Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish *Amphiprion percula*

Robert Lehmann[1], Damien J. Lightfoot[1], Celia Schunter[1], Craig T. Michell[2], Hajime Ohyanagi[3], Katsuhiko Mineta[3], Sylvain Foret[4,5], Michael L. Berumen[2], David J. Miller[4], Manuel Aranda[2], Takashi Gojobori[3], Philip L. Munday[4] and Timothy Ravasi[1,*]

[1] KAUST Environmental Epigenetic Program, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Kingdom of Saudi Arabia.

[2] Red Sea Research Center, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Kingdom of Saudi Arabia.

[3] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Kingdom of Saudi Arabia.

[4] ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland, 4811, Australia.

[5] Evolution, Ecology and Genetics, Research School of Biology, Australian National University, Canberra, Australian Capital Territory, 2601, Australia.

Keywords:

Orange Clownfish, *Amphiprion percula*, Nemo, Functional Genomics, Chromosome-Scale Assembly, Fish Genomics, Coral Reef Fish.

Running Title:

The Nemo Genome

[*]**Corresponding Author:**

Timothy Ravasi, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Kingdom of Saudi Arabia, timothy.ravasi@kaust.edu.sa

Abstract

The iconic orange clownfish, *Amphiprion percula*, is a model organism for studying the ecology and evolution of reef fishes, including patterns of population connectivity, sex change, social organization, habitat selection and adaptation to climate change. Notably, the orange clownfish is the only reef fish for which a complete larval dispersal kernel has been established and was the first fish species for which it was demonstrated that anti-predator responses of reef fishes could distinguish predators based on their olfactory cues and that this ability is impaired by ocean acidification. Despite its importance, molecular resources for this species remain scarce and until now it lacked a reference genome assembly. Here we present a *de novo* chromosome-scale assembly of the genome of the orange clownfish *Amphiprion percula*. We utilized single-molecule real-time sequencing technology from Pacific Biosciences to produce an initial polished assembly comprised of 1,414 contigs, with a contig N50 length of 1.86 Mb. Using Hi-C based chromatin contact maps, 98% of the genome assembly were placed into 24 chromosomes, resulting in a final assembly of 908.8 Mb in length with contig and scaffold N50s of 3.12 and 38.4 Mb, respectively. This makes it one of the most contiguous and complete fish genome assemblies currently available. The genome was annotated with 26,597 protein coding genes and contains 96% of the core set of conserved actinopterygian orthologs. The availability of this reference genome assembly as a community resource will further strengthen the role of the orange clownfish as a model species for research on the ecology and evolution of reef fishes.

Introduction

The orange clownfish, *Amphiprion percula*, which was immortalized in the film "Finding Nemo", is arguably the most recognized fish on Earth. It is also one of the most important species for studying the ecology and evolution of coral reef fishes. The orange clownfish is used as a model species to study patterns and processes of social organization (Buston, Bogdanowicz, Wong, & Harrison, 2007; Buston & Wong, 2014; Wong, Uppaluri, Medina, Seymour, & Buston, 2016), sex change (Buston, 2003), mutualism (Schmiege, D'Aloia, & Buston, 2017), habitat selection (Dixson et al., 2008; Elliott & Mariscal, 2001; Scott & Dixson, 2016), lifespan (Buston & García, 2007) and predator-prey interactions (Dixson, 2012; Manassa, Dixson, McCormick, & Chivers, 2013). It has been central to ground-breaking research into the scale of larval dispersal and population connectivity in marine fishes (Almany et al., 2017; Pinsky et al., 2017; Planes, Jones, & Thorrold, 2009; Salles et al., 2016) and how this influences the efficacy of marine protected areas (Berumen et al., 2012; Planes et al., 2009). It is also used to study the ecological effects of environmental disturbances in marine ecosystems (Hess, Wenger, Ainsworth, & Rummer, 2015; Wenger et al., 2014), including climate change (McLeod et al., 2013; Saenz-Agudelo, Jones, Thorrold, & Planes, 2011) and ocean acidification (Dixson, Munday, & Jones, 2010; Jarrold, Humphrey, McCormick, & Munday, 2017; Munday et al., 2009; Simpson et al., 2011). Perhaps more than any other species, the orange clownfish has become a mainstay of research into the chemical, molecular, behavioral, population, conservation and climate-change ecology of marine fishes.

The orange clownfish is one of 30 species of anemonefishes belonging to the subfamily Amphiprioninae within the family Pomacentridae (damselfishes). The two clownfishes, *A. percula* (orange clownfish or clown anemonefish) and *A. ocellaris* (false clownfish or western

clown anemonefish) form a separate clade, alongside *Premnas biaculeatus*, within the Amphiprioninae (J. Li, Chen, Kang, & Liu, 2015; Litsios, Pearman, Lanterbecq, Tolou, & Salamin, 2014; Litsios & Salamin, 2014). The two species of clownfish are easily distinguished from other anemonefishes by their bright orange body coloration and three vertical white bars. The orange clownfish and the false clownfish have similar body coloration, but largely distinct allopatric geographical distributions (Litsios & Salamin, 2014). The orange clownfish occurs in northern Australia, including the Great Barrier Reef (GBR), and in Papua New Guinea, Solomon Islands and Vanuatu, while the false clownfish occurs in the Indo-Malaysian region, from the Ryukyu Islands of Japan, throughout south-east Asia and south to north-western Australia (but not the GBR).

Like all anemonefishes, the orange clownfish has a mutualistic relationship with sea-anemones. Wild adults and juveniles live exclusively in association with a sea anemone, where they gain shelter from predators and benefit from food captured by the anemone (Fautin, 1991; Fautin & Allen, 1997; Mebs, 2009). In return, the sea-anemone benefits by gaining protection from predators (Fautin & Allen, 1997; Holbrook & Schmitt, 2005), from supplemental nutrition from the clownfish's waste (Holbrook & Schmitt, 2005) and from increased gas exchange as a result of increased water flow provided by clownfish movement and activity (Herbert, Bröhl, Springer, & Kunzmann, 2017; Szczebak, Henry, Al-Horani, & Chadwick, 2013). The orange clownfish associates with two species of anemone, *Stichodactyla gigantea* and *Heteractis magnifica* (Fautin & Allen, 1997). Clownfish social groups typically consist of an adult breeding pair and a variable number of smaller, size-ranked juveniles that queue for breeding rights (Buston, 2003). The breeding female is larger than the male. If the female disappears, the male changes sex to female and the largest non-

103 breeder matures into a breeding male. The breeding pair lays clutches of demersal eggs in
104 close proximity to their host anemone. Eggs hatch after 7-8 days and the larvae disperse into
105 the open ocean for a period of 11-12 days, at which time they return to the reef and settle to
106 an anemone.
107
108 The close association of clownfish and other anemonefishes with sea anemones makes them
109 excellent species for studying aspects of marine mutualisms and habitat selection. The easily
110 identified and delineated habitat they occupy, along with the ease with which the fish can be
111 observed in nature, makes them ideal candidates for behavioral and population ecology. The
112 unique capacity to collect juveniles immediately after they have settled to the reef from their
113 pelagic larval phase also makes them ideally suited to testing long-standing questions about
114 larval dispersal and population connectivity in reef fish populations. Using molecular
115 techniques to assign parentage between newly settled juveniles and adult anemonefishes,
116 recent studies have been able to describe for the first time the spatial scales of dispersal in reef
117 fish and its temporal consistency (Almany et al., 2017). The ability to map the connectivity of
118 clownfish populations in space and time has also opened the door to addressing challenging
119 questions about selection, fitness and adaptation in natural populations of marine fishes
120 (Pinsky et al., 2017; Salles et al., 2016). Finally, the orange clownfish is one of the relatively
121 few coral reef fishes that can easily be reared in captivity (Wittenrich, Turingan, & Creswell,
122 2007). Consequently, it has unrivalled potential for experimental manipulation to test
123 ecological and evolutionary questions in marine ecology (Dixson et al., 2014; Manassa et al.,
124 2013), including the impacts of climate change and ocean acidification (Nilsson et al., 2012).
125 Increasingly, genome-wide methods are being used to test ecological and evolutionary

