471 further crossed with the wild type ZZ male to produce all-male fry, which are preferred
472 over females in aquaculture.

474

475 **Conclusion**

476 In this work, we generated and analyzed the chromosome-level genomes of two
477 important aquaculture species *ONJp* and *OA*. We characterized their complex repetitive
478 regions including centromeres, rDNA loci, and the chromosome-specific
479 heterochromatic region on LG3. We showed that the acquisition of LG3 heterochromatin
480 is the result of two episodes of repeat expansions, accompanied by dramatically
481 reduced recombination rate [30] over two thirds of this chromosome. Within the LG3
482 heterochromatin, we identified a candidate female SD gene in *OA* that showed an
483 ovary-specific expression pattern during the critical stage of sex determination. Overall,
484 our work provides important genomic resources for studying SD mechanisms and
485 genome architectures of tilapias.

486

487 **Materials and Methods**

488 **DNA sampling and sequencing**

489 All animal experiments were conducted in accordance with the regulations of the Guide
490 for Care and Use of Laboratory Animals and were approved by the Committee of
491 Laboratory Animal Experimentation at Southwest University. High molecular weight
492 DNAs of *ONJp* (derived from Prof. Nagahama at National Institute for Basic Biology of

493 Japan) and *OA* (from Wuxi Freshwater Fisheries Center in China) were extracted from
494 muscle tissues using a Blood & Cell Culture DNA Midi Kit (Q13343, Qiagen, CA, USA).
495 We also obtained genomic DNAs from *ONJp* with a YY genotype, and *OA* with a WW
496 genotype, by crossing the XY male with the XY female of *ONJp*, and the ZW female
497 with the ZW male of *OA*. The sexually reversed XY female or ZW male individuals were
498 produced by treating fry with the aromatase inhibitor Fadrozole (Novartis)[23] or 17-
499 alpha-ethynodiol[76]. We performed the DNA quality and quantity assessment
500 using a Qubit double-stranded DNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific)
501 and an Agilent Bioanalyzer 2100 (Agilent Technologies). For each Nanopore library,
502 approximately 8 μ g of gDNAs from the female *ONJp* (XX genotype) and male *OA* (ZZ
503 genotype) were size-selected (10- 50 kb) with a Blue Pippin (Sage Science, Beverly,
504 MA), and processed using the Ligation sequencing 1D kit (SQK-LSK108, ONT, UK)
505 according to the manufacturer's instructions. Libraries were constructed and sequenced
506 on R9.4 FlowCells using the GridION X5 sequencer (ONT, UK) each at the Genome
507 Center of Nextomics (Wuhan, China). To acquire a chromosomal-level assembly of the
508 genome, one gram of gonad tissues collected from the same *ONJp* or *OA* strain of the
509 same genotype was used for Hi-C library construction. The Hi-C experiment consisted
510 of cell crosslinking, cell lysis, chromatin digestion, biotin label, proximity chromatin DNA
511 ligations and DNA purification, which were performed by Annoroad Genomics (Beijing,
512 China) following the standard procedure [77]. The purified and enriched DNA was used
513 for sequencing library construction. Illumina HiSeq X Ten platform (Illumina) was used
514 to perform sequencing with a read length of 150 bp for each end. To identify the
515 candidate SDR, we also performed Illumina sequencing for the YY *ONJp* and WW *OA*
516 individuals with a 250bp library insert size.

517 **Genome assembly**

518 We used flye (2.3.1) [78] to assemble the Nanopore raw reads, with default parameters.
519 The draft assembly was then polished by Racon (v1.3.1) [79]. To do so, we mapped the
520 raw Nanopore reads using minimap2 (2.15-r905) [80], with options '-x map-ont --
521 secondary=no'. We performed Racon polishing for two rounds with default parameters.
522 We then used purge_haplotype [81] to remove tentative haplotypes (alternative haploid

523 contig). Coverage distribution of Nanopore reads were calculated using the readhist
524 module in purge_haplotigs, after the reads were mapped against the assembly by
525 minimap2 [80]. We used the options ‘-j 80 -s 80’ to decide the classification of haplotigs,
526 and the haplotigs were subsequently removed. The 3D-DNA pipeline (180922) [82] was
527 used to join the contigs into chromosomes. First, we mapped the Hi-C reads against the
528 contigs using Juicer (1.7.6) [83] with default settings. After removing the duplicates, the
529 Hi-C contact map was directly taken as input for 3D-DNA. The parameters were set as
530 ‘--editor-coarse-resolution 500000 --editor-coarse-region 1000000 --editor-saturation-
531 centile 5 -r 0’. We subsequently used Juicebox Assembly Tools [84] to review and
532 manually curate scaffolding errors. We further used Pilon (1.22) [85] to polish the
533 assembly with Illumina sequencing reads. For the ZZ genome (OA), ~40X sequencing
534 data from a short-insert library was produced for polishing the assembly. Those options
535 were used by Pilon: ‘--minmq 30 --diploid --fix bases,gaps --mindepth 15’. To assess
536 the completeness of the assembled genome, we screened the assembly for BUSCO
537 genes (3.0.2) [86] of actinopterygii. The ‘geno’ model was used with default parameters.

538 **Genome annotation**

539 We used RepeatModeler (1.0.10) to predict repetitive elements throughout the genome
540 and to classify the repeats based on their similarity to known repeat families. The
541 unclassified repeats were labelled as ‘unknown’. We then combined the newly predicted
542 repeat family with an existing repeat library from RepBase, and used RepeatMasker
543 (4.0.7) to search for repeats in the genomes. We used the MAKER pipeline (2.31.10)
544 [87] to annotate gene models. The protein sequences of *Oreochromis niloticus*
545 (O_niloticus_UMD_NMBU) [22] and *Maylandia zebra* (M_zebra_UMD2a) [30] were
546 downloaded from NCBI RefSeq as the query to search for homologs. An initial set of
547 gene models were predicted by MAKER with the input of protein sequences alone. We
548 also used Trinity (2.4.0) [88] to assemble the transcriptomes with the parameters ‘--
549 min_glue 5 --path_reinforcement_distance 30 --min_contig_length 300’. We further built
550 a comprehensive transcript database by using the PASA pipeline (2.3.3) [89] with
551 options ‘--stringent_alignment_overlap 30 --ALIGNERS gmap --TRANSDECODER’.

552 Then we fed the MAKER-produced gene models into the PASA pipeline for gene-model
553 polishing (-A --gene_overlap 50).

554 The known tilapia rDNA sequences were retrieved from NCBI (accession GU289229.1
555 and MF460358.1) and were mapped against the genome using blastn (2.10.0). For the
556 mapped locus, dotplots were produced by flexidot (1.06) [90] with the parameters -f 1 -k
557 39 -S 2.

558 **mRNA sequencing and gene expression analysis**

559 We extracted the total gonad RNAs of each sex of *OA* at 5, 30, and 180 dah
560 (**Supplementary Table S5**) using the Trizol Reagent (Invitrogen, Carlsbad, CA), and
561 eliminated the genomic DNA using DNasel. RNA qualities were monitored on 1%
562 agarose gels and a Nanodrop spectrophotometer. Illumina sequencing was carried out
563 at Novogene Bioinformatics Technology Co., Ltd., in Beijing, China. Sequencing
564 libraries were constructed using the NEBNext® Ultra™ RNA Library Prep Kit for
565 Illumina® (NEB, USA), according to the manufacturer's protocol. Index codes were
566 added to attribute sequences to each sample. The prepared libraries were sequenced
567 on an Illumina Hiseq 2500 platform, and 150bp paired-end reads were generated.
568 Gonad transcriptome data of *ONJp* at 5, 30 and 180 dah were from the previous studies
569 [50, 91]. The raw RNA-seq reads were mapped against the genomes using HISAT2
570 (2.1.0) [92]. The number of reads for each gene was counted using featureCounts
571 (1.6.2) [93] according to the gene model annotation. To quantify the expression levels,
572 read counts were normalised using the TPM (transcripts per million) method. The mean
573 expression levels were calculated for biological replicates.

