Genome sequence of the euryhaline Javafish medaka, Oryzias javanicus: a small aquarium fish model for studies on adaptation to salinity

Yusuke Takehana, Margot Zahm, Cédric Cabau, Christophe C. Klopp, Céline Roques, Olivier Bouchez, Cécile Donnadieu, Célia Barrachina, Laurent Journot, Mari Kawaguchi, et al.

► **To cite this version:**

Yusuke Takehana, Margot Zahm, Cédric Cabau, Christophe C. Klopp, Céline Roques, et al.. Genome sequence of the euryhaline Javafish medaka, Oryzias javanicus: a small aquarium fish model for studies on adaptation to salinity. 2019. hal-02790965

HAL Id: hal-02790965
https://hal.inrae.fr/hal-02790965

Preprint submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Genome sequence of the euryhaline Javafish medaka, *Oryzias javanicus*: a small aquarium fish model for studies on adaptation to salinity.

Yusuke Takehana[1], Margot Zahm[2], Cédric Cabau[3], Christophe Klopp[2, 3], Céline Roques[4], Olivier Bouchez[4], Cécile Donnadieu[4], Celia Barrachina[5], Laurent Journot[5], Mari Kawaguchi[6], Shigeki Yasumasu[6], Satoshi Ansai[7], Kiyoshi Naruse[7], Koji Inoue[8], Chuya Shinzato[8], Manfred Schartl[9, 10, 11], Yann Guiguen[12, *] and Amaury Herpin[12, *, ¶].

AFFILIATIONS

[1] Department of Animal Bio-Science, Faculty of Bio-Science, Nagahama Institute of Bioscience and Technology, 1266 Tamura, Nagahama 526-0829, Japan.

[2] Plate-forme bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRA, Castanet Tolosan, France.

[3] SIGENAE, GenPhySE, Université de Toulouse, INRA, ENVT, Castanet Tolosan, France.

[4] INRA, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France

[5] MGX, Biocampus Montpellier, CNRS, INSERM, University of Montpellier, Montpellier, France.

[6] Department of Materials and Life Sciences, Faculty of Science and Technology, Sophia University, 7-1 Kioi-cho Chiyoda-ku, Tokyo 102-8554, Japan.

[7] Laboratory of Bioresources, National Institute for Basic Biology, 38 Nishigonaka, Myodaiji-cho, Okazaki 444-8585, Japan.

[8] Atmosphere and Ocean Research Institute, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8564, Japan.

[9] University of Wuerzburg, Developmental Biochemistry, Biocenter, 97074 Wuerzburg, Germany.

[10] Comprehensive Cancer Center Mainfranken, University Hospital, 97080 Wuerzburg, Germany.

[11] Hagler Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, Texas 77843, USA.

[12] INRA, UR 1037 Fish Physiology and Genomics, F-35000 Rennes, France.

** These authors contributed equally to this work. ¶ Corresponding author.*

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

41 ## ABSTRACT

42 **Background**: The genus *Oryzias* is constituted of 35 medaka-fish species each exhibiting various
43 ecological, morphological and physiological peculiarities and adaptations. Beyond of being a
44 comprehensive phylogenetic group for studying intra-genus evolution of several traits like sex
45 determination, behaviour, morphology or adaptation through comparative genomic approaches,
46 all medaka species share many advantages of experimental model organisms including small size
47 and short generation time, transparent embryos and genome editing tools for reverse and
48 forward genetic studies. The Java medaka, *Oryzias javanicus*, is one of the two species of medaka
49 perfectly adapted for living in brackish/sea-waters. Being an important component of the
50 mangrove ecosystem, *O. javanicus* is also used as a valuable marine test-fish for ecotoxicology
51 studies. Here, we sequenced and assembled the whole genome of *O. javanicus*, and anticipate this
52 resource will be catalytic for a wide range of comparative genomic, phylogenetic and functional
53 studies. **Findings**: Complementary sequencing approaches including long-read technology and
54 data integration with a genetic map allowed the final assembly of 908 Mbp of the *O. javanicus*
55 genome. Further analyses estimate that the *O. javanicus* genome contains 33% of repeat
56 sequences and has a heterozygosity of 0.96%. The achieved draft assembly contains 525 scaffolds
57 with a total length of 809.7 Mbp, a N50 of 6,3 Mbp and a L50 of 37 scaffolds. We identified 21454
58 expressed transcripts for a total transcriptome size of 57, 146, 583 bps. **Conclusions**: We provide
59 here a high-quality draft genome assembly of the euryhaline Javafish medaka, and give emphasis
60 on the evolutionary adaptation to salinity.
61
62
63
64 ## KEYWORDS

65 Medaka, evolution, whole genome sequencing, long reads, genetic map, transcriptome,
66 adaptation, salinity.
67
68
69
70
71
72
73
74
75
76
77
78

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

79 **DATA DESCRIPTION**

80 Introduction/Background information:

81 Medaka fishes belong to the genus *Oryzias* and are an emerging model system for studying the
82 molecular basis of vertebrate evolution. This genus contains approximately 35 species,
83 individually exhibiting numerous morphological, ecological and physiological differences and
84 specificities (1–4). In addition, they all share many advantages of experimental model organisms,
85 such as their small size, easy breeding, short generation time, transparent embryos, transgenic
86 technology and genome-editing tools, with the "flag ship" species of this genus, the Japanese rice
87 fish, *Oryzias latipes* (5, 6). Such phenotypic variations together with cutting edge molecular
88 genetic tools make possible to identify major loci that contribute to evolutionary differences, and
89 to dissect the roles of individual genes and regulatory elements by functional tests. For example,
90 a recent genetic mapping approach using interspecific hybrids identified the major chromosome
91 regions that underlie the different hyperosmotic tolerance between species of the *Oryzias* genus
92 (7). Medaka fishes are also excellent models to study evolution of sex chromosomes and sex-
93 determining loci among species (8–11), with the advantage of being also suitable models for
94 providing functional evidences for these novel sex-determining genes by gain-of-function and/or
95 loss-of-function experiments (12, 13).

