Bicolor angelfish (*Centropyge bicolor*) provides the first chromosome-level genome of the Pomacanthidae family

3

4 Chunhua Li (ORCID: 0000-0002-9947-6855)^{1, †}, Xianwei Yang (ORCID:
5 0000-0003-4388-9674)^{1,3, †}, Libin Shao (ORCID: 0000-0002-1789-6005)^{1, †}, Rui
6 Zhang (ORCID: 0000-0003-1921-9435)¹, Qun Liu¹(ORCID:0000-0002-5772-4929)¹,
7 Mengqi Zhang(ORCID: 0000-0002-5641-0557)¹, Shanshan
8 Liu(ORCID:0000-0002-5756-1728)¹, Shanshan Pan (ORCID:000-0001-5977-4914)
9 ¹, Weizhen Xue¹(ORCID: 0000-0003-4254-2476)¹, Congyan Wang¹(ORCID:
10 0000-0002-0734-5461)¹, Chunyan Mao¹(ORCID: 0000-0001-5394-2884)¹, He Zhang
11 (ORCID: 0000-0001-9294-1403)^{1,2,*}, Guangyi Fan (ORCID: 0000-0001-7365-1590)¹,

12 *

13

14 ¹BGI-Qingdao, BGI-Shenzhen, Qingdao 26655-5, China

15 ²Department of Biology, Hong Kong Baptist University, Hong Kong, China

16 ³College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049,
17 China

18

19 *fanguangyi@genomics.cn; zhanghe@genomics.cn

20 †Contributed equally

21

22 Abstract

23 The Bicolor Angelfish, *Centropyge bicolor*, is a tropical coral reef fish. It is named for
 24 its striking two-color body. However, a lack of high-quality genomic data means little is
 25 known about the genome of this species. Here, we present a chromosome-level *C.*
 26 *bicolor* genome constructed using Hi-C data. The assembled genome is 650 Mbp in
 27 size, with a scaffold N50 value of 4.4 Mbp, and a contig N50 value of 114 Kbp.
 28 Protein-coding genes numbering 21,774 were annotated. Our analysis will help others
 29 to choose the most appropriate *de novo* genome sequencing strategy based on resources
 30 and target applications. To the best of our knowledge, this is the first chromosome-level
 31 genome for the Pomacanthidae family, which might contribute to further studies
 32 exploring coral reef fish evolution, diversity and conservation.

33

34 Data Description

35 Background

36 *Centropyge bicolor* (NCBI:txid109723; FishbaseID: 5454;
 37 urn:lsid:marinespecies.org:taxname:211780) (Figure 1), also known as the Bicolor,
 38 Two-Colored, or Pacific Rock Beauty Angelfish, is a showy coral reef fish commonly

distributed in the Indo-Pacific ocean (from East Africa to the Samoan and Phoenix Islands, north to southern Japan, south to New Caledonia; throughout Micronesia). As a member of the Pomacanthidae family, it is similar to those of the Chaetodontidae (Butterflyfishes) but is distinguished by the presence of strong preopercle spines. *C. bicolor* has clear boundaries between its body colors, so might be a good model in which to study body color development in coral fish^[1].

Context

Although the availability of genetic, and especially genomic resources, remains limited for the Pomacanthidae family, we assembled the first *C. bicolor* reference genome. This will provide valuable information for genetic studies of this coral reef fish, and will contribute to studies in body color diversity. With the whole genome sequence of *C. bicolor*, it might be possible to explore the genetic mechanisms of body color development in coral reef fish by comparative genomic methods.

Methods and Results

A protocols collection for BGISEQ-500, stLFR and Hi-C library construction is available in protocols.io (Figure 2)^[2].

57

58 **Figure 2.** Protocols for BGISEQ-500, stLRF and Hi-C library preparation and
59 construction, and genome assembly, for the Bicolor Angelfish, *Centropyge bicolor*^[2].