126 questions and this is particularly true for coral reef species in the wake of anthropomorphic
127 climate change and its effects on these sensitive ecosystems (Stillman & Armstrong, 2015)
128
129 To date, genome assemblies of two anemonefish, *A. frenatus* (Marcionetti, Rossier, Bertrand,
130 Litsios, & Salamin, 2018) and *A. ocellaris* (Marcionetti et al., 2018), have been published.
131 Both of these were based on short-read Illumina technology with genome scaffolding
132 provided by shallow coverage of PacBio (Marcionetti et al., 2018) or Oxford Nanopore (Tan
133 et al., 2018) long reads. While the use of long reads to scaffold Illumina-based assemblies
134 improves contiguity, both genome assemblies are highly fragmented with respective contig
135 and scaffold N50s of 14.9 and 244.5 kb for *A. frenatus* and 323.6 and 401.7 kb for *A.*
136 *ocellaris*. Here we present a chromosome-scale genome assembly of the orange clownfish,
137 which was assembled using a primary PacBio long read strategy, followed by scaffolding
138 with Hi-C-based chromatin contact maps. The resulting final assembly is highly contiguous
139 with contig and scaffold N50 values of 3.12 and 38.4 Mb, respectively. This assembly will be
140 a valuable resource for the research community and will further establish the orange
141 clownfish as a model organism for genetic and genomic studies into ecological, evolutionary
142 and environmental aspects of reef fishes. To facilitate the use of this resource, we have
143 developed an integrated database, the Nemo Genome DB (http://nemogenome.org/), which
144 allows for the interrogation and mining of genomic and transcriptomic data described here.
145
146
147
148
149 **Materials and Methods**

150 Specimen collection and DNA extraction

151 Adult orange clownfish breeding pairs were collected on the northern GBR in Australia. Fish
152 were bred at the Experimental Aquarium Facility of James Cook University (JCU) and one
153 individual offspring was sacrificed at the age of 8 months. The whole brain was excised, snap
154 frozen and kept at -80°C until processing. High molecular weight DNA was extracted from
155 whole brain tissue using the Qiagen Genomic-tip 100/G extraction kit. The tissue was first
156 homogenized in lysis buffer G2 supplemented with 200 µg/mL RNase A using sterile beads
157 for 30 sec. After homogenization, proteinase K was added and the homogenate was incubated
158 at 50°C overnight. DNA extraction was then performed according to the manufacturer's
159 protocol with a final elution volume of 200 µl. DNA fragment size and quality was assessed
160 using pulsed-field gel electrophoresis. This study was completed under JCU animal ethics
161 permits A1961 and A2255.

162

163 PacBio library preparation and sequencing

164 For Pacific Biosciences (PacBio) long read sequencing, the extracted orange clownfish DNA
165 was first sheared using a g-TUBE (Covaris, MA, USA) (target size of 20 kb) and then
166 converted into SMRTbell template libraries according to the manufacturer's protocol (Pacific
167 Biosciences, CA, USA). Size selection was performed using BluePippin (Sage Science, MA,
168 USA) to generate two libraries with a minimum size of 10 and 15 kb, respectively.
169 Sequencing was performed using P6-C4 chemistry on the PacBio RS II instrument at the King
170 Abdullah University of Science and Technology (KAUST) Bioscience Core Laboratory
171 (BCL) with 360 mins movies. A total of 113 SMRT cells were sequenced.

172

173 Genome assembly

174 The genome sequence was assembled from the unprocessed PacBio reads (Table S1) using
175 the hierarchical diploid aware PacBio assembler FALCON v0.4.0 (Chin et al., 2016). To
176 obtain the optimal assembly, different parameters were tested (Table S2) to generate 12
177 candidate assemblies. The contiguity of these assemblies was assessed with QUAST v3.2
178 (Gurevich, Saveliev, Vyahhi, & Tesler, 2013), while assembly completeness was determined
179 with BUSCO v2.0 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015).
180 Assembly "A7" exhibits the highest contiguity and single copy orthologous gene
181 completeness and was selected for further improvement. The FALCON_Unzip algorithm was
182 then applied to the initial A7 assembly obtain a haplotype-resolved, phased assembly, termed
183 "A7-phased". Contigs less than 20 kb in length were removed from the assembly. This phased
184 assembly was polished with Quiver to achieve final consensus sequence accuracies
185 comparable to Sanger sequencing (Chin et al., 2013) using default settings, which produced
186 the "A7-phased-polished" assembly.
187
188 Genome assembly scaffolding with chromatin contact maps
189 The flash-frozen brain tissue was sent to Phase Genomics (Seattle, WA, USA) for the
190 construction chromatin contact maps. Tissue fixation, chromatin isolation, library preparation
191 and 80-bp paired end sequencing were performed by Phase Genomics. The sequencing reads
192 were aligned to the A7-phased-polished version of the assembly with BWA (H. Li & Durbin,
193 2010) and uniquely mapping read pairs were retained. Contigs from the A7-phased-polished
194 assembly were clustered, ordered and then oriented using Proximo (Bickhart et al., 2017;
195 Burton et al., 2013), with settings as previously described (Peichel, Sullivan, Liachko, &
196 White, 2017). Briefly, contigs were clustered into chromosomal groups using a hierarchical
197 clustering algorithm based on the number of read pairs linking scaffolds, with the final

198 number of groups specified as the number of the haploid chromosomes. The haploid
199 chromosome number was set as 24, which is consistent with the observed haploid
200 chromosome number of the Amphiprioninae, as published for *A. ocellaris* (Arai, Inoue, & Ida,
201 1976), *A. frenatus*, (Molina & Galetti, 2004; Takai & Kosuga, 2007), *A. clarkii* (Arai &
202 Inoue, 1976; Takai & Kosuga, 2007), *A. perideraion* (Supiwong et al., 2015) and *A. polymnus*
203 (Tanomtong et al., 2012). After clustering into chromosomal groups, the scaffolds were
204 ordered based on Hi-C link densities and then oriented with respect to the adjacent scaffolds
205 using a weighted directed acyclic graph of all possible orientations based on the exact
206 locations of the Hi-C links between scaffolds. Gaps between contigs were represented with
207 100 Ns and the proximity-guided assembly was named "A7-PGA". Gaps in the scaffolded
208 assembly were subsequently closed using PBJelly from PBSuite v15.8.24 (English et al.,
209 2012) with the entire PacBio read dataset and Blasr (Chaisson & Tesler, 2012) (parameters: --
210 minMatch 8 --minPctIdentity 70 --bestn 1 --nCandidates 20 --maxScore -500 --nproc 32 --
211 noSplitSubreads), to give rise to the final version of the assembly, "Nemo v1".
212
213 Genome assembly validation
214 Genomic DNA was extracted from a second individual and Illumina sequencing libraries
215 were prepared using the NEBNext Ultra II DNA library prep kit for Illumina following the
216 manufacturer's protocol. Three cycles of PCR were used to enrich the library. The sequencing
217 libraries were sequenced on two lanes of a HiSeq 2500 at the KAUST BCL. A total of
218 1,199,533,204 paired reads were generated, covering approximately 181 Gb. The 151-bp
219 paired end reads were processed with Trimmomatic v0.33 to remove adapter sequences and
220 low-quality stretches of nucleotides (parameters: 2:30:10 LEADING:20 TRAILING:20
221 SLIDINGWINDOW:4:20 MINLEN:75) (Bolger, Lohse, & Usadel, 2014).