96

97 Among these species, the Java medaka, *Oryzias javanicus* (Figure 1), is unique as being the
98 prototypic species of this genus with respect to adaptation to seawater. Previous phylogenetic
99 studies divided the genus *Oryzias* into three monopyletic groups: *(i) javanicus*, *(ii) latipes* and *(iii)*
100 *celebensis* species groups (4, 14). Most of the *Oryzias* species inhabit mainly freshwater biotopes
101 while only two species, which belong to the *javanicus* group, live in sea- or brackish waters. One
102 is *O. javanicus*, found in mangrove swamps from Thailand to Indonesia, and the other is *O.*
103 *dancena* (previously named *O. melastigma*) living both in sea- and freshwaters from India to
104 Malaysia. Although both species are highly adaptable to seawater, *O. javanicus* prefers
105 hyperosmotic conditions while *O. dancena* favours hypoosmotic conditions at the west coast of
106 Malaysian peninsula where their distribution ranges overlap (15). In addition, *O. javanicus* is an
107 important component of the mangrove ecosystem (16), and has been used as a valuable marine
108 test fish in several ecotoxicology studies (17, 18).

109

110 In this study, we sequenced and assembled the whole genome of *O. javanicus*, a model fish
111 species for studying molecular mechanisms of seawater adaptation. In teleost fish, the major
112 osmoregulatory organs i.e., gills, intestine and kidney, play different roles for maintaining body
113 fluid homeostasis. Many genes encoding hormones, receptors, osmolytes, transporters, channels
114 and cellular junction proteins are potentially involved in this osmotic regulation. In addition to
115 osmoregulation, hatching enzyme activity dramatically fluctuates and adjusts at different salt
116 conditions. At hatching stage, fish embryos secrete a specific cocktail of enzymes in order to
117 dissolve the egg envelope, or chorion. In the medaka *O. latipes*, digestion of the chorion occurs

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

118 through the cooperative action of two kinds of hatching enzymes, ***(i)*** the high choriolytic enzyme
119 (HCE) and ***(ii)*** the low choriolytic enzyme (LCE) (19). The HCE displays a higher activity in fresh-
120 than in brackish waters (20). Thus, availability of a high-quality reference genome in *O. javanicus*
121 would facilitate further research for investigating the molecular basis of physiological
122 differences, including the osmotic regulation and the hatching enzyme activity, among *Oryzias*
123 species.
124

125 SAMPLING AND SEQUENCING

126

127 Animal samplings

128 The wild stock of *O. javanicus* used in this study was supplied by the National Bio-Resource
129 Project (NBRP) medaka in Japan. This stock (strain ID: RS831) was originally collected at
130 Penang, Malaysia, and maintained in aquaria under an artificial photoperiod of 14 hours light:10
131 hours darkness at 27±2ºC. Genomic DNA was extracted from the whole body of a female (having
132 ZW sex chromosome) using a conventional phenol/chloroform method, and was subjected to
133 PacBio and 10X Genomics sequencings. For RNA-sequencing, total RNAs were extracted from
134 nine female tissues (brain, bone, gill, heart, intestine, kidney, liver, muscle and ovary), and one
135 male tissue (testis) using the RNeasy Mini Kit (Qiagen). For genetic mapping, we used a DNA
136 panel consisting of 96 F1 progeny with their parents (originally described in a previous study
137 (21)). Phenotypic sex was determined by secondary sex characteristics of adult fish, namely, the
138 shapes of dorsal and anal fins. All animal experiments performed in this study complied with the
139 guideline of National Institute for Basic Biology, and have been approved by the Institutional
140 Animal Care and Use Committee of National Institute of Natural Science (16A050 and 17A048).
141

142 Libraries construction and sequencing

143 ***PacBio genome sequencing***

144 Library construction and sequencing were performed according to the manufacturer's
145 instructions (Shared protocol-20kb Template Preparation Using BluePippin Size Selection
146 system (15kb size Cutoff)). When required, DNA was quantified using the Qubit dsDNA HS Assay
147 Kit (Life Technologies). DNA purity was assessed by spectrophotometry using the nanodrop
148 instrument (Thermofisher), and size distribution and absence of degradation were monitored
149 using the Fragment analyzer (AATI) (8–11). Purification steps were performed using 0.45X
150 AMPure PB beads (PacBio). 80µg of DNA was purified and then sheared at 40kb using the
151 megaruptor system (diagenode). DNA and END damage repair step was further performed for 5
152 libraries using the SMRTBell template Prep Kit 1.0 (PacBio). Blunt hairpin adapters were then
153 ligated to the libraries. Libraries were subsequently treated with an exonuclease cocktail in order
154 to digest unligated DNA fragments. Finally, a size selection step using a 15kb cutoff was
155 performed on the BluePippin Size Selection system (Sage Science) using 0.75% agarose cassettes,

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

156 Marker S1 high Pass 15-20kb. Conditioned sequencing primer V2 was annealed to the size-
157 selected SMRTbell. The annealed libraries were then bound to the P6-C4 polymerase using a
158 ratio of polymerase to SMRTbell set at 10:1. After performing a magnetic bead-loading step
159 (OCPW), SMRTbell libraries were sequenced on 48 SMRTcells (RSII instrument at 0.25nM with a
160 360-min movie resulting in a total of 61.8Gb of sequence data (1.28Gb/SMRTcell).

161 ***10X Genomics genome sequencing***

162 Chromium library was prepared according to 10X Genomics' protocol using the Genome Reagent
163 Kits v1. Sample quantity and quality controls were further validated on Qubit, Nanodrop and
164 Femto. Optimal performance has been characterized on input gDNA with a mean length greater
165 than 50 kb. The library was prepared using 3 µg of high molecular weight (HMW) gDNA (cut off
166 at 50kb using BluePippin system). In details, for the microfluidic Genome Chip, a library of
167 Genome Gel Beads was combined with HMW template gDNA in Master Mix and partitioning oil in
168 order to create Gel Bead-In-EMulsions (GEMs) in the Chromium. Each Gel Bead was
169 functionalized with millions of copies of a 10x™ Barcoded primer. Upon dissolution of the
170 Genome Gel Bead in the GEM, primers containing (*i*) an Illumina R1 sequence (Read 1 sequencing
171 primer), (*ii*) a 16 bp 10x Barcode, and (*iii*) a 6 bp random primer sequence were released. Read 1
172 sequence and the 10x™ Barcode were added to the molecules during the GEM incubation. P5 and
173 P7 primers, Read 2, and Sample Index were added during library construction. 8 cycles of PCR
174 were performed for amplifying the library. Library quality was assessed using a Fragment
175 analyzer. Finally, the library was sequenced on an Illumina HiSeq3000 using a paired-end read
176 length of 2x150 pb with the Illumina HiSeq3000 sequencing kits resulting in 101.6Gb of raw
177 sequence data.