60

61 Sample collection and genome sequencing

62 A *C. bicolor* individual was collected from the market in Xiamen, Fujian Province,
63 China. DNA was extracted from fresh muscle tissue according to a standard protocol.
64 Single-tube long fragment read (stLFR)^[2] and Hi-C libraries were constructed
65 following the manufacturers' instructions^[2,3] to sequence and assemble the genome.
66 We obtained 130.47 Gbp (gigabase pairs; ~197×) raw stLFR data and 134.57 Gbp
67 (~203.20×) raw Hi-C data (Table 1) using the BGISEQ-500 platform in 100-bp
68 (basepair) paired-end mode.

69 Low-quality reads (sequences with more than 40% of bases with a quality score
70 lower than 8), polymerase chain reaction (PCR) duplications, adaptor sequences and
71 reads with a high (greater than 10%) proportion of ambiguous bases (Ns) occurring in
72 stLFR data were filtered using SOAPnuke (v1.6.5; RRID:SCR_015025)^[4]. We obtained

62.6 Gbp (~91.67×) clean data (Table 1) to assemble the draft genome. Meanwhile, HiC-Pro (v. 2.8.0)^[5] was used for the quality control of raw Hi-C data, and 42.51 Gbp (~ 64.19×) valid data were used to assemble the genome to the chromosome-level (Table 1).

Table 1. Statistics of DNA sequencing data.

Libraries	Read length	Raw data			Valid data		
		Total	bases	Sequencing	Total	bases	Sequencing depth
		(Gbp)		depth (×)	(Gbp)		(×)
stLFR	100:100	130.47		197.00	60.71		91.67
Hi-C	100:100	134.57		203.20	42.51		64.19

Sequencing depth = Total bases / Genome size, where the genome size is the result of *k*-mer estimation, as shown in Table 2.

Genome assembly

Using GenomeScope software with stLFR clean data, *k*-mer distribution was used to understand the genome complexity before genome assembly^[6]. The genome size of *C. bicolor* was estimated as 662.27 Mbp (megabase pairs), with 37.6% repeat sequences

86 and 1.16% heterozygous sites (Table 2, Figure 3).

87

88 **Table 2.** Statistical information of 17-mer analysis.

<i>k</i> -mer	<i>k</i> -mer number	<i>k</i> -mer Depth	Heterozygosity (%)	Genome size (Mbp)
17	50,994,645,240	77	1.16	662.27

89 The genome size, G, was defined as $G = K_num / K_depth$, where K_num is the total number of *k*-mers, and

90 K_depth is the most frequently occurring *k*-mer.

91

92 We reformatted the clean stLFR data into 10× Genomics format using an in-house
 93 script(https://github.com/BGI-Qingdao/stlfr2supernova_pipeline) and assembled the
 94 draft genome using Supernova (v.2.0.1 , RRID:SCR_016756)^[7] with default
 95 parameters. The draft genome was 681 Mbp, with a contig N50 of 115.5 Kbp (kilobase
 96 pairs) and scaffold N50 of 4.4 Mbp (Table 3), which is similar to the estimated genome
 97 size.

98

99 **Table 3.** Statistics of the draft assembly with stLFR data.

Statistics	Contig	Scaffold
------------	--------	----------

Total number (#)	40,442	29,065
Total length (bp)	655,705,062	681,285,455
Gap (N) (bp)	0	25,580,393
Average length (bp)	16,213.47	23,440.06
N50 length (bp)	115,524	4,424,004
N90 length (bp)	6,029	7,618
Maximum length (bp)	1,148,507	21,943,074
Minimum length (bp)	48	940
GC content (%)	41.74	41.74

100

101 To obtain the chromosome-level genome, we used Juicer (v3, RRID:SCR_017226)^[8]
102 to build a contact matrix and 3dDNA(v. 170123)^[9]to sort and anchor scaffolds with
103 the parameters: “-m haploid -s 4 -c 24”. There are 24 distinct contact blocks, which
104 correspond to 24 chromosomes, representing 96% of the whole genome (Figure 4A,
105 Figure 5, Table 4). On evaluating the completeness of the genome and gene set using
106 Benchmarking Universal Single-Copy Orthologs (BUSCO,(v.3.0.2 ,
107 RRID:SCR_015008))^[10]and a vertebrata database, our assembly maintained a score
108 of 96.2% (Table 5). We also identified putative homologous chromosomal regions

109 between *C. bicolor* and *Oryzias latipes* by MCscanx^[11](Figure 6).