574 **Small RNA analysis**

575 The small RNA (sRNA) sequencing data of *ONJp* was retrieved from [94] and [95].
576 Trimmomatic (0.36) [96] was used to trim adaptors and low-quality bases with the
577 parameter ILLUMINACLIP:adapter:2:30:7 MINLEN:18. The sequencing reads were
578 collapsed using seqcluster (1.2.7) [97] prior to mapping. The aligner bowtie (1.2.1.1)
579 [98] was used to map the reads to the genomes with the parameters --best --strata -k1 -
580 m 1000. The sRNA sequences were compared against small RNA Rfam with cmscan to
581 filter out those that were annotated as microRNA, tRNA, rRNA and other known small

582 RNAs. We further classified the small RNA according to the sequence lengths: piRNA
583 between 26 bp and 32 bp, siRNA between 16 bp and 23 bp. The expression levels of
584 putative piRNA and siRNA loci were quantified by counting read counts at each locus
585 followed by normalization using the TPM method. We used proTRAC (2.4.3) [99] to
586 detect piRNA clusters with default parameters. Prior to running proTRAC, piRNA
587 transcripts were mapped to the genome using the script sRNAmapper.pl following the
588 recommendations by proTRAC.

589 **Sex determining region**

590 Whole genome resequencing data of multiple males and females (**Supplementary**
591 **Table S6**) were produced to infer the SDR in both Nile tilapia and blue tilapia. The
592 sequencing reads were mapped against the genome with BWA-mem (0.7.16a). For
593 each sample the variants were called using GATK (3.8.1.0) HaplotypeCaller [100]. The
594 variants were then genotyped together, combining all samples (join calling). The single
595 nucleotide variants were selected and filtered using the criteria QD < 2.0 || FS > 60.0 ||
596 MQRankSum < -12.5 || RedPosRankSum < -8.0 || SOR > 3.0 || MQ < 40.0. For *ONJp*,
597 we selected SNPs that were heterozygous (0/1) in males but homozygous (0/0) in
598 females, and for *OA*, we selected female-heterozygous (0/1) but male-homozygous
599 (0/0) SNPs. For pooled resequencing data, we used LoFreq (2.1.2)[101] to call variants,
600 with default parameters. We required the heterozygous SNPs to have an allele
601 frequency between 0.35 and 0.65. The regions contained those sex-linked SNPs were
602 defined as sex-determining regions. For blue tilapia, we further genotyped the SNPs by
603 PCR across the SDR (**Supplementary Table S3**), and discarded the variants that failed
604 to exhibit the sex-linked pattern across all the inspected male and female individuals
605 from different *OA* populations. Sequencing coverage was calculated by Samtools depth
606 (1.3.1) [102] (only for the sites with mapping quality of least 60) followed by calculating
607 the mean coverage in 50k sliding windows along the chromosomes.

608 **Histological analysis**

609 We sampled XX and XY fish at 5, 30 and 180 dah (days after hatching). Briefly, the fish
610 were anesthetized using an overdose of MS222 (Sigma-Aldrich, St. Louis, USA).
611 Histological analysis was performed as described [103]. We dissected gonads and fixed

612 the gonads in Bouin's solution for at least 24 hours at room temperature, dehydrated,
613 and embedded in paraffin. All tissue blocks were sectioned at 5 μ m using the Leica
614 microtome (Leica Microsystems, Wetzlar, Germany) and stained with hematoxylin and
615 eosin. Photographs were taken under Olympus BX51 light microscope (Olympus,
616 Tokyo, Japan).

617

618 **Data availability**

619 The sequencing reads have been deposited at NCBI SRA, under PRJNA609616. The
620 genome assemblies have been deposited at DDBJ/ENA/GenBank under the accession
621 JAAMTG000000000 and JAAMTF000000000

622

623 **Code availability**

624 The codes used in this study have been deposited at

625 <https://github.com/lurebgi/tilapiaSexChr>.

626

627 **Acknowledgement**

628 Q.Z. is supported by the National Natural Science Foundation of China (Grant nos.
629 31722050 and 31671319), the Natural Science Foundation of Zhejiang Province
630 (LD19C190001), and the European Research Council Starting Grant (Grant agreement
631 677696). D.W. is supported by the National Natural Science Foundation of China
632 (31861123001 and 31630082), and the National Key Research and Development
633 Program of China (2018YFD0900202). L.X. is supported by uni:docs fellowship from
634 University of Vienna.

635

636 **Competing interests**

637 The authors declare that they have no competing interests.

638

639

640

641 **Figure legend**

642 **Figure 1 High-quality genome assemblies at the chromosome level.** a) In this
643 study, the genomes of the Japanese strain of Nile tilapia (*ONJp*) and blue tilapia (*OA*)
644 were assembled. The genomes of the Egyptian strain of Nile tilapia (*ONEg*)[22] and
645 Zebra mbuna (*MZ*)[30] have been published. The heterogamety (XY or ZW) and the sex
646 chromosome (linkage group, or LG) were shown next to the fish photos. b) Contig N50,
647 the percentage of sequences anchored into chromosome, and genome completeness
648 (BUSCO) were compared among the four genomes. c) The genome synteny between
649 *ONJp* and *ONEg* is highly conserved except for LG3 which is much larger in terms of
650 anchored sequences for *ONJp* relative to *ONEg*. d) LG3 has a much higher repeat
651 content than the other LGs. e) GO enrichment (FDR < 0.05) for gene families that have
652 expanded in the tilapia lineage. Redundant GO terms were removed. f) One example of
653 gene duplication at the ancestor of tilapias. g) The density of centromeric repeats
654 (length per 50k) is shown along each chromosome.

655 **Figure 2 The genomic organization of rDNA loci.** a) The length of rDNA sequence
656 per 50kb along the chromosomes of *OA*. We selected one 45S locus (b) and one 5S
657 locus (d) for demonstration. b) the dotplot showing an array of 11 copies of 45S genes
658 and the intergenic spacers (IGSs). The green colors represent reversed alignments.
659 One IGS was selected for a zoom-in view. c) Each 45S locus contains one 18S, one
660 5.8S and one 28S gene. The second IGS forms a higher order repeat (HOR), consisting
661 of 9 repeats of a 3.7kb element which itself consists of tandem duplications of a 102 bp
662 sequence. d) The dotplot showing an array of 25 copies of 5S rRNA gene loci. Each
663 locus contains two 5S rRNAs with opposite coding directions and one SINE element
664 shown in e).

665 **Figure 3 Heterochromatin region of LG3** a) The synteny between the LG3 of *MZ* and
666 *ONJp*. Each grey band represents a synteny block. b) Comparison of the composition of
667 repeats on two heterochromatic parts (Het-1 and Het-2) of LG3 and other LGs. c) The
668 distribution of repeat content (pink) and GC content (blue) along the LG3 of *ONJp*. The
669 heterochromatic part is divided into Het-1 and Het-2 which show differential degrees of

670 heterochromatinization. d) The density (length per 5kb) of three transposable elements
671 along the LG3. e-f) The distribution of A/B compartments on the LG3 of *OA* and *ONJp*.
672 The A compartment usually corresponds to the active chromatin domain, and the B
673 compartment corresponds to the repressive or heterochromatin domain. They are
674 derived from eigenvector analyses of Hi-C data. g) Expression profiles of six tissues of
675 *ONJp*. Both Het-1 and Het-2 regions have significantly lower expression levels
676 compared with genes from other chromosomes.