178 ***Transcriptome RNA-seq sequencing***

179 RNA-seq libraries were prepared according to Illumina's protocols using the Illumina TruSeq
180 Stranded mRNA sample prep kit. Briefly, mRNAs were selected using poly-T beads, reverse-
181 transcribed and fragmented. The resulting cDNAs were then subjected to adaptor ligation. 10
182 cycles of PCR were performed for amplifying the libraries. Quality of the libraries was assessed
183 using a Fragment Analyser. Quantification was performed by qPCR using the Kapa Library
184 Quantification Kit. RNA-seq libraries were sequenced on an Illumina HiSeq3000 using a paired-
185 end read length of 2x150 pb with the Illumina HiSeq3000 sequencing kits resulting in 95Gb of
186 sequence data (28.9M reads pairs/library).

187

188 ***RAD-library construction***

189 RAD-seq library was built following the Baird et al. (22) protocol with minor modifications.
190 Briefly, between 400 to 500 ng of gDNA per fish were digested with SbfI-HF enzyme (R3642S,
191 NEB). Digested DNA was purified using AMPure PX magnetic beads (Beckman Coulters) and
192 ligated to indexed P1 adapters (1 index per sample) using concentrated T4 DNA ligase (M0202T,

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

193 NEB). After quantification (Qubit dsDNA HS assay kit, Thermofisher) all samples were pooled in
194 equal amounts. The pool was then fragmented on a S220 sonicator (Covaris) and purified with
195 Minelute column (Qiagen). Finally, the sonicated DNA was size selected (250 to 450 bps) on a
196 Pippin HT (Sage science) using a 2 % agarose cassette, repaired using the End-It DNA-end repair
197 kit (Tebu Bio) and adenylated at its 3' ends using Klenow (exo-) (Tebu-Bio). P2 adapters were
198 then ligated using concentrated T4 DNA ligase, and 50 ng of the ligation product were engaged in
199 a 12 cycles PCR for amplification. After AMPure XP beads purification, the resulting library was
200 checked on a Fragment Analyzer (Agilent) using the HS NGS kit (DNF-474-33) and quantified by
201 qPCR using the KAPA Library Quantification Kit (Roche, ref. KK4824). Ultimately the whole
202 library was denatured, diluted to 10 pM, clustered and sequenced using the rapid mode v2
203 SR100nt lane of a Hiseq2500 device (Illumina).
204

205 ## Assembly results and quality assessment

206

207 ### Genome Characteristics

208 To estimate size and other genome characteristics, 10X reads were processed with Jellyfish
209 v1.1.11 (23) to produce 21-mer distribution. The k-mer histogram was uploaded to
210 GenomeScope (24) with the max k-mer coverage parameter set to 10,000. Genome size was
211 estimated around 908 Mbp, which is slightly higher than the 850 Mbp (0.87pg) estimated size
212 reported on the Animal Genome Size Database (25). Furthermore, this analysis estimates that the
213 *O. javanicus* genome contains 33% of repeat sequences (around 303 Mbp) and has a
214 heterozygosity of 0.96% (Table 1).
215

Property	min	max
Heterozygosity	0.960 %	0.964 %
Genome Haploid Length	908,146,324 bp	908,641,143 bp
Genome Repeat Length	303,610,795 bp	303,776,222 bp
Genome Unique Length	604,535,529 bp	604,864,921 bp
Model Fit	95.95 %	99.72 %
Read Error Rate	1.50 %	1.50 %

216 **Table 1:** GenomeScope outputs on *O. javanicus* genome statistics.

217

218 ### Genome assembly with long PacBio reads and short 10X reads

219 PacBio reads were corrected and trimmed using Canu v1.5 (26). Contigs were then assembled
220 using SMARTdenovo version of May 2017 (27). The draft assembly produced contains 729
221 contigs with a total genome size of 807.5 Mbp, an N50 of 3,9 Mbp and a L50 of 59 contigs (Figure
222 2). To improve the assembly base pair quality two polishing steps were run. First, BLASR aligned
223 PacBio reads were processed with Quiver from the Pacific Biosciences SMRT link software
224 v.4.0.0. Second, 10X reads were realigned to the genome using Long Ranger v2.1.1 and the

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

alignment file was processed with Pilon v1.22 (28). Third, the same 10X reads were aligned to the genome with BWA-MEM v0.7.12-r1039 (29) and the alignment file was processed with ARCS v1.0.1 (30) to scaffold the genome. Both tools were run with default parameters. The final draft assembly contains 525 scaffolds with a total length of 809.7 Mbp, a N50 of 6,3 Mbp and a L50 of 37 scaffolds. This represents 89.1% of the k-mer estimated genome size. Given the high percentage of repeats in the *O. javanicus* genome (33%), it is possible that the PacBio assembly did not totally succeed in completing all repeated regions. The genome completeness was estimated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0 (31) based on 4,584 BUSCO orthologs derived from the Actinopterygii lineage leading to BUSCO scores of 4,327 (94.4%) complete BUSCOs, 176 (3.8%) fragmented BUSCOs and 81 (1.8%) missing BUSCOs.

Integration with the genetic map.

RAD reads were trimmed by Trim Galore 0.4.3 (32) with Cutadapt 1.12 (33) and then mapped to the assembled scaffolds using BWA-MEM v0.7.17 (29). Uniquely mapped reads were extracted from the read alignments, and then called variant bases using uniquely mapped reads by samtools *mpileup* and bcftools *call* (34). Indels and variants with a low genotyping quality (GQ < 20), a low read depth (DP < 5), a low frequency of the minor allele (< 5%), more than four alleles in the family, no more than 5% individuals missing were removed by vcftools v0.1.15 (35). After quality filtering, 6,375 variant sites were kept for the following analysis. Linkage map was constructed using this genotype information using Lep-MAP3 (36). Briefly, the filtered vcf file was loaded and the markers removed with high segregation distortion (*Filtering2*: dataTolerance=0.001). Markers were then separated into 24 linkage groups with a LOD score threshold set at 9 and a fixed recombination fraction of 0.08 (*SeparateChromosomes2*: lodlimit=9 and theta=0.08). Two linkage groups were then excluded because of their small numbers of contained markers (less than 10). Classification of the markers was determined after maximum likelihood score indexing with 100 iterations (*OrderMarkers2*: numMergeIterations=100) in each linkage group. The final map had 5,738 markers dispatched amongst 24 linkage groups spanning a total genetic distance of 1,221 cM.