222

223 The genome assembly size was validated by comparison to a k-mer based estimate of genome
224 size. The first half of the paired-end reads of one sequencing lane (~ 25 Gb of data) was used
225 for the k-mer estimate of genome size. Firstly, KmerGenie (Chikhi & Medvedev, 2014) was
226 used to determine the optimal k-value for a k-mer based estimation. Following that, Jellyfish
227 v2.2.6 (Marçais & Kingsford, 2011) was used with k=71 to obtain the frequency distribution
228 of all k-mers with this length. The resulting distribution was analyzed with Genomescope
229 (Vurture et al., 2017) to estimate genome size, repeat content and the level of heterozygosity.
230 To further validate the assembly, we determined the proportion of trimmed Illumina short
231 reads that mapped to the Nemo v1 assembly with BWA v0.7.10 (H. Li & Durbin, 2010) and
232 SAMtools v1.1 (H. Li et al., 2009). Additionally, the completeness of the genome assembly
233 annotation as determined by the conservation of a core set of genes was measured using
234 BUSCO with default parameters.

235

236 Repeat annotation

237 A species-specific *de novo* repeat library was assembled by combining the results of three
238 distinct repeat annotation methods. Firstly, RepeatModeler v1.08 (Smit & Hubley, 2008) was
239 used to build an initial repeat library. Secondly, we used LtrHarvest (Ellinghaus, Kurtz, &
240 Willhoeft, 2008) and LTRdigest (Steinbiss, Willhoeft, Gremme, & Kurtz, 2009), both
241 accessed *via* genometools 1.5.6 (Gremme, Steinbiss, & Kurtz, 2013), with the following
242 parameters: -seed 76 -xdrop 7 -mat 2 -mis -2 -ins -3 -del -3 -mintsd 4 -maxtsd 20 -minlenltr
243 100 -maxlenltr 6000 -maxdistltr 25000 -mindistltr 1500 -similar 90. The resulting hits were
244 filtered with LTRdigest, accepting only sequences featuring a hit to one of the hidden markov
245 models in the GyDB 2.0 database. Thirdly, TransposonPSI v08222010 (Haas, 2018) was used

246 to detect sequences with similarities to known families of transposon open reading frames. To
247 remove duplicated sequences in the combined result from all three methods a clustering with
248 USEARCH (Edgar, 2010) was performed requiring at least 90% sequence identity, and only
249 cluster representatives were retained. The resulting representative sequences were classified
250 by RepeatClassifier (part of RepeatModeler), Censor v4.2.29 (Jurka, Klonowski, Dagman, &
251 Pelton, 1996) and Dfam v2.0 (Wheeler et al., 2012), and were then blasted against the
252 Uniprot/Swissprot database (release 2017_12) to obtain a unified classification. Furthermore,
253 these three classification methods and the blast result was used to filter out spurious matches
254 to protein-coding sequence. Specifically, putative repeat sequences were only retained when
255 at least one classification method recognized the sequence as a repeat and the best match in
256 Swissprot/Uniprot was not a protein-coding gene (default blastx settings). Furthermore,
257 sequences were retained if two of the three identification methods classified the sequence as
258 repeat, but the best blast hit was not a transposable element. This *de novo* library was
259 combined with the thoroughly curated zebrafish repeat library provided by Repbase v22.05
260 (Bao, Kojima, & Kohany, 2015) and this combined library was employed for repeat masking
261 in the Nemo v1 assembly using RepeatMasker (Smit, Hubley, & Green, 2010).
262
263 <u>RNA extraction, library construction, sequencing and read processing</u>
264 Tissues for RNA extraction were dissected from one eight-month old orange clownfish
265 individual. RNA was extracted from skin, eye, muscle, gill, liver, kidney, gallbladder,
266 stomach and fin tissues using the Qiagen AllPrep kit following manufacturer's instructions.
267 Sequencing libraries were prepared using the TruSeq Stranded mRNA Library Preparation kit
268 and 150 bp paired-end sequencing was performed on one lane of an Illumina HiSeq 4000
269 machine in the KAUST BCL. The RNA-seq reads were trimmed with Trimmomatic v0.33

270 (Bolger et al., 2014) (parameters: 2:30:10 LEADING:3 TRAILING:3
271 SLIDINGWINDOW:4:15 MINLEN:40) and contamination was removed with Kraken (Wood
272 & Salzberg, 2014) by retaining only unclassified reads.
273
274 Genome assembly annotation
275 After mapping the RNA-seq data with STAR v2.5.2b (Dobin et al., 2013) to the final
276 assembly, an *ab-initio* annotation with BRAKER1 v1.9 (Hoff, Lange, Lomsadze,
277 Borodovsky, & Stanke, 2016) was performed. This initial annotation identified 49,881 genes.
278 This annotation was then integrated with external evidence using the MAKER2 v2.31.8 (Holt
279 & Yandell, 2011) gene annotation pipeline. First, the transcriptome of the orange clownfish
280 was provided to MAKER2 as EST evidence in two forms, a *de novo* assembly of the
281 preprocessed RNA-seq reads obtained with Trinity v2.4.0 (Grabherr et al., 2011), and a
282 genome-guided assembly performed with the Hisat2 v2.1.0/Stringtie v1.3.3b workflow
283 (Pertea, Kim, Pertea, Leek, & Salzberg, 2016). Second, we combined the proteomes of
284 zebrafish (GCF_000002035.6_GRCz11), Nile tilapia (GCF_001858045.1_ASM185804v2)
285 and bicolor damselfish (*Stegastes partitus*) (GCA_000690725.1), together with the
286 Uniprot/Swissprot database (release 2017_12: 554,515 sequences) and the successfully
287 detected BUSCO genes to generate a reference protein set for homology based gene
288 prediction. In the initial MAKER2 run, the annotation edit distances (AED) were calculated
289 for the BRAKER1-obtained annotation, and only gene annotations with an AED of less than
290 0.1 and a corresponding protein length of greater than 50 amino acids were retained for
291 subsequent training of the gene prediction program SNAP v2013.11.29 (Korf, 2004).
292 Similarly, the AUGUSTUS v3.2.3 (Stanke et al., 2006) gene prediction program was trained
293 on 1,850 gene annotations that possessed: an AED score of less than 0.01; an initial start