677 **Figure 4 LG3 heterochromatin contains tandem arrays of mRNA and sRNA genes.**

678 a) LG3 has a larger portion of duplicated genes compared with other LGs. b) Tandem
679 duplications of genes with at least two duplicated copies are shown along the LG3 of
680 *OA* and *ONJp*. The homologous genes to tilapia duplicates are also shown on *MZ* LG3.
681 Homologous genes of the same family are in the same color across species. For tilapias
682 only the regions from 40 to 100 Mb of LG3 are shown. c) A disproportionately larger
683 number of piRNA clusters and siRNA genes on the LG3. d) log1p transformed density
684 of piRNA and siRNA genes over 100 kb windows. e) The density of piRNA and siRNA
685 on the positive (blue) and negative (black) strand. The black triangles indicate the
686 locations of piRNA clusters. f) log transformed expression levels (TPM) of piRNA and
687 siRNA of gonads and livers on the positive (blue) and negative (black) strand. Genes
688 with low expression (TPM < 1) were filtered out.

689 **Figure 5 Sex-determining region of Nile tilapia and blue tilapia.** a) Distribution of
690 male-specific SNP (number of SNPs per 50k window) in *ONJp*. b) The zoom-in view of
691 the sex-determining region on LG23. The coverage of YY male was calculated for each
692 5kb window. The location of the sex-determining gene *Amhy* is indicated by a vertical
693 dashed line. c) Distribution of female-specific SNP (number of SNPs per 50kb window)
694 in *OA*. d) The zoom-in view for the sex-determining region showing all female-specific
695 variants. e) The zoom-in view for the region that contains the verified female-specific
696 variants. The verified SNP is highlighted in red. The ratio of coverage of WW female
697 and ZZ male was calculated for every 5kb window. Windows with less than 60% base
698 pairs mapped are not shown. f) We examined oogonia and spermatogonia in the XX
699 and XY gonads of *ONJp* at 5 dah, when no morphological differences can be found

700 between sexes. At 30 dah, oogonia and oocytes can be observed in the XX gonads,
701 indicating the initiation of meiosis. But only spermatogonia can be found in the XY
702 gonad at 30 dah. At 180 dah, the XX gonads display large previtellogenetic oocytes, while
703 the XY gonads are characterized by the appearance of spermatogonia, spermatocytes
704 and spermatids. OG, oogonia; SG, spermatogonia; OC, oocytes; SC, spermatocytes;
705 ST, spermatids; SZ, spermatozoa. g) The expression profiles over three stages of
706 gonad development are shown for six known teleost SD genes and two candidate SD
707 genes of blue tilapia.

708

709

710 **References**

- 711 1. Salzburger W, Meyer A: **The species flocks of East African cichlid fishes: recent**
712 **advances in molecular phylogenetics and population genetics.**
713 *Naturwissenschaften* 2004, **91**:277-290.
- 714 2. Kocher TD: **Adaptive evolution and explosive speciation: the cichlid fish model.**
715 *Nat Rev Genet* 2004, **5**:288-298.
- 716 3. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW,
717 **Bezaula E, et al: The genomic substrate for adaptive radiation in African cichlid**
718 **fish.** *Nature* 2014, **513**:375-381.
- 719 4. Xiao J, Zhong H, Liu Z, Yu F, Luo Y, Gan X, Zhou Y: **Transcriptome analysis revealed**
720 **positive selection of immune-related genes in tilapia.** *Fish Shellfish Immunol* 2015,
721 **44**:60-65.
- 722 5. Pouyaud L, Agnèse JF: **Phylogenetic relationships between 21 species of three**
723 **tilapiine genera Tilapia, Sarotherodon and Oreochromis using allozyme data.** *J*
724 *Fish Biol* 1995.
- 725 6. Beardmore JA, Mair GC, Lewis RL: **Monosex male production in finfish as**
726 **exemplified by tilapia: applications, problems, and prospects.** In *Reproductive*
727 *Biotechnology in Finfish Aquaculture*. Edited by Lee C-S, Donaldson EM. Amsterdam:
728 Elsevier; 2001: 283-301
- 729 7. Mair GC, Abucay JS, Abella TA, Beardmore JA, Skibinski DOF: **Genetic manipulation**
730 **of sex ratio for the large-scale production of all-male tilapia Oreochromis**
731 **niloticus.** *Can J Fish Aquat Sci* 1997, **54**:396-404.
- 732 8. Cnaani A, Lee BY, Zilberman N, Ozouf-Costaz C, others: **Genetics of sex**
733 **determination in tilapiine species.** *Sexualities* 2008.
- 734 9. Baroiller JF, D'Cotta H, Bezaula E, Wessels S, Hoerstgen-Schwark G: **Tilapia sex**
735 **determination: Where temperature and genetics meet.** *Comp Biochem Physiol A Mol*
736 *Integr Physiol* 2009, **153**:30-38.
- 737 10. Gammerdinger WJ, Kocher TD: **Unusual diversity of sex chromosomes in African**
738 **cichlid fishes.** *Genes* 2018, **9**.
- 739 11. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW,
740 Kitano J, Mayrose I, Ming R, et al: **Sex determination: why so many ways of doing**
741 **it?** *PLoS Biol* 2014, **12**:e1001899.
- 742 12. Beukeboom LW, Perrin N: **The evolution of sex determination.** 2014.
- 743 13. Mank JE, Avise JC: **Evolutionary diversity and turn-over of sex determination in**
744 **teleost fishes.** *Sex Dev* 2009, **3**:60-67.
- 745 14. Kottler VA, Schartl M: **The Colorful Sex Chromosomes of Teleost Fish.** *Genes* 2018,
746 **9**.
- 747 15. Kikuchi K, Hamaguchi S: **Novel sex-determining genes in fish and sex chromosome**
748 **evolution.** *Dev Dyn* 2013.
- 749 16. Cnaani A: **The tilapias' chromosomes influencing sex determination.** *Cytogenet*
750 *Genome Res* 2013, **141**:195-205.
- 751 17. Campos-Ramos R, Harvey SC, Penman DJ: **Sex-specific differences in the**
752 **synaptonemal complex in the genus Oreochromis (Cichlidae).** *Genetica* 2009,
753 **135**:325-332.
- 754 18. Carrasco LAP, Penman DJ, Bromage N: **Evidence for the presence of sex**
755 **chromosomes in the Nile tilapia (Oreochromis niloticus) from synaptonemal**
756 **complex analysis of XX, XY and YY genotypes.** *Aquaculture* 1999.
- 757 19. Ocalewicz K, Mota-Velasco JC, Campos-Ramos R, others: **FISH and DAPI staining of**
758 **the synaptonemal complex of the Nile tilapia (Oreochromis niloticus) allow**

759 orientation of the unpaired region of bivalent 1 observed during *Chromosome*
 760 2009.

761 20. Gammerdinger WJ, Conte MA, Acquah EA, Roberts RB, Kocher TD: **Structure and**
 762 **decay of a proto-Y region in Tilapia, *Oreochromis niloticus*.** *BMC Genomics* 2014,
 763 15:975.