The linkage map exhibited discrepancies between genomic scaffolds and genetic markers. Among 525 genomics scaffolds, 32 were linked to more than one linkage group. To split chimeric scaffolds with a higher precision and to rebuild chromosomes with a higher fidelity, we used a cross-species synteny map between the Java medaka (*O. javanicus*) scaffolds and the medaka *(O. latipes*) chromosomes in order to combine marker locations from genetic and synteny maps. To build the synteny map, medaka cDNAs were aligned to the Java medaka scaffolds using BLAT v36 (37), and a list of pairwise correspondence of gene positions on Java medaka scaffolds and medaka chromosomes was established. 13,796 markers were added to the 5,738 markers of the genetic map. Java medaka chromosomes were then reconstructed using ALLMAPS from the JCVI utility libraries v0.5.7 (38). This package was used to combine genetic and synteny maps, to split chimeric scaffolds, to anchor, order and orient genomic scaffolds. The resulting chromosomal

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

265 assembly consists of 321 scaffolds anchored on 24 chromosomes (97.7% of the total bases) and
266 231 unplaced scaffolds
267
268 **Transcriptome assembly**
269 The read quality of the RNA-seq libraries was evaluated using FastQC (39). *De novo* and
270 reference-based transcriptome assemblies were produced. Reads were cleaned, filtered and *de*
271 *novo* assembled using the DRAP pipeline v1.91 (40) with the Oases assembler (41). Assembled
272 contigs were filtered in order to keep only those with at least one fragment per kilobase of
273 transcript per million reads (FPKM). In the reference-based approach, all clean reads were
274 mapped to the chromosomal assembly using STAR v2.5.1b (42) with outWigType and
275 outWigStrand options to output signal wiggle files. Cufflinks v2.2.1 (43) was used to assemble the
276 transcriptome.
277
278 **Annotation results**
279 The first annotation step was identifying repetitive DNA content using RepeatMasker v4.0.7 (44),
280 Dust (45) and TRF v4.09 (46). A species-specific *de novo* repeat library was built with
281 RepeatModeler v1.0.11 (47). Repeated regions were located using RepeatMasker with the *de*
282 *novo* and the Zebrafish (*Danio rerio*) libraries. Bedtools v2.26.0 (48) was used to merge repeated
283 regions identified with the three tools and to soft mask the genome. Repeats were estimated to
284 account for 43.16% (349 Mbp) of our chromosomal assembly. The MAKER3 genome annotation
285 pipeline v3.01.02-beta (49) combined annotations and evidences from three approaches:
286 similarity with known fish proteins, assembled transcripts and *de novo* gene predictions. Protein
287 sequences from 11 other fish species found in Ensembl were aligned to the masked genome using
288 Exonerate v2.4 (50). Previously assembled transcripts were used as RNA-seq evidence. A *de novo*
289 gene model was built using Braker v2.0.4 (51) with wiggle files provided by STAR as hints file for
290 training GeneMark and Augustus. The best supported transcript for each gene was chosen using
291 the quality metric Annotation Edit Distance (AED) (52). The genome annotation gene
292 completeness was assessed by BUSCO using the Actinopterygii group (Table 2). Finally, the
293 predicted genes were subjected to similarity searches against the NCBI NR database using
294 Diamond v0.9.22 (53). The top hit with a coverage over 70% and identity over 80% was retained.
295
296
297
298
299
300
301
302
303

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

Gene annotation	
Number of genes	21,454
Number of transcripts	21,454
Transcriptome size	57,146,583 bp
Mean transcript length	2,663 bp
Longest transcript	42,733 bp
Number of genes with significant hit against NCBI NR	17,412 (81.2%)
Gene completeness	
Complete BUSCOs	4,289 (93.6%)
Fragmented BUSCOs	187 (4.1%)
Missing BUSCOs	108 (2.3%)

304 **Table 2: Java medaka assembly and annotation statistics.**

305

306 **Mitochondrial genome and annotation**

307 The previously sequenced *Oryzias javanicus* mitochondrial genome (NC_012981) (54) was
308 aligned to the chromosomal assembly using Blat. All hits were supported by a single scaffold.
309 This scaffold was removed from the assembly, circularised and annotated using MITOS (55). This
310 new *Oryzias javanicus* mitochondrial genome is 16,789 bp long and encodes 13 genes, 2 rRNAs
311 and 19 tRNAs.

312

313 **Phylogenetic relationship**

314 To precisely determine the phylogenetic position of *O. javanicus* within the genus *Oryzias*, we
315 estimated the phylogenetic relationship using published whole genome datasets as references.
316 Reference assemblies and annotations of *O. latipes* (Hd-rR: ASM223467v1), *O. sakaizumii* (HNI-II:
317 ASM223471v1), *Oryzias* sp. (HSOK: ASM223469v1), *O. melastigma* (Om_v0.7.RACA), and
318 southern platyfish *Xiphophorus maculatus* (X_maculatus-5.0-male) were obtained from Ensembl
319 Release 94 (http://www.ensembl.org/). Among the six genomes, orthologous groups were
320 classified and 10,852 single-copy orthologous genes were identified using OrthoFinder 2.2.6 (56).
321 For every single gene, codon alignment based on translated peptide sequences was generated by
322 PAL2NAL (57) and then trimmed by trimAl with '-autometed1' option (58). All multi-sample
323 fasta files were concatenated into a single file using AMAS *concat* by setting each gene as a
324 separate partition (59). A maximum likelihood tree was then inferred using IQ-TREE v1.6.6 (60)
325 with the GTR+G substitution model for each codon, followed by an ultrafast bootstrap analysis of
326 1,000 replicates (61). This tree (Figure 3) indicates that *O. javanicus* forms a monophyletic group
327 with *O. melastigma* but not with the *O. latipes* species complex (Hd-rR, HNI-II, and HSOK), being
328 consistent with previous trees inferred from two mitochondrial genes and a nuclear gene (14).

329

330

331

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

332 **Adaptation to salinity and hatching enzymes**

333 To gain insight into gene family evolution associated with osmoregulation, we used HMMER
334 version 3.1b2 (62) to identify Pfam domain (Pfam 32, El-Gebali et al., 2019) containing proteins
335 in the *O. javanicus* genome. We used protein sequences based on our gene model of *O. javanicus*
336 combined with Ensembl genes of the *O. latipes* species complex (Hd-rR, HNI-II and HSOK) and *O.*
337 *melastigma* (a synonym of *O. dancena*) for the Pfam search, and focused on 147 domains found in
338 224 proteins whose functions were related to osmoregulation (Additional Tables 1 and 2).
339 Similar numbers of proteins were observed among species for each domain, suggesting that the
340 osmoregulation gene repertoires are relatively conserved in *Oryzias* species. However, further
341 detailed comparisons are required because gene annotation methods are different among data.