294 codon, a terminal stop codon and no in-frame stop codons; more than one exon; and no
295 introns greater than 10 kb. The hidden markov gene model of GeneMark v4.32 (Ter-
296 Hovhannisyan, Lomsadze, Chernoff, & Borodovsky, 2008) was trained by BRAKER1. The
297 final annotation was then obtained in the second run of MAKER2 with the trained models for
298 SNAP, GeneMark, and AUGUSTUS. InterProScan 5 was then used to obtain the Pfam
299 protein domain annotations for all genes. The standard gene builds were then generated. The
300 output was filtered to include all annotated genes with evidence (AED less than 1) or with a
301 Pfam protein domain, as recommended (Campbell, Holt, Moore, & Yandell, 2014).
302
303 Functional annotation
304 The protein sequences produced from the genome assembly annotation were aligned to the
305 UniProtKB/Swiss-Prot database (release 2017_12) with blastp v2.2.29 (parameters: -outfmt 5
306 -evalue 1e-3 -word_size 3 -show_gis -num_alignments 20 -max_hsps 20) and protein
307 signatures were annotated with InterProScan 5. The results were then integrated with
308 Blast2GO v4.1.9 (Gotz et al., 2008).
309
310 Genome assembly comparisons
311 For genome assembly comparisons, we compared the Nemo v1 genome assembly to the 26
312 previously reported fish chromosome-scale genome assemblies (Table S3). Comparisons were
313 made for genome assembly contiguity and completeness. Contig N50 values are reported for
314 the scaffold-scale versions of each assembly and are taken from the indicated publication
315 (Table S3), database description (Table S3) or were generated with the Perl
316 assemblathon_stats_2.pl script (Bradnam et al., 2013). Genome assembly completeness was
317 assessed by determining the proportion of the genome size that is contained within the

318 chromosome content of each assembly. It should be noted that this comparison is relative to
319 the estimated genome size and not the published assembly size. The estimated genome size
320 was taken as either the published estimated genome size in the relevant paper (Table S3), or
321 from the Animal Genome Size Database (Gregory, 2018). Where possible, k-mer derived or
322 flow cytometry-based estimates of genome size were used. Before calculation, we remove
323 stretches of Ns from the genome assemblies as these are used to arbitrarily space scaffolds
324 and do not contain actual genome information. However, this step was not possible for the
325 Asian arowana, southern platyfish, yellowtail or croaker genomes as the chromosome-scale
326 assemblies have not been made publicly available. Genome assembly completeness was
327 determined with BUSCO (Simão et al., 2015) using the Actinopterygii set of 4,584 genes and
328 the AUGUSTUS zebrafish gene model provided with the software.
329
330 Gene homology
331 To investigate the gene space of the orange clownfish genome assembly, we used
332 OrthoFinder v1.1.4 (Emms & Kelly, 2015) to identify orthologous gene relationships between
333 the orange clownfish and four related fish species. The following four fish species were
334 utilized in addition to the orange clownfish: Asian seabass
335 GCF_001640805.1_ASM164080v1 (45,223 sequences), Nile tilapia
336 GCF_001858045.1_ASM185804v2 (58,087 sequences), southern platyfish
337 GCF_000241075.1_*Xiphophorus_maculatus*-4.4.2 (23,478 sequences), and zebrafish
338 GCF_000002035.6_GRCz11 (52,829 sequences). The longest isoform of each gene was
339 utilized in the analysis, which corresponded to 25,050, 28,497, 23,043, and 32,420 sequences,
340 respectively. 26,597 sequences were used for the orange clownfish. These protein sequences
341 were reciprocally blasted against each other and clusters of orthologous genes were then

342 defined using OrthoFinder with default parameters. As part of OrthoFinder, the concatenated
343 sequences of single-copy orthologs present in all species were then used to construct a
344 phylogenetic tree, which was rooted using STRIDE (Emms & Kelly, 2017).
345
346 Database system architecture and software
347 The Nemo Genome DB database (http://nemogenome.org) was implemented on a UNIX
348 server with CentOS version 7, Apache web server and MySQL Database server. JBrowse
349 (Buels et al., 2016) was employed to visualize the genome assembly and genomic features
350 graphically and interactively. JavaScript was adopted to implement client-side rich
351 applications. The JavaScript library, jQuery (http://jquery.com) was employed. Other
352 conventional utilities for UNIX computing were appropriately installed on the server if
353 necessary. All of the Nemo Genome DB resources are stored on the server and are available
354 through HTTP access.
355
356
357 **Results and Discussion**
358 Sequencing and assembly of the orange clownfish genome
359 Genomic DNA of an individual orange clownfish (Fig. 1A) was sequenced with the PacBio
360 RS II platform to generate 1,995,360 long reads, yielding 113.8 Gb, which corresponds to a
361 121-fold coverage of the genome (Table S1). After filtering with the read pre-assembly step
362 of the Falcon assembler, 5,764,748 reads, covering 54.3 Gb and representing a 58-fold
363 coverage of the genome, were available for assembly.
364

365 To optimize the assembly parameters, we performed 12 trial assemblies using a range of
366 parameters for different stages of the Falcon assembler (Table S2). The assembly quality was
367 assessed by considering assembly contiguity (contig N50 and L50), total assembly size, and
368 also gene completeness (BUSCO) (Table 1). Assembly A7 exhibited the highest contig N50
369 (1.80 Mb), lowest contig L50 (138 contigs), lowest number of missing BUSCO genes (132)
370 and is only slightly surpassed in the longest contig metric (15.8 Mb) by the highly similar
371 assemblies A8 and A9 (16.5 Mb) (Table 1).
372
373 Genome assemblies represent a mixture of the two possible haplotypes of a diploid individual
374 at each locus. This collapsing of haplotypes may result in a loss of important sequence
375 information. However, diploid-aware assembly algorithms such as the Falcon_Unzip
376 assembler are designed to detect single nucleotide polymorphisms (SNPs) as well as structural
377 variations and to use this information to phase ("unzip") heterozygous regions into distinct
378 haplotypes (Chin et al., 2016). This procedure results in a primary assembly and a set of
379 associated haplotype contigs (haplotigs) capturing the divergent sequences. Having
380 established the parameter set that gave the best assembly metrics with Falcon, we used
381 Falcon_Unzip to produce a phased assembly ("A7-phased") of the orange clownfish (Table
382 2). The phased assembly was 905.0 Mb in length with a contig N50 of 1.85 Mb. As has been
383 seen in previous genome assembly projects (Chin et al., 2016), Falcon_Unzip produced a
384 smaller assembly with fewer contigs than the assembly produced by Falcon (Table 2). The
385 phased primary assembly was then polished with Quiver, which yielded an assembly ("A7-
386 phased-polished") with 1,414 contigs spanning 903.6 Mb with an N50 of 1.86 Mb (Table 2).
387 This polishing step closed 91 gaps in the assembly and improved the N50 by approximately
388 14.3 kb. After polishing of the "unzipped" A7-phased-polished assembly, 9,971 secondary

389 contigs were resolved, covering 340.1 Mb of the genome assembly. The contig N50 of these
390 secondary contigs was 38.2 kb, with over 99% of them being longer than 10 kb in size.
391 Relative to the 903.6 Mb A7-phased-polished primary contig assembly, the secondary contigs
392 covered 38% of the assembly size. To the best of our knowledge, this is the first published
393 fish genome assembly that has been resolved to the haplotype level with Falcon_Unzip.
394
395 Scaffolding of the orange clownfish genome assembly into chromosomes
396 To build a chromosome-scale reference genome assembly of the orange clownfish, chromatin
397 contact maps were generated by Phase Genomics (Fig. S1). Scaffolding was performed by the
398 Proximo algorithm (Bickhart et al., 2017; Burton et al., 2013) on the A7-phased-polished
399 assembly using 231 million Hi-C-based paired-end reads to produce the proximity guided
400 assembly "A7-PGA" (Table 2). The contig clustering allowed the placement of 1,073 contigs
401 into 24 scaffolds (chromosomes) with lengths ranging from 23.4 to 45.8 Mb (Tables 2 and 3).
402 While only 76% of the contigs were assembled into chromosome clusters, this corresponds to
403 98% (885.4 Mb) of total assembly length and represents 95% of the estimated genome size of
404 938.9 Mb (Tables 2 and 3). This step substantially improved the overall assembly contiguity,
405 raising the N50 20-fold from 1.86 to 38.1 Mb.
406
407 A quality score for the order and orientation of contigs within the A7-PGA assembly was
408 determined. This metric is based on the differential log-likelihood of the contig orientation
409 having produced the observed log-likelihood, relative to its neighbors (Burton et al., 2013).
410 The orientation of a contig was deemed to be of high quality if its placement and orientation,
411 relative to neighbors, was 100 times more likely than alternatives (Burton et al., 2013). In A7-
412 PGA, the placements of 524 (37%) of the scaffolds were deemed to be of high quality,