764 21. Lee B-Y, Coutanceau J-P, Ozouf-Costaz C, D'Cotta H, Baroiller J-F, Kocher TD:
 765 **Genetic and physical mapping of sex-linked AFLP markers in Nile tilapia**
 766 **(*Oreochromis niloticus*).** *Mar Biotechnol* 2011, 13:557-562.

767 22. Conte MA, Gammerdinger WJ, Bartie KL, Penman DJ, Kocher TD: **A high quality**
 768 **assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure**
 769 **of two sex determination regions.** *BMC Genomics* 2017, 18:341.

770 23. Sun Y-L, Jiang D-N, Zeng S, Hu C-J, Ye K, Yang C, Yang S-J, Li M-H, Wang D-S:
 771 **Screening and characterization of sex-linked DNA markers and marker-assisted**
 772 **selection in the Nile tilapia (*Oreochromis niloticus*).** *Aquaculture* 2014, 433:19-27.

773 24. Eshel O, Shirak A, Dor L, Band M, others: **Identification of male-specific amh**
 774 **duplication, sexually differentially expressed genes and microRNAs at early**
 775 **embryonic development of Nile tilapia.** *Biomed Chromatogr* 2014.

776 25. Li M, Sun Y, Zhao J, Shi H, Zeng S, Ye K, Jiang D, Zhou L, Sun L, Tao W, et al: **A**
 777 **tandem duplicate of anti-Müllerian hormone with a missense SNP on the Y**
 778 **chromosome is essential for male sex determination in Nile tilapia, *Oreochromis***
 779 ***niloticus*.** *PLoS Genet* 2015, 11:e1005678.

780 26. Cáceres G, López ME, Cádiz MI, others: **Fine mapping using whole-genome**
 781 **sequencing confirms anti-Müllerian hormone as a major gene for sex**
 782 **determination in farmed Nile tilapia (*Oreochromis niloticus* L** *G3: Genes,*
 783 *Genomes* 2019.

784 27. Li M-H, Yang H-H, Li M-R, Sun Y-L, Jiang X-L, Xie Q-P, Wang T-R, Shi H-J, Sun L-N,
 785 Zhou L-Y, Wang D-S: **Antagonistic roles of Dmrt1 and Foxl2 in sex differentiation**
 786 **via estrogen production in tilapia as demonstrated by TALENs.** *Endocrinology*
 787 2013, 154:4814-4825.

788 28. Wang D-S, Kobayashi T, Zhou L-Y, Paul-Prasanth B, Ijiri S, Sakai F, Okubo K,
 789 Morohashi K-i, Nagahama Y: **Foxl2 up-regulates aromatase gene transcription in a**
 790 **female-specific manner by binding to the promoter as well as interacting with Ad4**
 791 **binding protein/steroidogenic factor 1.** *Molecular Endocrinology* 2007, 21:712-725.

792 29. Conte MA, Kocher TD: **An improved genome reference for the African cichlid,**
 793 ***Metriaclima zebra*.** *BMC Genomics* 2015, 16:724.

794 30. Conte MA, Joshi R, Moore EC, Nandamuri SP, Gammerdinger WJ, Roberts RB,
 795 Carleton KL, Lien S, Kocher TD: **Chromosome-scale assemblies reveal the**
 796 **structural evolution of African cichlid genomes.** *Gigascience* 2019, 8.

797 31. Foresti F, Oliveira C, Galetti Junior PM, Almeida-Toledo LF: **Synaptonemal complex**
 798 **analysis in spermatocytes of tilapia, *Oreochromis niloticus* (Pisces, Cichlidae).**
 799 *Genome* 1993, 36:1124-1128.

800 32. Nikaido M, Suzuki H, Toyoda A, Fujiyama A, Hagino-Yamagishi K, Kocher TD, Carleton
 801 K, Okada N: **Lineage-specific expansion of vomeronasal type 2 receptor-like (OlfC)**
 802 **genes in cichlids may contribute to diversification of amino acid detection**
 803 **systems.** *Genome Biol Evol* 2013, 5:711-722.

804 33. Martins C, Oliveira C, Wasko AP, Wright JM: **Physical mapping of the Nile tilapia**
 805 **(*Oreochromis niloticus*) genome by fluorescent in situ hybridization of repetitive**
 806 **DNAAs to metaphase chromosomes—a review.** *Aquaculture* 2004, 231:37-49.

807 34. Ferreira IA, Poletto AB, Kocher TD, Mota-Velasco JC, Penman DJ, Martins C:
 808 **Chromosome evolution in African cichlid fish: contributions from the physical**
 809 **mapping of repeated DNAAs.** *Cytogenet Genome Res* 2010, 129:314-322.

810 35. Chew JSK, Oliveira C, Wright JM, Dobson MJ: **Molecular and cytogenetic analysis of**
811 **the telomeric (TTAGGG)n repetitive sequences in the Nile tilapia, Oreochromis**
812 **niloticus (Teleostei: Cichlidae)**. *Chromosoma* 2002, **111**:45-52.

813 36. Franck JP, Wright JM, McAndrew BJ: **Genetic variability in a family of satellite DNAs**
814 **from tilapia (Pisces: Cichlidae)**. *Genome* 1992, **35**:719-725.

815 37. Franck JP, Kornfield I, Wright JM: **The utility of SATA satellite DNA sequences for**
816 **inferring phylogenetic relationships among the three major genera of tilapiine**
817 **cichlid fishes**. *Mol Phylogen Evol* 1994, **3**:10-16.

818 38. Oliveira C, Wright JM: **Molecular cytogenetic analysis of heterochromatin in the**
819 **chromosomes of tilapia, Oreochromis niloticus (Teleostei: Cichlidae)**.
Chromosome Res 1998, **6**:205-211.

820 39. Muller H, Gil J, Drinnenberg IA: **The impact of centromeres on spatial genome**
821 **architecture**. *Trends in Genetics* 2019, **35**:565-578.

822 40. Ichikawa K, Tomioka S, Suzuki Y, Nakamura R, Doi K, Yoshimura J, Kumagai M, Inoue
823 Y, Uchida Y, Irie N, et al: **Centromere evolution and CpG methylation during**
824 **vertebrate speciation**. *Nat Commun* 2017, **8**:1833.

825 41. Supiwong W, Tanomtong A, Supanuam P, Seetapan K, Khakhong S, Sanoamuang L-O:
826 **Chromosomal characteristic of Nile tilapia (Oreochromis niloticus) from mitotic**
827 **and meiotic cell division by T-Lymphocyte cell culture**. *CYTOTOLOGIA* 2013, **78**:9-14.

828 42. Poletto AB, Ferreira IA, Cabral-de-Mello DC, Nakajima RT, Mazzuchelli J, Ribeiro HB,
829 Venere PC, Nirchio M, Kocher TD, Martins C: **Chromosome differentiation patterns**
830 **during cichlid fish evolution**. *BMC Genet* 2012, **13**:2.

831 43. Symonová R: **Integrative rDNAomics—importance of the oldest repetitive fraction**
832 **of the eukaryote genome**. *Genes* 2019, **10**:345.

833 44. Willard HF, Waye JS: **Hierarchical order in chromosome-specific human alpha**
834 **satellite DNA**. *Trends Genet* 1987, **3**:192-198.

835 45. Gammerdinger WJ, Conte MA, Sandkam BA, others: **Characterization of sex**
836 **chromosomes in three deeply diverged species of Pseudocrenilabrinae (Teleostei:**
837 **Cichlidae)**. *Hydrobiologia* 2019.