110

111 **Table 4.** Statistics of the chromosome-level genome.

Statistics	Contig	Scaffold
Total number (#)	40,778	28,555
Total length (bp)	655,705,062	680,873,932
Gap (N) (bp)	0	25,168,870
Average length (bp)	16,079.87	23,844.30
N50 length (bp)	113,563	21,943,074
N90 length (bp)	5,988	7,542
Maximum length (bp)	1,148,507	28,105,280
Minimum length (bp)	43	43
GC content (%)	41.74	41.74

112

113 **Table 5.** Statistics of the BUSCO assessment.

Types of BUSCOs	Gene set	Assembly
-----------------	----------	----------

	Number	Percentage	Number	Percentage
		(%)		(%)
Complete BUSCOs	2,408	93.1	2,486	96.2
Complete BUSCOs	2,348	90.8	2,438	94.3
Fragmented BUSCOs	81	3.1	64	2.5
Missing BUSCOs	97	3.8	36	1.3
Total BUSCO groups searched	2,586	100	2,586	100

114

115 In addition, we cut off partial stLFR reads (25 M) for assembly by MitoZ with
 116 default parameters^[12], and obtained a 16,961-bp circular mitochondrial genome of *C.*
 117 *bicolor*. Thirteen protein-coding genes, 24 tRNA genes and three rRNA genes were
 118 annotated by GeSeq^[13] (Figure 4B).

119

120 Genomic annotation

121 For the annotation of repeats, we carried out homolog annotation and *ab initio*
 122 prediction independently. RepeatMasker (v.4.0.6 , RRID:SCR_012954)^[14],
 123 RepeatProteinMask (a module from RepeatMasker) and trf (Tandem Repeats Finder,

v.4.07b)^[15] were used to identify known repetitive sequences by comparing the whole genome with RepBase^[16]. LTR_FINDER (v.1.06, RRID:SCR_015247)^[17][15] and RepeatModeler (v.1.0.8, RRID:SCR_015027)^[18] were used in *de novo* prediction. We also classified transposable elements (TEs) from the integration of all repeats. In total, we identified 124 Mbp (18.32% of the entire genome) of repetitive sequences (Figure 4A, Table 6), including 110 Mbp of TEs (Figure 4A, Table 7).

130

131 **Table 6.** Statistics of repetitive sequences.

Type	Repeat size (bp)	Percentage of genome (%)
TRF	14,165,095	2.08
RepeatMasker	43,423,877	6.38
RepeatProteinMask	12,503,750	1.84
<i>De novo</i>	110,871,693	16.28
Total	124,708,977	18.32

132

133 **Table 7.** Statistics of transposable elements.

Repbases	TEs, n	Protein	TEs, n	<i>De novo</i>	TEs, n	Combined	TEs, n
----------	--------	---------	--------	----------------	--------	----------	--------

	(%)	(%)	(%)	(%)
DNA	27,163,851 (3.990)	1,068,990 (0.157)	61,731,447 (9.067)	70,925,963 (10.417)
LINE	10,228,332 (1.502)	6,956,340 (1.022)	20,006,579 (2.938)	26,714,285 (3.924)
SINE	856,125 (0.126)	0 (0.000)	497,024 (0.073)	1,187,676 (0.174)
LTR	10,971,817 (1.611)	4,485,808 (0.659)	16,270,071 (2.390)	23,101,529 (3.393)
Other	10,041 (0.001)	0	0	10,041 (0.001)
Unknown	0	0	14,054,230 (2.064)	14,054,230 (2.064)
Total	43,423,877 (6.378)	12,503,750 (1.836)	99,265,690 (14.579)	109,868,166 (16.136)

134

135 Homolog-based and *ab initio* prediction were used to identify the protein-coding
136 genes. Augustus (v.3.3, RRID:SCR_008417)^[19] was used in *ab initio* prediction basing
137 on a repeat-masked genome^[20]. Protein sequences of *Astatotilapia calliptera*, *Danio*
138 *rerio*, *Larimichthys crocea*, and *Oreochromis niloticus* were downloaded from the
139 National Center for Biotechnology Information (NCBI) GenBank database and aligned

140 to the *C. bicolor* genome for homolog gene annotation with Genewise (v2.4.1,
141 RRID:SCR_015054)^[21]. Finally, we used GLEAN^[22] to integrate all the above
142 evidence and obtained a total of 21,774 genes, which contained 11 exons on average
143 and had an average coding sequence (CDS) length of 1,575 bp (Table 8).