413 accounting for 775.5 Mb (87%) of the scaffolded chromosomes, indicating the robustness of
414 the assembly.
415
416 A final polishing step was performed with PBJelly to generate the final Nemo v1 assembly.
417 This polishing step closed 369 gaps, thereby improving the contig N50 by 68% and increasing
418 the total assembly length by 5.21 Mb (Tables 2 and 3). The length of each chromosome was
419 increased, with a range of 23.7 to 46.1 Mb (Fig. 1B). Gaps were closed in each chromosome
420 except for chromosome 14, leaving an average of only 28 gaps per chromosome (Table 3).
421 The final assembly is 908.9 Mb in size and has contig and scaffold N50s of 3.12 and 38.4 Mb,
422 respectively. The assembly is highly contiguous as can be observed by the fact that 50% of
423 the genome length is contained within the largest 84 contigs. 890.2 Mb (98%) of the genome
424 assembly size was scaffolded into 24 chromosomes, with only 18.8 Mb of the assembly
425 failing to be grouped. The 18.8 Mb of unscaffolded assembly is comprised of 341 contigs
426 with a contig N50 of only 57.8 kb.
427
428 Validation of the orange clownfish genome assembly size
429 The final assembly size of 908.9 Mb is consistent with the results of a Feulgen image analysis
430 densitometry-based study, which determined a C-value of 0.96 pg and thus a genome size of
431 938.9 Mb for the orange clownfish (Hardie & Hebert, 2004). Furthermore, our assembly size
432 is in keeping with estimates of genome size for other fish of the *Amphiprion* genus, which
433 range from 792 to 1193 Mb (Gregory, 2018). We additionally validated the observed
434 assembly size by using a k-mer based approach. Specifically, the k-mer coverage and
435 frequency distribution were plotted and fitted with a four-component statistical model with
436 GenomeScope (Fig. S2A). This allowed us to generate an estimate of genome size as well as

437 the repeat content and level of heterozygosity. However, varying the k-value from the
438 recommended value of 21 up to 27 yielded a corresponding increase of the estimated genome
439 size. We therefore used KmerGenie to determine the optimal k-mer length of 71 to capture the
440 available sequence information. The utilization of small k-values might partially explain the
441 reported tendency of GenomeScope to underestimate the genome size (Vurture et al., 2017).
442 The final estimate of the haploid genome length by k-mer analysis was 906.6 Mb, with 732.8
443 Mb (80%) of unique sequence and a repeat content of 173.8 Mb (19%). Furthermore, the
444 estimated heterozygosity level of 0.12% is low considering that an F1 offspring of wild
445 caught fish was sequenced (Fig. S2B). While the short-read k-mer based genome size
446 estimate of 906.6 Mb matches the final assembly size of 908.9 Mb very well, the C-value
447 derived genome size estimate is slightly larger (938.9 Mb). As an additional validation of the
448 accuracy of the genome assembly, we mapped the trimmed Illumina short reads to the Nemo
449 v1 assembly and observed that 95% of the reads mapped to the assembly and that 84% of the
450 reads were properly paired.
451
452 Based on the C-value derived genome size estimate, there is approximately 29.9 Mb (3.3%)
453 of sequence length absent from our genome assembly. It seems likely that our assembly is
454 nearly complete for the euchromatic regions of the genome given our assessment of genome
455 size and gene content completeness. However, genomic regions such as the proximal and
456 distal boundaries of euchromatic regions contain heterochromatic and telomeric repeats,
457 respectively, are refractory to currently available sequencing techniques and are typically
458 absent from genome assemblies (Bickhart et al., 2017; Hoskins et al., 2007).
459
460 Chromosome-scale fish genome assembly comparisons

461 To date, chromosome-scale genome assemblies have been released for 26 other fish species
462 (Table S3). Here, we present the first chromosome-scale assembly of a tropical coral reef fish,
463 the orange clownfish. As a measure of genome assembly quality, we assessed the contiguity
464 and completeness of these 27 chromosome-scale genome assemblies. We investigated
465 genome contiguity with the contig N50 metric and characterized genome completeness for
466 each genome assembly by calculating the proportion of the estimated genome size that was
467 assigned to chromosomes. As shown in Fig. 1C, the orange clownfish genome assembly is
468 highly contiguous, with a scaffold-scale contig N50 of 1.86 Mb, which is only surpassed by
469 the contig N50 of the Nile tilapia genome assembly. Interestingly, even though different
470 assembler algorithms were utilized, the three genome assemblies based primarily on long read
471 PacBio technology were the most contiguous, with only Nile tilapia (3.09 Mb, Canu), orange
472 clownfish (1.86 Mb, Falcon) and Asian seabass (1.19 Mb, HGAP) genome assemblies
473 yielding contig N50s in excess of 1 Mb.
474
475 While the use of long read sequencing technologies facilitates the production of highly
476 contiguous genome assemblies, scaffold sizes are still much shorter than the length of the
477 underlying chromosomes. The use of further scaffolding technologies such as genetic linkage
478 maps, scaffolding based on synteny with genome assemblies from related organisms, as well
479 as *in vitro* and *in vivo* Hi-C based methods has allowed for the production of assemblies with
480 chromosome-sized scaffolds. Here, the use of Hi-C based chromatin contact maps allowed for
481 the placement of 98% of the Nemo v1 assembly length (890.2 of 908.9 Mb) into
482 chromosomes, yielding a final assembly with a scaffold N50 of 38.4 Mb. This corresponds to
483 95% of the estimated genome size (938.9 Mb), which suggests that the Nemo v1 assembly is
484 one of the most complete fish genome assemblies published to date (Fig. 1C). Only the

485 zebrafish (94%) and Atlantic cod (91%) genome assemblies had a comparably high
486 proportion of their estimated genome sizes scaffolded into chromosome-length scaffolds (Fig.
487 1C). It is likely that the use of both PacBio long reads and Hi-C based chromatin contact maps
488 contributed to the very high proportion of the orange clownfish genome that we were able to
489 both sequence and assemble into chromosomes.
490
491 While assembly contiguity is important, genome completeness with respect to gene content is
492 also vital for producing a genome assembly that will be utilized by the research community.
493 We evaluated the completeness of the 27 chromosome-scale assemblies with BUSCO and the
494 Actinopterygii lineage, which encompasses 4,584 highly-conserved genes. When ranked by
495 the total of complete (single copy and duplicate) genes, the orange clownfish assembly is the
496 second most complete, with 4,456 (97.2%) of the orthologs identified (Fig. 2). The top ranked
497 assembly, Nile tilapia, contains only 9 more of the core set of orthologs such that it contains
498 4,465 of the orthologs (97.4%). While the assemblies based on PacBio long read technology
499 are again amongst the most complete, it should also be noted that most of the assemblies
500 analyzed showed a very high level of completeness.
501
502 Genome annotation
503 To annotate repetitive sequences and transposable elements, we constructed an orange
504 clownfish-specific library by combining the results of Repeatmodeler, LTRharvest and
505 TransposonPSI. Duplicate sequences were removed and false positives were identified using
506 three classification protocols (Censor, Dfam, RepeatClassifier) as well as comparisons to
507 Uniprot/Swissprot databases. After these filtering steps, we identified 21,644 repetitive
508 sequences. These sequences, in combination with the zebrafish library of RepBase, were then