838 46. Bolívar P, Mugal CF, Nater A, Ellegren H: **Recombination rate variation modulates**
839 **gene sequence evolution mainly via GC-biased gene conversion, not Hill-**
840 **Robertson interference, in an avian system**. *Mol Biol Evol* 2016, **33**:216-227.

841 47. Senti K-A, Brennecke J: **The piRNA pathway: a fly's perspective on the guardian of**
842 **the genome**. *Trends Genet* 2010, **26**:499-509.

843 48. Shirak A, Zak T, Dor L, Benet-Perlberg A, Weller JI, Ron M, Seroussi E: **Quantitative**
844 **trait loci on LGs 9 and 14 affect the reproductive interaction between two**
845 **Oreochromis species, O. niloticus and O. aureus**. *Heredity* 2019, **122**:341-353.

846 49. Kashimada K, Koopman P: **Sry: the master switch in mammalian sex determination**.
847 *Development* 2010, **137**:3921-3930.

848 50. Tao W, Yuan J, Zhou L, Sun L, Sun Y, Yang S, Li M, Zeng S, Huang B, Wang D:
849 **Characterization of gonadal transcriptomes from Nile tilapia (Oreochromis**
850 **niloticus) reveals differentially expressed genes**. *PLoS One* 2013, **8**:e63604.

851 51. Zhang X, Li M, Ma H, Liu X, Shi H, Li M, others: **Mutation of foxI2 or cyp19a1a results**
852 **in female to male sex reversal in XX Nile tilapia**. *Endocrinology* 2017.

853 52. Jiang D-N, Yang H-H, Li M-H, Shi H-J, Zhang X-B, Wang D-S: **gsdf is a downstream**
854 **gene of dmrt1 that functions in the male sex determination pathway of the Nile**
855 **tilapia**. *Mol Reprod Dev* 2016, **83**:497-508.

856 53. Wei L, Li X, Li M, Tang Y, Wei J, Wang D: **Dmrt1 directly regulates the transcription**
857 **of the testis-biased Sox9b gene in Nile tilapia (Oreochromis niloticus)**. *Gene* 2019,
858 **687**:109-115.

859

860 54. Tang Y, Li X, Xiao H, Li M, Li Y, Wang D, Wei L: **Transcription of the Sox30 Gene Is**
861 **Positively Regulated by Dmrt1 in Nile Tilapia.** *International Journal of Molecular*
862 *Sciences* 2019, **20**:5487.

863 55. Lin Y-T, Capel B: **Cell fate commitment during mammalian sex determination.**
864 *Current Opinion in Genetics & Development* 2015, **32**:144-152.

865 56. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R: **Whole-**
866 **genome sequences of Malawi cichlids reveal multiple radiations interconnected**
867 **by gene flow.** *Nat Ecol Evol* 2018, **2**:1940-1955.

868 57. Svardal H, Quah FX, Malinsky M, Ngatunga BP, Miska EA, Salzburger W, Genner MJ,
869 Turner GF, Durbin R: **Ancestral hybridization facilitated species diversification in**
870 **the Lake Malawi cichlid fish adaptive radiation.** *Molecular Biology and Evolution*
871 2019.

872 58. Meier JI, Stelkens RB, Joyce DA, Mwaiko S, Phiri N, Schliewen UK, Selz OM, Wagner
873 CE, Katongo C, Seehausen O: **The coincidence of ecological opportunity with**
874 **hybridization explains rapid adaptive radiation in Lake Mweru cichlid fishes.** *Nat*
875 *Commun* 2019, **10**:5391.

876 59. Herpin A, Schartl M: **Plasticity of gene-regulatory networks controlling sex**
877 **determination: of masters, slaves, usual suspects, newcomers, and usurpators.**
878 *EMBO Rep* 2015, **16**:1260-1274.

879 60. O'Meally D, Ezaz T, Georges A, Sarre SD, Graves JAM: **Are some chromosomes**
880 **particularly good at sex? Insights from amniotes.** *Chromosome Res* 2012, **20**:7-19.

881 61. Graves JAM, Marshall Graves JA, Peichel CL: **Are homologies in vertebrate sex**
882 **determination due to shared ancestry or to limited options?** *Genome Biology* 2010,
883 **11**:205.

884 62. Denton RD, Kudra RS, Malcom JW, Du Preez L, Malone JH: **The African Bullfrog**
885 **(*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making**
886 **vertebrate sex chromosomes.**

887 63. Parnell NF, Streelman JT: **Genetic interactions controlling sex and color establish**
888 **the potential for sexual conflict in Lake Malawi cichlid fishes.** *Heredity* 2013,
889 **110**:239-246.

890 64. Charlesworth D, Charlesworth B, Marais G: **Steps in the evolution of heteromorphic**
891 **sex chromosomes.** *Heredity* 2005, **95**:118-128.

892 65. Roberts RB, Ser JR, Kocher TD: **Sexual Conflict Resolved by Invasion of a Novel**
893 **Sex Determiner in Lake Malawi Cichlid Fishes.** *Science* 2009, **326**:998-1001.

894 66. Ser JR, Roberts RB, Kocher TD: **Multiple interacting loci control sex determination**
895 **in lake Malawi cichlid fish.** *Evolution* 2010, **64**:486-501.

896 67. Bergero R, Gardner J, Bader B, Yong L, Charlesworth D: **Exaggerated heterochiasmy**
897 **in a fish with sex-linked male coloration polymorphisms.** *Proceedings of the*
898 *National Academy of Sciences* 2019, **116**:6924-6931.

899 68. Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, Shan Z, Haaf T,
900 Shimizu N, Shima A, et al: **A duplicated copy of DMRT1 in the sex-determining**
901 **region of the Y chromosome of the medaka, *Oryzias latipes*.** *Proceedings of the*
902 *National Academy of Sciences* 2002, **99**:11778-11783.

903 69. Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, Kobayashi T, Morrey CE,
904 Shibata N, Asakawa S, Shimizu N, et al: **DMY is a Y-specific DM-domain gene**
905 **required for male development in the medaka fish.** *Nature* 2002, **417**:559-563.

906 70. Takehana Y, Matsuda M, Myosho T, Suster ML, Kawakami K, Shin-I T, Kohara Y, Kuroki
907 Y, Toyoda A, Fujiyama A, et al: **Co-option of Sox3 as the male-determining factor on**
908 **the Y chromosome in the fish *Oryzias dancena*.** *Nature Communications* 2014, **5**.

909 71. Bao L, Tian C, Liu S, Zhang Y, Elaswad A, Yuan Z, Khalil K, Sun F, Yang Y, Zhou T, et
910 al: **The Y chromosome sequence of the channel catfish suggests novel sex**
911 **determination mechanisms in teleost fish.** *BMC Biology* 2019, **17**.

912 72. Yano A, Guyomard R, Nicol B, Jouanno E, Quillet E, Klopp C, Cabau C, Bouchez O,
913 Fostier A, Guiguen Y: **An immune-related gene evolved into the master sex-**
914 **determining gene in rainbow trout, *Oncorhynchus mykiss*.** *Curr Biol* 2012, **22**:1423-
915 1428.

916 73. Bertho S, Herpin A, Branthonne A, Jouanno E, Yano A, Nicol B, Muller T, Pannetier M,
917 Pailhoux E, Miwa M, et al: **The unusual rainbow trout sex determination gene**
918 **hijacked the canonical vertebrate gonadal differentiation pathway.** *Proc Natl Acad*
919 *Sci U S A* 2018, **115**:12781-12786.