342

343 We then also focused on specific genes encoding hatching enzymes. In the genome of *O. latipes*,
344 five copies of HCE genes -including one pseudogene- are clustered tandemly with the same
345 transcriptional direction on chromosome 3 (chr. 3), while only one single copy of the LCE gene is
346 located on chromosome 24 (chr. 24) (63). In *O. javanicus* 5 copies of the HCE (OjHCE) gene are
347 located on chromosome 3 and one LCE (OjLCE) gene was found on chromosome 24. The amino
348 acid sequence similarities in the mature enzyme region of the 5 OjHCE genes are between 89-
349 99%. Only in comparison to *O. latipes*, within the five *O. javanicus* HCE genes, the fourth one
350 (OjHCE4) displays an opposite orientation compared to the others (Figure 4A) suggesting a re-
351 arrangement within the HCE gene cluster that has likely been occurring during the evolution of
352 *Oryzias* lineage.

353

354

355 While LCE's activity remains constant over various salinities, HCEs have been reported to show
356 salt-dependent activity (49). In contrast to other *Oryzias* species, *O. javanicus,* being a euryhaline
357 species, specifically adapted its physiology to higher water salinities. In order to test whether
358 such adaptive evolution would translate at the level of HCE activity, recombinant OjHCE3
359 (rOjHCE3) was generated in an *E. coli* expression system, refolded, and its activity regarding to
360 the digestion of the egg-envelope determined at various salt concentrations based on the method
361 described in Kawaguchi et al. (49). Although rOjHCE3 showed virtually no activity at 0 M NaCl, an
362 increased activity was apparent at elevated salt concentrations. Furtheron rOjHCE3 activity was
363 recorded to be highest at 0.25 M NaCl, while still maintaining high activity up to 0.75 M NaCl
364 (Figure 4B). In contrast, it has been reported that *O. latipes* HCEs show highest activity at 0 M
365 NaCl, and drastically decrease when salt concentrations increase ((20), Figure 4B). These results
366 suggest that salt preference of HCE enzymes is a species-specific adaptation to different salt
367 environments at hatching.

368

369

370

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

371 **CONCLUSION**

372 The Java medaka, *Oryzias javanicus*, is one of the two species of medaka perfectly adapted for
373 living in brackish/sea-waters. Being an important component of the mangrove ecosystem, *O.*
374 *javanicus* is also used as a valuable marine test-fish for ecotoxicology studies. Here, we sequenced
375 and assembled the whole genome of *O. javanicus*. Complementary sequencing approaches and
376 data integration with a genetic map allowed the final assembly of the 908 Mbp of the *O. javanicus*
377 genome. The final draft assembly contains 525 scaffolds with a total length of 809.7 Mbp, a N50
378 of 6,3 Mbp and a L50 of 37 scaffolds. Providing here a high-quality draft genome assembly of the
379 euryhaline Javafish medaka, we anticipate this resource will be catalytic for a wide range of
380 comparative genomic, phylogenetic and functional studies within the genus *Oryzias* and beyond.
381
382
383 **Availability of supporting data**
384 All genome and transcriptome datasets are available at the GigaDB repository [XX]. The genome
385 assembly has also been deposited at GenBank under whole genome shotgun sequencing project
386 accession number RWID00000000.1. Illumina genome and transcriptomes and PacBio genome
387 raw reads are also available in the Sequence Read Archive (SRA), under BioProject reference
388 PRJNA505405.
389
390 **Author contributions**
391 -Designed the project: YT, AH, YG, KN
392 -Collected the samples and prepared the quality control: YT
393 -Sequencing data production: CR, OB, CD, CB, LJ
394 -Data analysis: MZ, CC, CK, MK, SY, SA, KI, CS
395 -Wrote the manuscript: YT, AH, MK, MS, SY, SA
396 -Supervision, project administration and funding acquisition: YT, AH, YG, MK, KN
397 All the authors read and approved the final manuscript.
398
399 **Competing interests**
400 All authors declare no competing interest.
401
402 **Acknowledgements**
403 This work was supported by a “Projet Incitatif PHASE department 2015" grant (Grant ID
404 ACI_PHASE, Institut National de la Recherche Agronomique) to AH, a NIBB Collaborative
405 Research Initiative to KN, NIBB individual Collaboration Research Project to MK, and Grants-in-
406 Aid for Young Scientists to YT (Grant ID 16K18590) and MK (Grant ID 16K18593). The GeT and
407 MGX core facilities were supported by France Génomique National infrastructure, funded as part
408 of “Investissement d’avenir” program managed by Agence Nationale pour la Recherche (contract

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

409 ANR-10-INBS-09). The GeT core facility was also supported by the GET-PACBIO program
410 (« Programme operationnel FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020 »).
411
412
413
414
415
416

417 Figure Legends

418 **Figure 1: A couple of Java medakas, *Oryzias javanicus*.** Picture from K. Naruse, NBRP Medaka
419 stock centre (https://shigen.nig.ac.jp/medaka/top/top.jsp).

420 **Figure 2: Oryzias javanicus assembly pipeline.** Sequencing data are represented by coloured
421 rectangles with waved bases. Tools used are in grey rectangles. Assembly metrics are in grey and
422 white rectangles. This pipeline is divided in stages symbolized by the frame.

423

424 **Figure 3: Phylogenetic position of** *O. javanicus*. Maximum likelihood tree was inferred from the
425 concatenated codon-alignment of 10,852 single-copy genes among 5 reference assemblies of
426 *Oryzias* species with Southern platyfish (*Xiphophorus maculatus*) as outgroup. All nodes were
427 supported by 100% bootstrap values.
428

429 **Figure 4. Hatching enzyme of Oryzias javanicus.** (A) HCE gene cluster of *O. latipes* (MHCE1-5)
430 and *O. javanicus* (OjHCE1-5). Arrowheads indicate direction of transcription. (B) Salt dependency
431 of *O. javanicus* HCE (black circle) and *O. latipes* (white circle). Activities are shown as % of
432 relative activity with respect to highest activity, which is considered as 100% in each species.
433
434
435

436 References

437 1. Parenti,L.R. (2008) A phylogenetic analysis and taxonomic revision of ricefishes,
438 Oryzias and relatives (Beloniformes, Adrianichthyidae). *Zool J Linn Soc*, **154**,
439 494–610.