509 used for genome masking with RepeatMasker. This lead to a total of 28% of the assembly
510 being identified as repetitive (Fig. 3A and Table S4). It was observed that there is a general
511 trend for increased repeat density towards the ends of chromosome arms (Fig. 3B and S2).
512 The total fraction of repetitive genomic sequence is in good agreement with other related fish
513 species (Chalopin, Naville, Plard, Galiana, & Volff, 2015). Similarly, the high fraction of
514 DNA transposons (~10%) is in line with DNA transposon content in other fish species
515 (Chalopin et al., 2015) but is substantially higher that what has been reported in mammals
516 (~3%) (Chalopin et al., 2015; Lander et al., 2001).
517
518 Following the characterization of repetitive sequences in the Nemo v1 genome assembly,
519 gene annotation was performed with the BRAKER1 pipeline, which trained the AUGUSTUS
520 gene predictor with supplied RNA-seq data, and a successive refinement with the MAKER2
521 pipeline. We provided BRAKER1 with mapped RNA-seq data from 10 different tissues. This
522 initial annotation comprised 49,881 genes with 55,273 transcripts. The gene finder models of
523 SNAP and AUGUSTUS were refined based on the initial annotation, and MAKER2 was then
524 used to improve the annotation using the new models and the available protein homology and
525 RNA-seq evidence. The resulting annotation contained 26,606 genes and 35,498 transcripts,
526 which feature a low mean AED of 0.12, indicating a very good agreement with the provided
527 evidence. After retaining only genes with evidence support (AED of less than 1) or an
528 annotated Pfam protein domain, the filtered annotation was comprised of 26,597 genes,
529 corresponding to 35,478 transcripts (Table 4). This result is broadly consistent with the
530 average number of genes (23,475) found in the 22 diploid fish species considered in this study
531 (Table S3). Compared to the initial annotation, genes in the final annotation are 61% longer
532 (13,049 bp) and encode mRNAs that are 80% longer (17,727 bp). The proportion of the

533 genome that is covered by coding sequences also increased to 8.1% in the final annotation.
534 Together with the observed reduction in the gene number by 47%, this indicates a substantial
535 reduction of likely false positive gene annotations of short length and/or few exons. The gene
536 density across the 24 chromosomes of our assembly varied from 23.6 genes/Mb (chromosome
537 21) to 36.5 genes/Mb (chromosome 14), with a genome-wide average of one gene every 29.7
538 Mb (Table 3). The spatial distribution of genes across all 24 chromosomes is relatively even
539 (Fig. 1B), with regions of very low gene density presumably corresponding to centromeric
540 regions. We observed that the longest annotated gene was APERC1_00006329 (26.5 kb),
541 which encodes the extracellular matrix protein FRAS1, while the gene coding for the longest
542 protein sequence was APERC1_00011517, which codes for the 18,851 amino acid protein,
543 Titin. Functional annotation was carried out using Blast2GO and yielded annotations for
544 22,507 genes (85%) after aligning the protein sequences to the UniProt/Swissprot database
545 and annotating protein domains with InterProScan.
546
547 Identification of orange clownfish-specific genes
548 To investigate the gene space of the orange clownfish relative to other fishes, we used
549 OrthoFinder v1.1.4 (Emms & Kelly, 2015) to identify orthologous relationships between the
550 protein sequences of the orange clownfish and four other fish species (Asian seabass, Nile
551 tilapia, southern platyfish and zebrafish) from across the teleost phylogenetic tree (Betancur-
552 R. et al., 2013). The vast majority of sequences (89%) could be assigned to one of 19,838
553 orthogroups, with the remainder identified as “singlets” with no clear orthologs. We observed
554 a high degree of overlap of protein sequence sets between all five species, with 75% of all
555 orthogroups (14,783) shared amongst all species (Fig. 4A). The proteins within these
556 orthogroups presumably correspond to the core set of teleost genes. Of the 14,783

557 orthogroups with at least one sequence from each species, a subset of 8,905 orthogroups
558 contained only a single sequence from each species. The phylogeny obtained from these
559 single-copy orthologous gene sequences (Fig. 4B) is consistent with the known phylogenetic
560 tree of teleost fishes (Betancur-R. et al., 2013). Interestingly, we identified a total of 4,429
561 sequences that are specific to the orange clownfish, 2,293 (49%) of which possess functional
562 annotations (Fig. 4A). Future investigations will focus on the characterization of these unique
563 genes and what roles they may play in orange clownfish phenotypic traits.
564
565
566 **Conclusion**
567 Here, we present a reference-quality genome assembly of the iconic orange clownfish, *A.*
568 *percula*. We sequenced the genome to a depth of 121X with PacBio long reads and performed
569 a primary assembly with these reads utilizing the Falcon_Unzip algorithm. The primary
570 assembly was polished to yield an initial assembly of 903.6 Mb with a contig N50 value of
571 1.86 Mb. These contigs were then assembled into chromosome-sized scaffolds using Hi-C
572 chromatin contact maps, followed by gap-filling with the PacBio reads, to produce the final
573 reference assembly, Nemo v1. The Nemo v1 assembly is highly contiguous, with contig and
574 scaffold N50s of 3.12 and 38.4 Mb, respectively. The use of Hi-C chromatin contact maps
575 allowed us to scaffold 890.2 Mb (98%) of the 908.2 Mb final assembly into the 24
576 chromosomes of the orange clownfish. An analysis of the core set of Actinopterygii genes
577 suggests that our assembly is nearly complete, containing 97% of the core set of highly
578 conserved genes. The Nemo v1 assembly was annotated with 26,597 genes with an average
579 AED score of 0.12, suggesting that most gene models are highly supported.
580

581 The high-quality Nemo v1 reference genome assembly described here will facilitate the use of
582 this now genome-enabled model species to investigate ecological, environmental and
583 evolutionary aspects of reef fishes. To assist the research community, we have created the
584 Nemo Genome DB database, http://nemogenome.org/ (Fig. 5), where researchers can access,
585 mine and visualize the genomic and transcriptomic resources of the orange clownfish.
586
587

588 **Acknowledgements**

589 This study was supported by the Competitive Research Funds OCRF-2014-CRG3-62140408
590 from the King Abdullah University of Science and Technology (KAUST) to T.R., M.L.B. and
591 P.L.M., as well as KAUST baseline support to M.L.B., M.A., T.G. and T.R. This project was
592 completed under JCU Ethics A1233 and A1415. We thank Dr. Jennifer Donelson and staff at
593 JCU's MARFU facility for assistance with animal husbandry, Dr. Susanne Sprungala for
594 DNA extraction for Illumina library preparation, KAUST BCL for the PacBio sequencing,
595 Dr. Hicham Mansour for sequencing advice and Dr. Rita Bartossek for the PacBio library
596 preparations. We thank Dr. Salim Bougouffa for stimulating discussions. We also
597 acknowledge Mr. Tane Sinclair-Taylor for providing the photograph of the orange clownfish
598 (Fig. 1A). This paper is dedicated to our good friend and colleague, Dr. Sylvain Foret.