920 74. Mazzuchelli J, Kocher TD, Yang F, Martins C: **Integrating cytogenetics and genomics**
921 **in comparative evolutionary studies of cichlid fish.** *BMC Genomics* 2012, **13**:463.

922 75. Xie Q-P, He X, Sui Y-N, Chen L-L, Sun L-N, Wang D-S: **Haploinsufficiency of SF-1**
923 **Causes Female to Male Sex Reversal in Nile Tilapia, *Oreochromis niloticus*.**
924 *Endocrinology* 2016, **157**:2500-2514.

925 76. Hopkins KD, Shelton WL, Engle CR: **Estrogen sex-reversal of Tilapia aurea.**
926 *Aquaculture* 1979, **18**:263-268.

927 77. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit
928 I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive mapping of long-range**
929 **interactions reveals folding principles of the human genome.** *Science* 2009,
930 **326**:289-293.

931 78. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: **Assembly of long, error-prone reads**
932 **using repeat graphs.** *Nat Biotechnol* 2019, **37**:540-546.

933 79. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly**
934 **from long uncorrected reads.** *Genome Res* 2017, **27**:737-746.

935 80. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
936 **34**:3094-3100.

937 81. Roach MJ, Schmidt SA, Borneman AR: **Purge Haplotype: allelic contig reassignment**
938 **for third-gen diploid genome assemblies.** *BMC Bioinformatics* 2018, **19**:460.

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

942 83. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL: **Juicer**
943 **provides a one-click system for analyzing loop-resolution Hi-C experiments.** *Cell*
944 *Syst* 2016, **3**:95-98.

945 84. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostafa R, Pham M, St
946 Hilaire BG, Yao W, Stamenova E, et al: **The Juicebox Assembly Tools module**
947 **facilitates de novo assembly of mammalian genomes with chromosome-length**
948 **scaffolds for under \$1000.**

949 85. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
950 Wortman J, Young SK, Earl AM: **Pilon: an integrated tool for comprehensive**
951 **microbial variant detection and genome assembly improvement.** *PLoS One* 2014,
952 **9**:e112963.

953 86. Seppey M, Manni M, Zdobnov EM: **BUSCO: Assessing Genome Assembly and**
954 **Annotation Completeness.** *Methods Mol Biol* 2019, **1962**:227-245.

955 87. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A,
956 Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging**
957 **model organism genomes.** *Genome Res* 2008, **18**:188-196.

958 88. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
959 Raychowdhury R, Zeng Q, et al: **Full-length transcriptome assembly from RNA-Seq**
960 **data without a reference genome.** *Nature Biotechnology* 2011, **29**:644-652.

961 89. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R,
962 Ronning CM, Rusch DB, Town CD, et al: **Improving the *Arabidopsis* genome**
963 **annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res*
964 2003, **31**:5654-5666.

965 90. Seibt KM, Schmidt T, Heitkam T: **FlexiDot: highly customizable, ambiguity-aware**
966 **dotplots for visual sequence analyses.** *Bioinformatics* 2018, **34**:3575-3577.

967 91. Tao W, Chen J, Tan D, Yang J, Sun L, Wei J, Conte MA, Kocher TD, Wang D:
968 **Transcriptome display during tilapia sex determination and differentiation as**
969 **revealed by RNA-Seq analysis.** *BMC Genomics* 2018, **19**:363.

970 92. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL: **Graph-based genome alignment**
971 **and genotyping with HISAT2 and HISAT-genotype.** *Nature Biotechnology* 2019,
972 **37**:907-915.

973 93. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for**
974 **assigning sequence reads to genomic features.** *Bioinformatics* 2014, **30**:923-930.

975 94. Qiang J, Bao WJ, Tao FY, He J, Li XH, Xu P, Sun LY: **The expression profiles of**
976 **miRNA-mRNA of early response in genetically improved farmed tilapia**
977 **(*Oreochromis niloticus*) liver by acute heat stress.** *Sci Rep* 2017, **7**:8705.

978 95. Tao W, Sun L, Shi H, Cheng Y, Jiang D, Fu B, Conte MA, Gammerdinger WJ, Kocher
979 TD, Wang D: **Integrated analysis of miRNA and mRNA expression profiles in tilapia**
980 **gonads at an early stage of sex differentiation.** *BMC Genomics* 2016, **17**:328.

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

983 97. Pantano L, Estivill X, Martí E: **A non-biased framework for the annotation and**
984 **classification of the non-miRNA small RNA transcriptome.** *Bioinformatics* 2011,
985 **27**:3202-3203.

986 98. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient**
987 **alignment of short DNA sequences to the human genome.** *Genome Biol* 2009,
988 **10**:R25.

989 99. Rosenkranz D, Zischler H: **proTRAC - a software for probabilistic piRNA cluster**
990 **detection, visualization and analysis.** *BMC Bioinformatics* 2012, **13**:1-10.

991 100. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA,
992 Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al: **Scaling accurate genetic**
993 **variant discovery to tens of thousands of samples.**

994 101. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd
995 ML, Nagarajan N: **LoFreq: a sequence-quality aware, ultra-sensitive variant caller**
996 **for uncovering cell-population heterogeneity from high-throughput sequencing**
997 **datasets.** *Nucleic Acids Res* 2012, **40**:11189-11201.

998 102. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
999 Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format**
1000 **and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.

1001 103. Yan L, Feng H, Wang F, Lu B, Liu X, Sun L, Wang D: **Establishment of three estrogen**
1002 **receptors (esr1, esr2a, esr2b) knockout lines for functional study in Nile tilapia.** *J*
1003 *Steroid Biochem Mol Biol* 2019, **191**:105379.