144

145 **Table 8.** Statistics of the predicted genes in the bicolor angelfish genome.

	Gene set	Gene	Average transcript length	Average CDS length	Average intron length	Average exon length	Average exons per gene
		number	(bp)	(bp)	(bp)	(bp)	
Homolog	<i>Astatotilapia calliptera</i>	51,174	21,762.29	2,259.23	1,691.33	180.29	12.53
	<i>Danio rerio</i>	22,005	27,982.75	1,570.36	3,438.82	180.90	8.68
	<i>Larimichthys crocea</i>	47,419	19,884.78	2,139.39	1,575.94	174.50	12.26
	<i>Oreochromis niloticus</i>	47,067	17,771.04	1,906.97	1,608.29	175.53	10.86
	<i>De novo</i>	Augustus	34,470	9,675.42	1,335.20	1,344.81	185.40

GLEAN	21,774	14,024.40	1,906.28	1,206.07	172.55	11.05
--------------	--------	-----------	----------	----------	--------	-------

146 The GLEAN gene set is the integrated result of *de novo* gene predictions and homolog gene predictions.

147

148 To predict gene functions, 21,774 genes were aligned against several public

149 databases, including TrEMBL^[23], SwissProt^[23], KEGGViewer^[24] and InterProScan^[25].

150 As a result, 99.67% of all genes were predicted functionally (Table 9, Figure 7).

151 **Table 9.** Statistics of the functional annotation.

Database	Number	Percentage (%)
Total	21,774	100.00
SwissProt	20,784	95.45
KEGG	19,168	88.03
TrEMBL	21,688	99.61
Interpro	20,153	92.56
Overall	21,702	99.67

152

153 **Phylogenetic analysis**

154 We downloaded the gene data of seven representative teleost fishes from NCBI to

study the phylogenetic relationships between *C. bicolor*. These seven fishes were: *Danio rerio*, *Gasterosteus aculeatus*, *Gadus morhua*, *Larimichthys crocea*, *Oryzias latipes*, *Oreochromis niloticus* and *Tetraodon nigroviridis*. For each dataset, the longest transcripts were selected and aligned to each other by BLASTP (v2.9.0, RRID:SCR_001010)^[26] (E-value $\leq 1e-5$). TreeFam (v.2.0.9, RRID:SCR_013401)^[27] was used to cluster gene families, with default parameters. Among all 20,706 clustered gene families, there were 4,450 common single-copy families and 57 families specific to *C. bicolor* (Table 10). With single-copy sequences, we used PhyML (v.3.3,RRID:SCR_014629)^[28] to construct the phylogenetic tree of *C. bicolor* and the seven other fishes mentioned above, setting *D. rerio* as an outgroup.

165

166 **Table 10. Statistics of gene family clustering.**

Species	Total genes	Unclustered genes	Families	Unique families	Average number of genes per family
<i>Centropyge</i>	21,774	694	16,219	57	1.3
<i>bicolor</i>					
<i>Danio rerio</i>	30,067	2,188	18,575	726	1.5
<i>Gasterosteus</i>	20,756	784	15,921	16	1.25
<i>aculeatus</i>					
<i>Gadus morhua</i>	19,987	535	15,630	9	1.24
<i>Larimichthys</i>	24,403	610	17,273	55	1.38
<i>crocea</i>					

<i>Oryzias latipes</i>	19,535	1,048	14,805	87	1.25
<i>Oreochromis niloticus</i>	21,431	180	15,780	14	1.35
<i>Tetraodon nigroviridis</i>	19,544	901	14,803	57	1.26

167

168 Based on the phylogenetic tree and single-copy sequences, the divergence time
 169 between different species was estimated by MCMCTREE with parameters of “--model
 170 0 --rootage 500 -clock 3”. The results showed that *C. bicolor* was formed
 171 ~34.95 million years ago, when differentiated from the common ancestor with *L. crocea*
 172 (Figure 8).