599

600

601 **Author contributions**

602 R.L. and D.J.L. designed and performed the computational analysis. R.L., T.R., C.S. and
603 D.J.L. interpreted the results. H.O., K.M. and T.G. created the database. C.T.M. and S.F.
604 produced sequencing libraries. R.L., D.J.L, T.R., P.L.M., M.L.B., M.A. and D.J.M. wrote the
605 manuscript and all authors approved the final version. T.R. supervised the project.

606

607

608

References

Almany, G. R., Planes, S., Thorrold, S. R., Berumen, M. L., Bode, M., Saenz-Agudelo, P., … Jones, G. P. (2017). Larval fish dispersal in a coral-reef seascape. *Nature Ecology & Evolution*, *1*(6), 148.

Arai, R., & Inoue, M. (1976). Chromosomes of seven species of Pomacentridae and two species of Acanthuridae from Japan. *Bulletin of the National Museum of Nature and Science, Series A*, *2*, 73–78.

Arai, R., Inoue, M., & Ida, H. (1976). Chromosomes of four species of coral fishes from Japan. *Bulletin of the National Museum of Nature and Science, Series A*, *2*, 137–141.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11.

Berumen, M. L., Almany, G. R., Planes, S., Jones, G. P., Saenz-Agudelo, P., & Thorrold, S. R. (2012). Persistence of self-recruitment and patterns of larval connectivity in a marine protected area network. *Ecology and Evolution*, *2*(2), 444–452.

Betancur-R., R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., … Ortí, G. (2013). The tree of life and a new classification of bony fishes. *PLoS Currents*, *5*, ecurrents.tol.53ba26640df0ccaee75bb165c8c26288.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., … Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nature Genetics*, *49*(4), 643–650.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15).

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 10.

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., … Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*(1), 66.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), 1119–1125.

Buston, P. M. (2003). Mortality is associated with social rank in the clown anemonefish (*Amphiprion percula*). *Marine Biology*, *143*(4), 811–815.

Buston, P. M., Bogdanowicz, S. M., Wong, A., & Harrison, R. G. (2007). Are clownfish groups composed of close relatives? An analysis of microsatellite DNA variation in *Amphiprion percula*. *Molecular Ecology*, *16*(17), 3671–3678.

Buston, P. M., & García, M. B. (2007). An extraordinary life span estimate for the clown anemonefish *Amphiprion percula*. *Journal of Fish Biology*, *70*(6), 1710–1719.

Buston, P. M., & Wong, M. (2014). Why some animals forgo reproduction in complex societies. *American Scientist*, *102*(4), 290.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. In *Current Protocols in Bioinformatics* (Vol. 48, p. 4.11.1-4.11.39). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, *13*(1), 238.

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, *7*(2), 567–580.

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31–37.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., … Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*, 1050–1054.

Dixson, D. L. (2012). Predation risk assessment by larval reef fishes during settlement-site selection. *Coral Reefs*, *31*(1), 255–261.

Dixson, D. L., Jones, G. P., Munday, P. L., Planes, S., Pratchett, M. S., Srinivasan, M., … Thorrold, S. R. (2008). Coral reef fish smell leaves to find island homes. *Proceedings. Biological Sciences*, *275*(1653), 2831–2839.

Dixson, D. L., Jones, G. P., Munday, P. L., Planes, S., Pratchett, M. S., & Thorrold, S. R.

673 (2014). Experimental evaluation of imprinting and the role innate preference plays in
674 habitat selection in a coral reef fish. *Oecologia*, *174*(1), 99–107.
675 Dixson, D. L., Munday, P. L., & Jones, G. P. (2010). Ocean acidification disrupts the innate
676 ability of fish to detect predator olfactory cues. *Ecology Letters*, *13*(1), 68–75.
677 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R.
678 (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.
679 Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST.
680 *Bioinformatics*, *26*(19), 2460–2461.
681 Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible
682 software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18.
683 Elliott, J. K., & Mariscal, R. N. (2001). Coexistence of nine anemonefish species: differential
684 host and habitat utilization, size and recruitment. *Marine Biology*, *138*(1), 23–36.
685 Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome
686 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*,
687 *16*(1), 157.
688 Emms, D. M., & Kelly, S. (2017). STRIDE: Species tree root inference from gene duplication
689 events. *Molecular Biology and Evolution*, *34*(12), 3267–3278.
690 English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., … Gibbs, R. A. (2012).
691 Mind the Gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing
692 technology. *PLoS ONE*, *7*(11), e47768.
693 Fautin, D. G. (1991). The anemonefish symbiosis: what is known and what is not. *Symbiosis*,
694 *10*, 23–46.
695 Fautin, D. G., & Allen, G. R. (1997). Life history of Anemonefishes. In: Anemone fishes and
696 their host sea anemones. *Western Australian Museum, Perth*, 1–142.
697 Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., …
698 Conesa, A. (2008). High-throughput functional annotation and data mining with the
699 Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435.
700 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev,
701 A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference
702 genome. *Nature Biotechnology*, *29*(7), 644–652.
703 Gregory, T. R. (2018). Animal Genome Size Database. Retrieved from
704 http://www.genomesize.com

705 Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: a comprehensive software
706 library for efficient processing of structured genome annotations. *IEEE/ACM*
707 *Transactions on Computational Biology and Bioinformatics*, *10*(3), 645–656.
708 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool
709 for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.
710 Haas, B. J. (2018). TransposonPSI. Retrieved from http://transposonpsi.sourceforge.net/
711 Hardie, D. C., & Hebert, P. D. (2004). Genome-size evolution in fishes. *Canadian Journal of*
712 *Fisheries and Aquatic Sciences*, *61*(9), 1636–1646.
713 Herbert, N. A., Bröhl, S., Springer, K., & Kunzmann, A. (2017). Clownfish in hypoxic
714 anemones replenish host O_2 at only localised scales. *Scientific Reports*, *7*(1), 6547.
715 Hess, S., Wenger, A. S., Ainsworth, T. D., & Rummer, J. L. (2015). Exposure of clownfish
716 larvae to suspended sediment levels found on the Great Barrier Reef: Impacts on gill
717 structure and microbiome. *Scientific Reports*, *5*(1), 10561.
718 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1:
719 Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and
720 AUGUSTUS. *Bioinformatics*, *32*(5), 767–769.
721 Holbrook, S. J., & Schmitt, R. J. (2005). Growth, reproduction and survival of a tropical sea
722 anemone (Actiniaria): benefits of hosting anemonefish. *Coral Reefs*, *24*(1), 67–73.
723 Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database
724 management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1),
725 491.
726 Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., …
727 Celniker, S. E. (2007). Sequence finishing and mapping of *Drosophila melanogaster*
728 heterochromatin. *Science (New York, N.Y.)*, *316*(5831), 1625–1628.
729 Jarrold, M. D., Humphrey, C., McCormick, M. I., & Munday, P. L. (2017). Diel CO2 cycles
730 reduce severity of behavioural abnormalities in coral reef fish under ocean acidification.
731 *Scientific Reports*, *7*(1), 10153.
732 Jurka, J., Klonowski, P., Dagman, V., & Pelton, P. (1996). CENSOR--a program for
733 identification and elimination of repetitive elements from DNA sequences. *Computers &*
734 *Chemistry*, *20*(1), 119–121.
735 Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59.
736 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., …

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, J., Chen, X., Kang, B., & Liu, M. (2015). Mitochondrial DNA Genomes Organization and Phylogenetic Relationships Analysis of Eight Anemonefishes (Pomacentridae: Amphiprioninae). *PLoS ONE*, *10*(4), e0123894.

Litsios, G., Pearman, P. B., Lanterbecq, D., Tolou, N., & Salamin, N. (2014). The radiation of the clownfishes has two geographical replicates. *Journal of Biogeography*, *41*(11), 2140–2149.