440 2. Inoue,K. and Takei,Y. (2002) Diverse adaptability in oryzias species to high
441 environmental salinity. *Zool. Sci.*, **19**, 727–734.

442 3. Inoue,K. and Takei,Y. (2003) Asian medaka fishes offer new models for studying
443 mechanisms of seawater adaptation. *Comp. Biochem. Physiol. B, Biochem.*
444 *Mol. Biol.*, **136**, 635–645.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

445 4. Mokodongan,D.F. and Yamahira,K. (2015) Origin and intra-island diversification
446 of Sulawesi endemic Adrianichthyidae. *Mol. Phylogenet. Evol.*, **93**, 150–160.

447 5. Wittbrodt,J., Shima,A. and Schartl,M. (2002) Medaka--a model organism from the
448 far East. *Nat. Rev. Genet.*, **3**, 53–64.

449 6. Kirchmaier,S., Naruse,K., Wittbrodt,J. and Loosli,F. (2015) The genomic and
450 genetic toolbox of the teleost medaka (Oryzias latipes). *Genetics*, **199**, 905–
451 918.

452 7. Myosho,T., Takahashi,H., Yoshida,K., Sato,T., Hamaguchi,S., Sakamoto,T. and
453 Sakaizumi,M. (2018) Hyperosmotic tolerance of adult fish and early embryos
454 are determined by discrete, single loci in the genus Oryzias. *Sci Rep*, **8**, 6897.

455 8. Tanaka,K., Takehana,Y., Naruse,K., Hamaguchi,S. and Sakaizumi,M. (2007)
456 Evidence for different origins of sex chromosomes in closely related Oryzias
457 fishes: substitution of the master sex-determining gene. *Genetics*, **177**, 2075–
458 2081.

459 9. Takehana,Y., Naruse,K., Hamaguchi,S. and Sakaizumi,M. (2007) Evolution of
460 ZZ/ZW and XX/XY sex-determination systems in the closely related medaka
461 species, Oryzias hubbsi and O. dancena. *Chromosoma*, **116**, 463–470.

462 10. Takehana,Y., Demiyah,D., Naruse,K., Hamaguchi,S. and Sakaizumi,M. (2007)
463 Evolution of different Y chromosomes in two medaka species, Oryzias
464 dancena and O. latipes. *Genetics*, **175**, 1335–1340.

465 11. Herpin,A. and Schartl,M. (2009) Molecular mechanisms of sex determination and
466 evolution of the Y-chromosome: insights from the medakafish (Oryzias
467 latipes). *Mol. Cell. Endocrinol.*, **306**, 51–58.

468 12. Myosho,T., Otake,H., Masuyama,H., Matsuda,M., Kuroki,Y., Fujiyama,A.,
469 Naruse,K., Hamaguchi,S. and Sakaizumi,M. (2012) Tracing the emergence of
470 a novel sex-determining gene in medaka, Oryzias luzonensis. *Genetics*, **191**,
471 163–170.

472 13. Takehana,Y., Matsuda,M., Myosho,T., Suster,M.L., Kawakami,K., Shin-I,T.,
473 Kohara,Y., Kuroki,Y., Toyoda,A., Fujiyama,A., *et al.* (2014) Co-option of
474 Sox3 as the male-determining factor on the Y chromosome in the fish Oryzias
475 dancena. *Nat Commun*, **5**, 4157.

476 14. Takehana,Y., Naruse,K. and Sakaizumi,M. (2005) Molecular phylogeny of the
477 medaka fishes genus Oryzias (Beloniformes: Adrianichthyidae) based on
478 nuclear and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.*, **36**, 417–
479 428.

480 15. Yusof,S., Ismail,A., Koito,T., Kinoshita,M. and Inoue,K. (2012) Occurrence of
481 two closely related ricefishes, Javanese medaka (Oryzias javanicus) and Indian
482 medaka (O. dancena) at sites with different salinity in Peninsular Malaysia.
483 *Environ Biol Fish*, **93**, 43–49.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

484 16. Zulkifli,S.Z., Mohamat-Yusuff,F., Ismail,A. and Miyazaki,N. (2012) Food
485 preference of the giant mudskipper Periophthalmodon schlosseri (Teleostei :
486 Gobiidae). *Knowl. Managt. Aquatic Ecosyst.*, 10.1051/kmae/2012013.

487 17. Koyama,J., Kawamata,M., Imai,S., Fukunaga,M., Uno,S. and Kakuno,A. (2008)
488 Java medaka: a proposed new marine test fish for ecotoxicology. *Environ.*
489 *Toxicol.*, **23**, 487–491.

490 18. Horie,Y., Kanazawa,N., Yamagishi,T., Yonekura,K. and Tatarazako,N. (2018)
491 Ecotoxicological Test Assay Using OECD TG 212 in Marine Java Medaka
492 (Oryzias javanicus) and Freshwater Japanese Medaka (Oryzias latipes). *Bull*
493 *Environ Contam Toxicol*, **101**, 344–348.

494 19. Yasumasu,S., Kawaguchi,M., Ouchi,S., Sano,K., Murata,K., Sugiyama,H.,
495 Akema,T. and Iuchi,I. (2010) Mechanism of egg envelope digestion by
496 hatching enzymes, HCE and LCE in medaka, Oryzias latipes. *J. Biochem.*,
497 **148**, 439–448.

498 20. Kawaguchi,M., Yasumasu,S., Shimizu,A., Kudo,N., Sano,K., Iuchi,I. and
499 Nishida,M. (2013) Adaptive evolution of fish hatching enzyme: one amino
500 acid substitution results in differential salt dependency of the enzyme. *J. Exp.*
501 *Biol.*, **216**, 1609–1615.

502 21. Takehana,Y., Hamaguchi,S. and Sakaizumi,M. (2008) Different origins of ZZ/ZW
503 sex chromosomes in closely related medaka fishes, Oryzias javanicus and O.
504 hubbsi. *Chromosome Res.*, **16**, 801–811.

505 22. Baird,N.A., Etter,P.D., Atwood,T.S., Currey,M.C., Shiver,A.L., Lewis,Z.A.,
506 Selker,E.U., Cresko,W.A. and Johnson,E.A. (2008) Rapid SNP discovery and
507 genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

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

510 24. Vurture,G.W., Sedlazeck,F.J., Nattestad,M., Underwood,C.J., Fang,H.,
511 Gurtowski,J. and Schatz,M.C. (2017) GenomeScope: fast reference-free
512 genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.