1004

1005

1006 **Table 1 Genome assembly and annotation statistics**

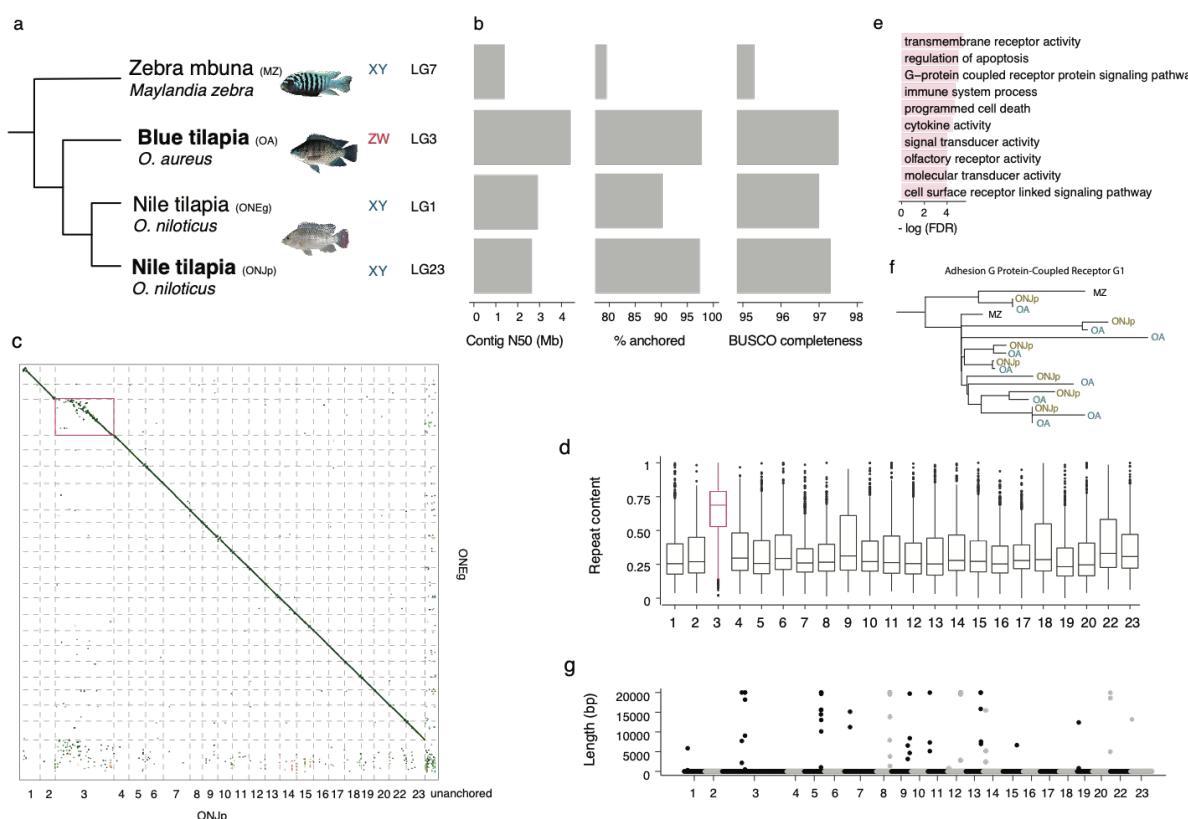
Assembly	ONEg	ONJp	OA
Sequencing platform	Pacbio	Nanopore	Nanopore
Coverage	44X	96X	85X
Assembly size	1,005,681,550	993,468,885	1,005,590,959
Contig N50	2,923,640	2,651,554	4,404,323
# contig	3,010	1,201	805
Scaffold N50	38,839,487	40,346,024	40,723,988
# scaffold	2,460	403	303
% anchored	90.20%	97.40%	97.80%
complete BUSCOs	97.00%	97.30%	97.50%
Fragmented BUSCOs	1.60%	1.40%	1.30%
Missing BUSCOs	1.40%	1.30%	1.20%
# gene	29,537	25,264	25,467
Repeat content	36.5	40.4	39.4

1007

1008

1009

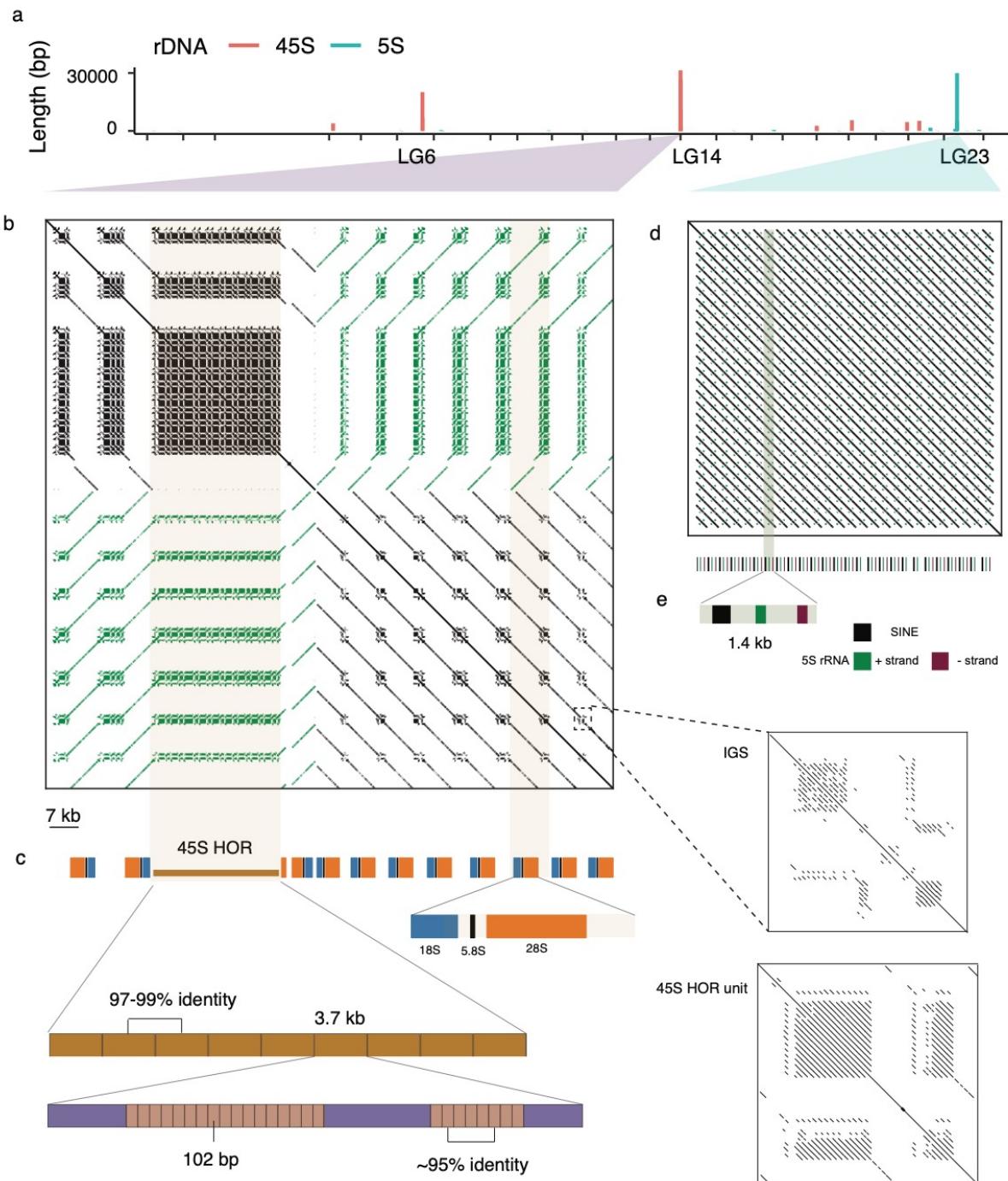
Figure 1 High-quality genome assemblies at the chromosome level.