173

174 Analysis of bicolor formation in teleosts

175 Current studies suggest that different pigment cells produce different pigments. Some
 176 types of pigment cells already have been identified in teleost^[29]. *C. bicolor* has an
 177 attractive body color with clear color boundaries, but the molecular mechanism
 178 underlying this remains unknown. Compared with other teleost, there are 1,081
 179 expanded gene families and 57 specific gene families in *C. bicolor* (Figure 9).
 180 Functional enrichment analysis showed that notable expansion occurred in those gene
 181 families related to visual development and enzyme metabolism (Figure 9).

182

183 **Re-use Potential**

184 Coral reef fishes, with distinctive color patterns and color morphs, are important for
185 understanding the adaptive evolution of fishes. In this study, we firstly assembled a
186 high-quality, chromosome-level genome of *C. bicolor*, with a length of 681 Mbp, and
187 annotated 21,774 genes. This is the first genome of a fish from the Pomacanthidae
188 family. These genomic data will be useful for genome-scale comparisons and further
189 studies on the mechanisms underlying colorful body development and adaptation.

190

191 **Data Availability**

192 The data sets supporting the results of this article are available in the GigaScience
193 Database, doi: 10.5524/100802. Raw reads from genome sequencing and assembly
194 are deposited at the China National Gene Bank under reference number CNP0001160,
195 which contains sample information (CNS0315939), Hi-C raw data (CNX0286336)
196 and stLFR raw data (CNX0286337). The project also has been deposited at NCBI
197 under accession ID PRJNA702283.

198

199 **Declarations**

200 **List of Abbreviations**

201 bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; Gbp:

gigabase pair; Kbp: kilobase pair; KEGG: Kyoto Encyclopedia of Genes and Genomes;
Mbp: megabase pair; NCBI: National Center for Biotechnology Information; stLFR:
single-tube long fragment reads; TE: transposable element.

205

206 **Ethical Approval**

All resources used in this study were approved by the Institutional Review Board of
BGI (IRB approval No. FT17007). This experiment has passed the ethics audit of the
Beijing Genomics Institute (BGI) Gene Bioethics and Biosecurity Review Committee.

210

211 **Consent for Publication**

Not applicable.

213

214 **Competing Interests**

The authors declare that they have no competing interests.

216

217 **Funding**

This work was supported by funding from the “Blue Granary” project for scientific
and technological innovation of China (2018YFD0900301-05).

220

221 **Authors' Contributions**

222 H.Z. and G.F. designed this project. M.Z. prepared the samples. S.L., S.P., W.X., C.W.
223 and C.M. conducted the experiments. C.L., X.Y., L.S., R.Z. and Q.L. did the analyses.
224 C.L., X.Y., L.S., R.Z. wrote and revised the manuscript. All authors read and
225 approved the final version of the manuscript.

226

227 **Acknowledgements**

228 We thank the China National Genebank for technical support in constructing and
229 sequencing the stLFR library.

230

231 **References**

- 232 1. Mendonça RC, Chen JY, Zeng C, Tsuzuki MY. Embryonic and early larval
233 development of two marine angelfish, *Centropyge bicolor* and *Centropyge bispinosa*.
234 *Zygote*. 2020; doi: 10.1017/S0967199419000789.
- 235 2. Li C, Yang X, Shao L, Zhang R, Zhang M, Liu S, et al.. Protocols for " Bicolor
236 Angelfish (*Centropyge bicolor*) genome provided first chromosome-level reference
237 of Pomacanthidae family and clues for bi-color body formation ". :10–12020;