513 25. Animal Genome Size Database:: Home.

514 26. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M.
515 (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer
516 weighting and repeat separation. *Genome Res.*, **27**, 722–736.

517 27. Ruan,J. (2018) Ultra-fast de novo assembler using long noisy reads.
518 **ruanjue/smartdenovo**.

519 28. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S.,
520 Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K., *et al.* (2014) Pilon: an
521 integrated tool for comprehensive microbial variant detection and genome
522 assembly improvement. *PLoS ONE*, **9**, e112963.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

523 29. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with
524 BWA-MEM. **arXiv:1303.3997v2 [q**-bio.**GN]**.

525 30. Yeo,S., Coombe,L., Chu,J., Warren,R.. and Birol,I. (2017) ARCS: Assembly
526 Roundup by Chromium Scaffolding. **BioRxiv 100750**.

527 31. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M.
528 (2015) BUSCO: assessing genome assembly and annotation completeness
529 with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

530 32. Babraham Bioinformatics - Trim Galore!

531 33. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput
532 sequencing reads. *EMBnet.journal*, **17**, 10–12.

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

536 35. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A.,
537 Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., *et al.* (2011) The variant
538 call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

539 36. Rastas,P. (2017) Lep-MAP3: robust linkage mapping even for low-coverage
540 whole genome sequencing data. *Bioinformatics*, **33**, 3726–3732.

541 37. Kent,W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, **12**, 656–
542 664.

543 38. Tang,H., Zhang,X., Miao,C., Zhang,J., Ming,R., Schnable,J.C., Schnable,P.S.,
544 Lyons,E. and Lu,J. (2015) ALLMAPS: robust scaffold ordering based on
545 multiple maps. *Genome Biol.*, **16**, 3.

546 39. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput
547 Sequence Data.

548 40. Cabau,C., Escudié,F., Djari,A., Guiguen,Y., Bobe,J. and Klopp,C. (2017)
549 Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies.
550 *PeerJ*, **5**, e2988.

551 41. Schulz,M.H., Zerbino,D.R., Vingron,M. and Birney,E. (2012) Oases: robust de
552 novo RNA-seq assembly across the dynamic range of expression levels.
553 *Bioinformatics*, **28**, 1086–1092.

554 42. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P.,
555 Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq
556 aligner. *Bioinformatics*, **29**, 15–21.

557 43. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J.,
558 Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and
559 quantification by RNA-Seq reveals unannotated transcripts and isoform
560 switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

561 44. RepeatMasker Home Page.

562 45. Morgulis,A., Gertz,E.M., Schäffer,A.A. and Agarwala,R. (2006) A fast and
563 symmetric DUST implementation to mask low-complexity DNA sequences. *J.*
564 *Comput. Biol.*, **13**, 1028–1040.

565 46. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences.
566 *Nucleic Acids Res.*, **27**, 573–580.

567 47. Smit,A.F.A. and Hubley,R. (2010) RepeatModeler Open-1.0.

568 48. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for
569 comparing genomic features. *Bioinformatics*, **26**, 841–842.

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

573 50. Slater,G.S.C. and Birney,E. (2005) Automated generation of heuristics for
574 biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

575 51. Hoff,K.J., Lange,S., Lomsadze,A., Borodovsky,M. and Stanke,M. (2016)
576 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with
577 GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.

578 52. Eilbeck,K., Moore,B., Holt,C. and Yandell,M. (2009) Quantitative measures for
579 the management and comparison of annotated genomes. *BMC Bioinformatics*,
580 **10**, 67.

581 53. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment
582 using DIAMOND. *Nat. Methods*, **12**, 59–60.

583 54. Setiamarga,D.H.E., Miya,M., Yamanoue,Y., Azuma,Y., Inoue,J.G., Ishiguro,N.B.,
584 Mabuchi,K. and Nishida,M. (2009) Divergence time of the two regional
585 medaka populations in Japan as a new time scale for comparative genomics of
586 vertebrates. *Biol. Lett.*, **5**, 812–816.

587 55. Bernt,M., Donath,A., Jühling,F., Externbrink,F., Florentz,C., Fritzsch,G., Pütz,J.,
588 Middendorf,M. and Stadler,P.F. (2013) MITOS: improved de novo metazoan
589 mitochondrial genome annotation. *Mol. Phylogenet. Evol.*, **69**, 313–319.

590 56. Emms,D.M. and Kelly,S. (2015) OrthoFinder: solving fundamental biases in
591 whole genome comparisons dramatically improves orthogroup inference
592 accuracy. *Genome Biol.*, **16**, 157.

593 57. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: robust conversion of
594 protein sequence alignments into the corresponding codon alignments. *Nucleic*
595 *Acids Res.*, **34**, W609-612.

596 58. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool
597 for automated alignment trimming in large-scale phylogenetic analyses.
598 *Bioinformatics*, **25**, 1972–1973.

bioRxiv preprint doi: https://doi.org/10.1101/593673. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.

599 59. Borowiec,M.L. (2016) AMAS: a fast tool for alignment manipulation and
600 computing of summary statistics. *PeerJ*, **4**, e1660.