1010

1011

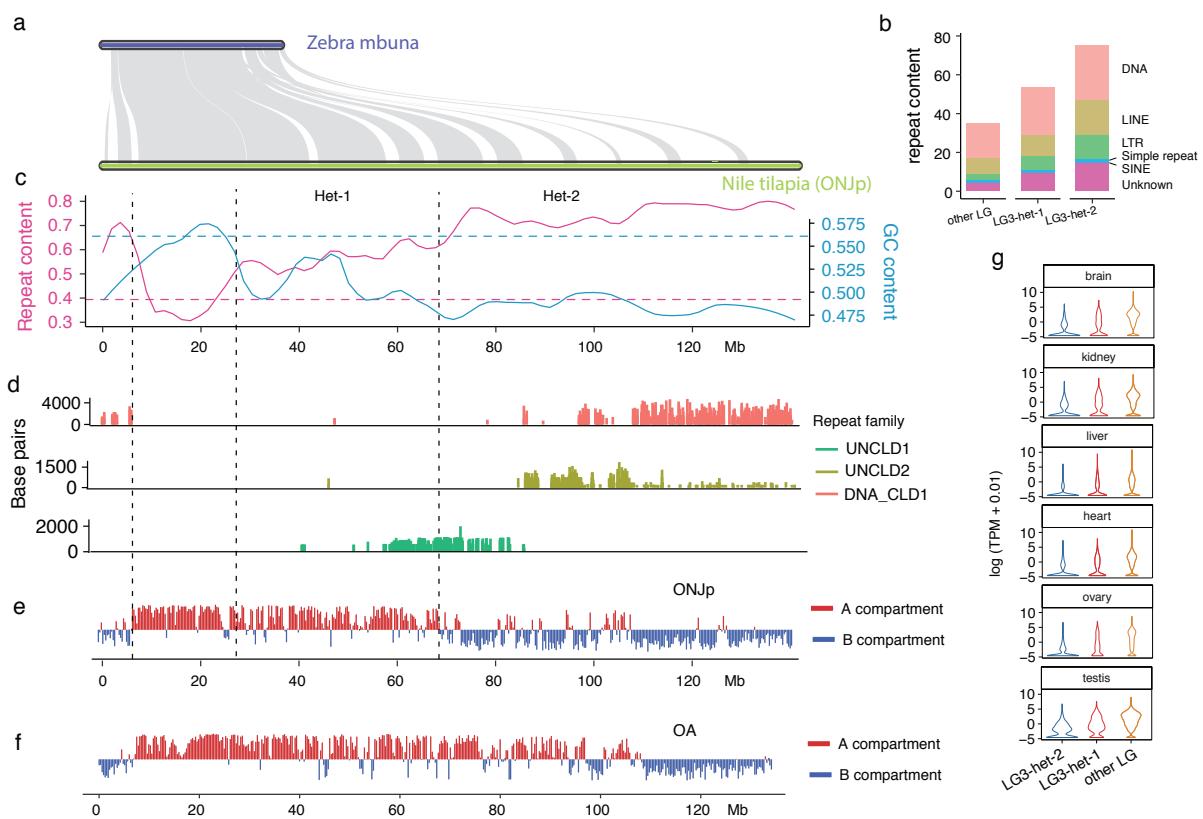
1012 **Figure 2 The genomic organization of rDNA loci**



1013

1014

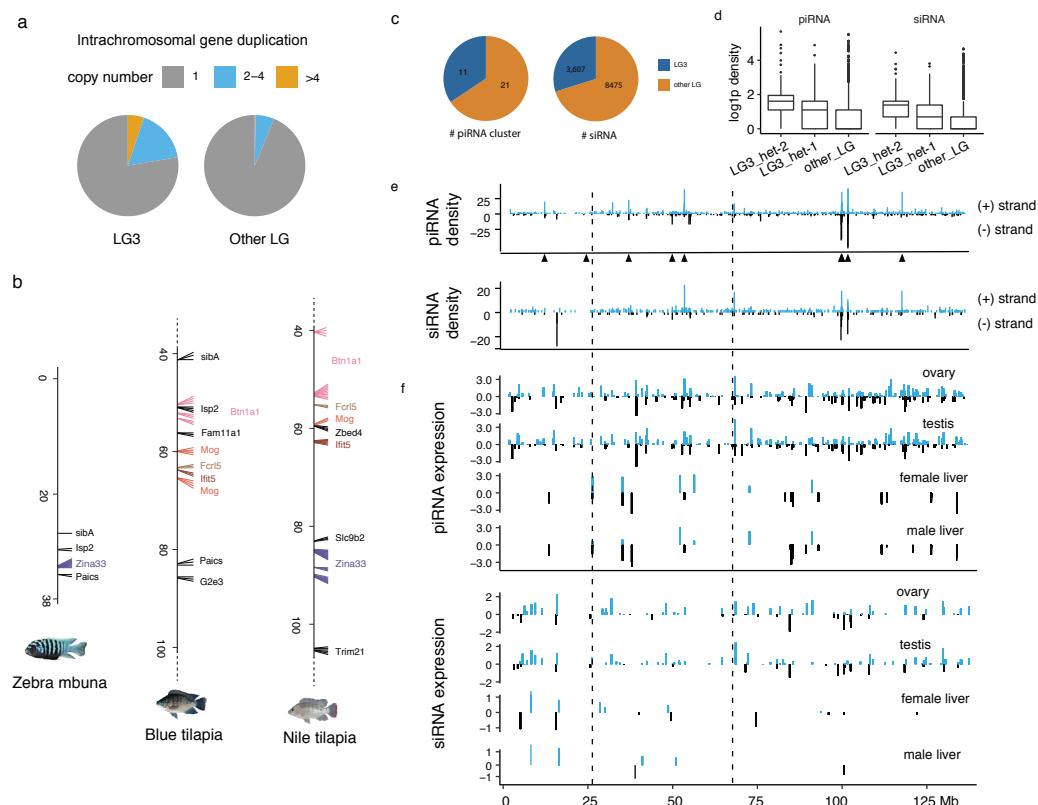
1015 **Figure 3 Heterochromatin region of LG3**



1016

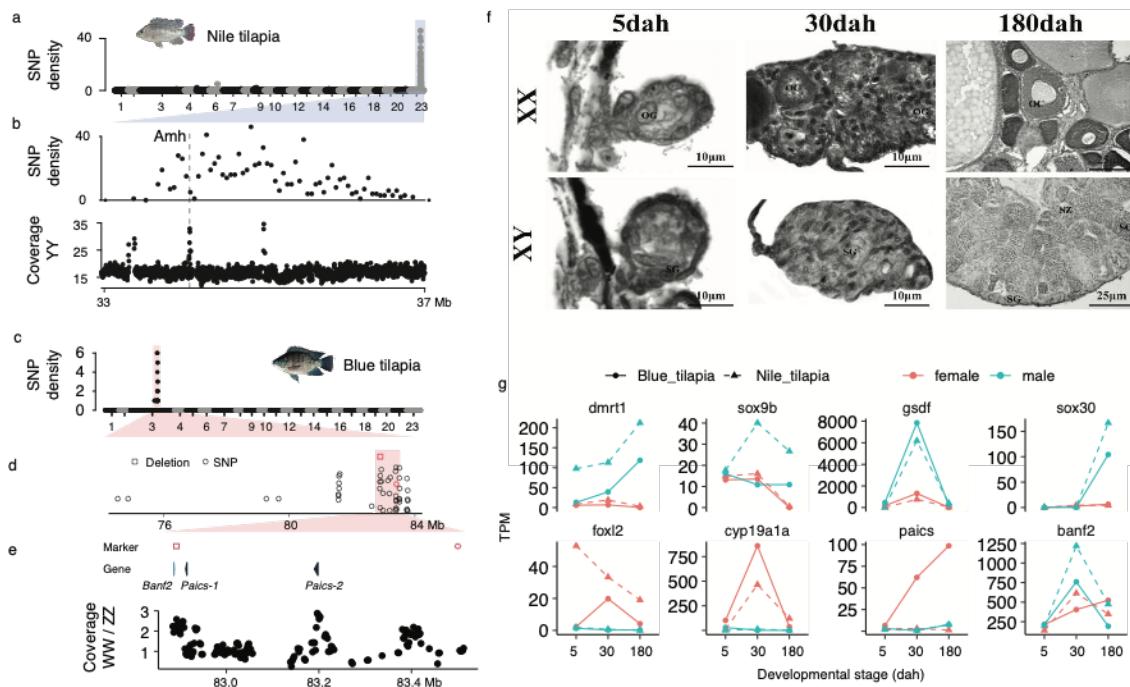
1017

1018 **Figure 4 LG3 heterochromatin contains tandem arrays of mRNA and sRNA genes**



1019

1020 **Figure 5 Sex-determining region of Nile tilapia and blue tilapia.**



1021