- 238 3. Wang O, Chin R, Cheng X, Yan Wu MK, Mao Q, Tang J, et al.. Efficient and
239 unique cobarcoding of second-generation sequencing reads from long DNA
240 molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo
241 assembly. *Genome Res.* 2019; doi: 10.1101/gr.245126.118.
- 242 4. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al.. SOAPnuke: A MapReduce
243 acceleration-supported software for integrated quality control and preprocessing of
244 high-throughput sequencing data. *Gigascience.* 2018; doi:
245 10.1093/gigascience/gix120.
- 246 5. Chen C-J, Servant N, Heard E, Lajoie BR, Viara E, Varoquaux N, et al.. HiC-Pro:
247 an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 2015; doi:
248 10.1186/s13059-015-0831-x.
- 249 6. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et
250 al.. GenomeScope: Fast reference-free genome profiling from short reads.
251 *Bioinformatics.* 2017; doi: 10.1093/bioinformatics/btx153.
- 252 7. Wong KHY, Levy-Sakin M, Kwok PY. De novo human genome assemblies reveal
253 spectrum of alternative haplotypes in diverse populations. *Nat Commun.* 2018; doi:
254 10.1038/s41467-018-05513-w.
- 255 8. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al..
256 Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C
257 Experiments. *Cell Syst.* 2016; doi: 10.1016/j.cels.2016.07.002.

- 258 9. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al.. De
259 novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length
260 scaffolds. *Science* (80-). 2017; doi: 10.1126/science.aal3327.
- 261 10. Waterhouse RM, Seppey M, Sim FA, Ioannidis P. BUSCO Applications from
262 Quality Assessments to Gene Prediction and Phylogenomics Letter Fast Track. 2017;
263 doi: 10.1093/molbev/msx319.
- 264 11. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al.. MCScanX: a toolkit
265 for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic
266 Acids Res.* 2012; doi: 10.1093/nar/gkr1293.
- 267 12. Meng G, Li Y, Yang C, Liu S. MitoZ: A toolkit for animal mitochondrial genome
268 assembly, annotation and visualization. *Nucleic Acids Res.* 2019; doi:
269 10.1093/nar/gkz173.
- 270 13. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al..
271 GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.*
272 2017; doi: 10.1093/nar/gkx391.
- 273 14. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements
274 in genomic sequences. *Curr Protoc Bioinforma.* 2009; doi:
275 10.1002/0471250953.bi0410s25.
- 276 15. Carrillo-Avila M, Resende EK, Marques DKS, Galetti PM. Tandem repeats finder:
277 a program to analyze DNA sequences. *Conserv Genet.* 2009; doi:

-
- 278 10.1590/S1679-62252007000200018.
- 279 16. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive
280 elements in eukaryotic genomes. *Mob DNA*. 2015; doi: 10.1186/s13100-015-0041-9.
- 281 17. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length
282 LTR retrotransposons. *Nucleic Acids Res*. 2007; doi: 10.1093/nar/gkm286.
- 283 18. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community
284 resource of transposable element families, sequence models, and genome annotations.
285 *Mob DNA*. 2021; doi: 10.1186/s13100-020-00230-y.
- 286 19. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in
287 eukaryotes with a generalized hidden Markov model that uses hints from external
288 sources. *BMC Bioinformatics*. 2006; doi: 10.1186/1471-2105-7-62.
- 289 20. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS:
290 ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006; doi:
291 10.1093/nar/gkl200.
- 292 21. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Systematic identification of
293 novel protein domain families associated with nuclear functions. *Genome Res*. 2002;
294 doi: 10.1101/gr.203201.
- 295 22. Lewis S, Searle S, Harris N, Gibson M, Iyer V, Richter J, et al.. Creating a honey
296 bee consensus gene set. *Genome Biol*. 2002; doi:

- 10.1186/gb-2002-3-12-research0082.
23. Bairoch A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 2000; doi: 10.1093/nar/28.1.45.
24. Habermann BH, Villaveces JM, Jimenez RC. KEGGViewer, a BioJS component to visualize KEGG Pathways. *F1000Research.* 2014; doi: 10.12688/f1000research.3-43.v1.
25. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al.. InterProScan 5: Genome-scale protein function classification. *Bioinformatics.* 2014; doi: 10.1093/bioinformatics/btu031.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; doi: 10.1016/S0022-2836(05)80360-2.
27. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, et al.. TreeFam: 2008 Update. *Nucleic Acids Res.* 2007; doi: 10.1093/nar/gkm1005.
28. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol.* 2010; doi: 10.1093/sysbio/syq010.
29. Kimura T, Nagao Y, Hashimoto H, Yamamoto-Shiraishi YI, Yamamoto S, Yabe T, et al.. Leucophores are similar to xanthophores in their specification and differentiation processes in medaka. *Proc Natl Acad Sci U S A.* 2014; doi:

316 10.1073/pnas.1311254111.

317

318

319 **Figure Legends**

320 **Figure 1.** Photograph of *Centropyge bicolor*.

321 **Figure 2.** Protocols for BGISEQ-500, stLRF and Hi-C library preparation and
322 construction, and genome assembly, for the Bicolor Angelfish, *Centropyge bicolor*^[2].

323 **Figure 3.** The 17-mer depth distribution of *Centropyge bicolor*.

324 The estimated genome size is 662.27 Mbp and the heterozygosity is 1.16%.

325 **Figure 4.** Annotation of the *Centropyge bicolor* genome. **(A)** Basic genomic elements
326 of the *Centropyge bicolor* genome. LTR, long terminal repeat; LINE, long
327 interspersed nuclear elements; SINE, short interspersed elements. **(B)** Physical map of
328 mitochondrial assembly.

329 **Figure 5.** Heat map of interactive intensity between chromosome sequences.

330 **Figure 6.** Homologous chromosomal regions between *Centropyge bicolor* and
331 *Oryzias latipes*.

332 **Figure 7.** Venn diagram of orthologous gene families.

333 Four teleost species (*Centropyge bicolor*, *Larimichthys crocea*, *Oreochromis niloticus*,
334 and *Danio rerio*) were used to generate the Venn diagram based on gene family
335 cluster analysis.

336 **Figure 8.** Comparative analysis of the *Centropyge bicolor* genome.

337 (A) The protein-coding genes of the eight species were clustered into 17,849 gene
338 families. Among these gene families, 4,450 were single-copy gene families. (B)
339 Phylogenetic analysis of *Centropyge bicolor* (Cbi.), *Danio rerio* (Dre.), *Gasterosteus*
340 *aculeatus* (Gac.), *Gadus morhua* (Gmo.), *Larimichthys crocea* (Lcr.), *Oryzias latipes*
341 (Ola.), *Oreochromis niloticus* (Oni.), and *Tetraodon nigroviridis* (Tni.) using
342 single-copy gene families. The species differentiation time between *Centropyge*
343 *bicolor* and *Larimichthys crocea* was ~34.95 million years.

344 **Figure 9.** Statistics of gene function enrichment (Gene Ontology) for expanded genes
345 of *Centropyge bicolor*.

346 Nodes are colored by *q*-value (adjusted *p*-value). Node size is shown according to its
347 enriched gene number.

20180609 XM 002

Nov 21, 2020

☆ Bookmark

📄 Copy / Fork

🔗 Protocols for "Bicolor Angelfish (*Centropyge bicolor*) genome provided first chromosome-level reference of Pomacanthidae family and clues for bi-color body formation" 🔗

Chunhua Li¹, Xianwei Yang¹, Libin Shao¹, Rui Zhang¹, QunLiu¹, Mengqi Zhang¹, Shanshan Liu², Shanshan Pan², Weizhen Xue¹, Congpan Wang¹, Chunyan Mao¹, He Zhang², Guangyi Fan¹

¹BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China,

²BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China, Department of Biology, Hong Kong Baptist University, Hong Kong, China

1 Works for me

🔗 Share

doi.org/10.17554/protocols.io.bpxhmpj8

👤 lichunhua 🏆

GenomeScope Profile

len:622,282,889bp uniq:62.4% het:1.16% kcov:39.6 err:0.105% dup:3.43% k:17

Number of gene families

Gene families
Expansion / Contraction

Combined Change
Across Lineages

MRCA
(17849)

million years ago

Statistics of Enrichment