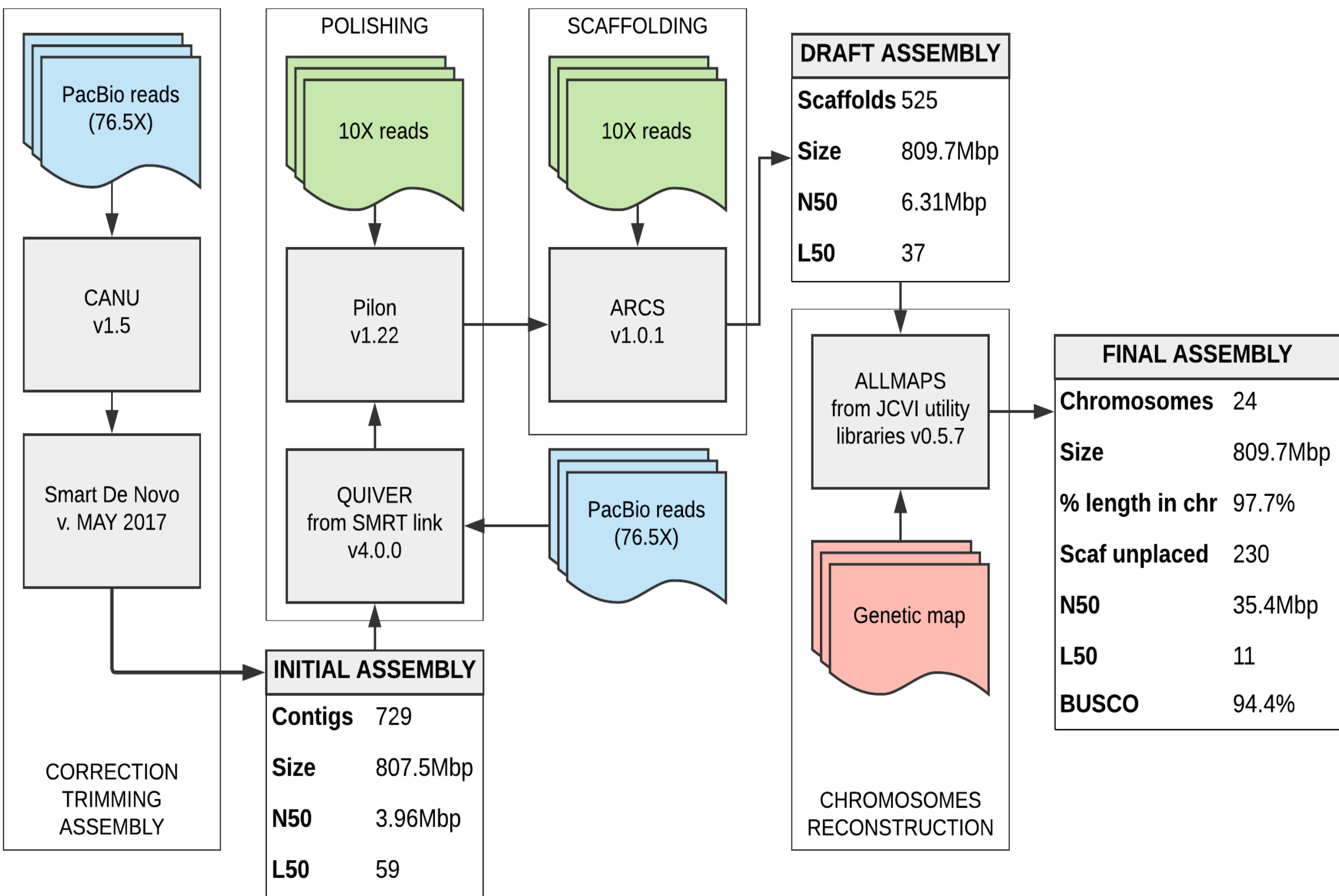
601 60. Nguyen,L.-T., Schmidt,H.A., von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a
602 fast and effective stochastic algorithm for estimating maximum-likelihood
603 phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

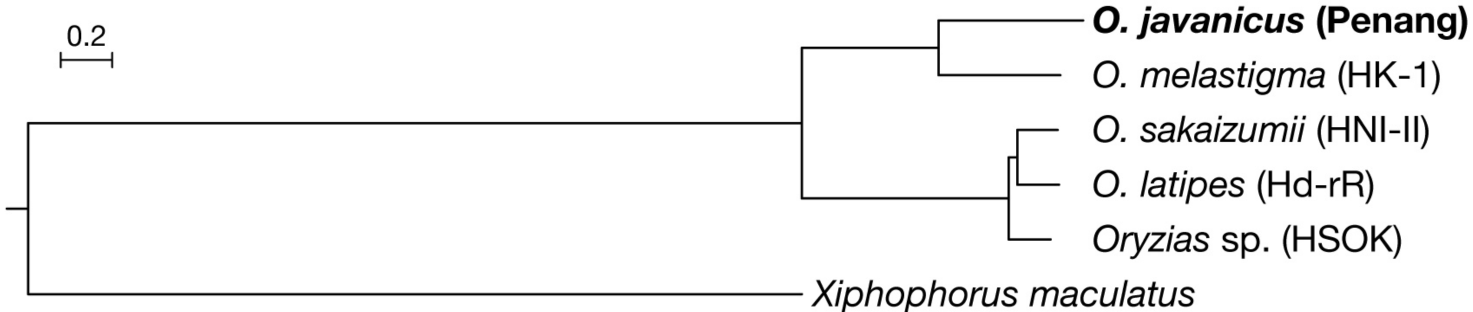
604 61. Hoang,D.T., Chernomor,O., von Haeseler,A., Minh,B.Q. and Vinh,L.S. (2018)
605 UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*,
606 **35**, 518–522.

607 62. HMMER.

608 63. Kawaguchi,M., Yasumasu,S., Hiroi,J., Naruse,K., Suzuki,T. and Iuchi,I. (2007)
609 Analysis of the exon-intron structures of fish, amphibian, bird and mammalian
610 hatching enzyme genes, with special reference to the intron loss evolution of
611 hatching enzyme genes in Teleostei. *Gene*, **392**, 77–88.

612


POLISHING
SCAFFOLDING
DRAFT ASSEMBLY
Scaffolds 525
Size 809.7Mbp
N50 6.31Mbp
L50 37
PacBio reads
(76.5X)
10X reads
10X reads
CANU
v1.5
Pilon
v1.22
ARCS
v1.0.1
ALLMAPS
from JCVI utility
libraries v0.5.7
FINAL ASSEMBLY
Chromosomes 24
Size 809.7Mbp
% length in chr 97.7%
Scaf unplaced 230
N50 35.4Mbp
L50 11
BUSCO 94.4%
Smart De Novo
v. MAY 2017
QUIVER
from SMRT link
v4.0.0
PacBio reads
(76.5X)
Genetic map
INITIAL ASSEMBLY
Contigs 729
Size 807.5Mbp
N50 3.96Mbp
L50 59
CORRECTION
TRIMMING
ASSEMBLY
CHROMOSOMES
RECONSTRUCTION


0.2
O. javanicus (Penang)
O. melastigma (HK-1)
O. sakaizumii (HNI-II)
O. latipes (Hd-rR)
Oryzias sp. (HSOK)
Xiphophorus maculatus

(A)

MHCE1
MHCE2
MHCE3
MHCE4
ψMHCE5

OjHCE1
OjHCE2
OjHCE3
OjHCE4
OjHCE5

10 kbp

(B)

activity (%)
100
50
0
0
0.25
0.5
0.75
NaCl (M)

rOjHCE3
rMHCE3