Litsios, G., & Salamin, N. (2014). Hybridisation and diversification in the adaptive radiation of clownfishes. *BMC Evolutionary Biology*, *14*(1), 245.

Manassa, R. P., Dixson, D. L., McCormick, M. I., & Chivers, D. P. (2013). Coral reef fish incorporate multiple sources of visual and chemical information to mediate predation risk. *Animal Behaviour*, *86*(4), 717–722.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Marcionetti, A., Rossier, V., Bertrand, J. A. M., Litsios, G., & Salamin, N. (2018). First draft genome of an iconic clownfish species (*Amphiprion frenatus*). *Molecular Ecology Resources*.

McLeod, I. M., Rummer, J. L., Clark, T. D., Jones, G. P., McCormick, M. I., Wenger, A. S., & Munday, P. L. (2013). Climate change and the performance of larval coral reef fishes: the interaction between temperature and food availability. *Conservation Physiology*, *1*(1), cot024.

Mebs, D. (2009). Chemical biology of the mutualistic relationships of sea anemones with fish and crustaceans. *Toxicon : Official Journal of the International Society on Toxinology*, *54*(8), 1071–1074.

Molina, W. F., & Galetti, P. M. (2004). Karyotypic changes associated to the dispersive potential on Pomacentridae (Pisces, Perciformes). *Journal of Experimental Marine*

Biology and Ecology, *309*(1), 109–119.

Munday, P. L., Dixson, D. L., Donelson, J. M., Jones, G. P., Pratchett, M. S., Devitsina, G. V, & Døving, K. B. (2009). Ocean acidification impairs olfactory discrimination and homing ability of a marine fish. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 1848–1852.

Nilsson, G. E., Dixson, D. L., Domenici, P., McCormick, M. I., Sørensen, C., Watson, S.-A., & Munday, P. L. (2012). Near-future carbon dioxide levels alter fish behaviour by interfering with neurotransmitter function. *Nature Climate Change*, *2*(3), 201–204.

Peichel, C. L., Sullivan, S. T., Liachko, I., & White, M. A. (2017). Improvement of the Threespine Stickleback genome using a Hi-C-Based proximity-guided assembly. *The Journal of Heredity*, *108*(6), 693–700.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*(9), 1650–1667.

Pinsky, M. L., Saenz-Agudelo, P., Salles, O. C., Almany, G. R., Bode, M., Berumen, M. L., … Planes, S. (2017). Marine dispersal scales are congruent over evolutionary and ecological time. *Current Biology*, *27*(1), 149–154.

Planes, S., Jones, G. P., & Thorrold, S. R. (2009). Larval dispersal connects fish populations in a network of marine protected areas. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5693–5697.

Saenz-Agudelo, P., Jones, G. P., Thorrold, S. R., & Planes, S. (2011). Detrimental effects of host anemone bleaching on anemonefish populations. *Coral Reefs*, *30*(2), 497–506.

Salles, O. C., Pujol, B., Maynard, J. A., Almany, G. R., Berumen, M. L., Jones, G. P., … Planes, S. (2016). First genealogy for a wild marine fish population reveals multigenerational philopatry. *Proceedings of the National Academy of Sciences*, *113*(46), 13245–13250.

Schmiege, P. F. P., D'Aloia, C. C., & Buston, P. M. (2017). Anemonefish personalities influence the strength of mutualistic interactions with host sea anemones. *Marine Biology*, *164*(1), 24.

Scott, A., & Dixson, D. L. (2016). Reef fishes can recognize bleached habitat during settlement: sea anemone bleaching alters anemonefish host selection. *Proceedings. Biological Sciences*, *283*(1831), 20152694.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Simpson, S. D., Munday, P. L., Wittenrich, M. L., Manassa, R., Dixson, D. L., Gagliano, M., & Yan, H. Y. (2011). Ocean acidification erodes crucial auditory behaviour in a marine fish. *Biology Letters*, *7*(6), 917–920.

Smit, A. F. A., & Hubley, R. (2008). RepeatModeler Open-1.0.

Smit, A. F. A., Hubley, R., & Green, P. (2010). RepeatMasker Open-4.0.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server), W435–W439.

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Research*, *37*(21), 7002–7013.

Stillman, J. H., & Armstrong, E. (2015). Genomics Are Transforming Our Understanding of Responses to Climate Change. *BioScience*, *65*(3), 237–246.

Supiwong, W., Tanomtong, A., Pinthong, K., Kaewmad, P., Poungnak, P., & Jangsuwan, N. (2015). The first chromosomal characteristics of nucleolar organizer regions and karyological analysis of pink anemonefish, *Amphiprion perideraion* (Perciformes, Amphiprioninae). *Cytologica*, *80*(3), 271–278.

Szczebak, J. T., Henry, R. P., Al-Horani, F. A., & Chadwick, N. E. (2013). Anemonefish oxygenate their anemone hosts at night. *Journal of Experimental Biology*, *216*(6), 970–976.

Takai, A., & Kosuga, S. (2007). Karyotypes and banded chromosomal features in two anemonefishes (Pomacentridae, Perciformes). *Chromosome Science*, *10*(3), 71–74.

Tan, M. H., Austin, C. M., Hammer, M. P., Lee, Y. P., Croft, L. J., & Gan, H. M. (2018). Finding Nemo: Hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the Clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience*, gix137.

Tanomtong, A., Supiwong, W., Chaveerach, A., Khakhong, S., Tanee, T., & Sanoamuang, L. (2012). First report of chromosome analysis of saddleback anemonefish, *Amphiprion polymnus* (*Perciformes*, *Amphiprioninae*), in Thailand. *Cytologica*, *77*(4), 441–446.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene

833 prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised
834 training. *Genome Research*, *18*(12), 1979–1990.
835 Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., &
836 Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short
837 reads. *Bioinformatics*, *33*(14), 2202–2204.
838 Wenger, A. S., McCormick, M. I., Endo, G. G. K., McLeod, I. M., Kroon, F. J., & Jones, G.
839 P. (2014). Suspended sediment prolongs larval development in a coral reef fish. *The*
840 *Journal of Experimental Biology*, *217*(Pt 7), 1122–1128.
841 Wheeler, T. J., Clements, J., Eddy, S. R., Hubley, R., Jones, T. A., Jurka, J., … Finn, R. D.
842 (2012). Dfam: a database of repetitive DNA based on profile hidden Markov models.
843 *Nucleic Acids Research*, *41*, D70–D82.
844 Wittenrich, M. L., Turingan, R. G., & Creswell, R. L. (2007). Spawning, early development
845 and first feeding in the gobiid fish *Priolepis nocturna*. *Aquaculture*, *270*(1–4), 132–141.
846 Wong, M., Uppaluri, C., Medina, A., Seymour, J., & Buston, P. M. (2016). The four elements
847 of within-group conflict in animal societies: an experimental test using the clown
848 anemonefish, *Amphiprion percula*. *Behavioral Ecology and Sociobiology*, *70*(9), 1467–
849 1475.
850 Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification
851 using exact alignments. *Genome Biology*, *15*(3), R46.
852
853

854 **Data Accessibility**

855 The assembled and annotated genome as well as the raw PacBio reads and Illumina reads are
856 available at the Nemo Genome DB (http://nemogenome.org). Furthermore, the assembled
857 genome will be available on GenBank as BioProject PRJNA436093 and BioSample accession
858 SAMN08615572. All raw sequencing data described in this study will be available via the
859 NCBI Sequencing Read Archive.

Tables and Figures

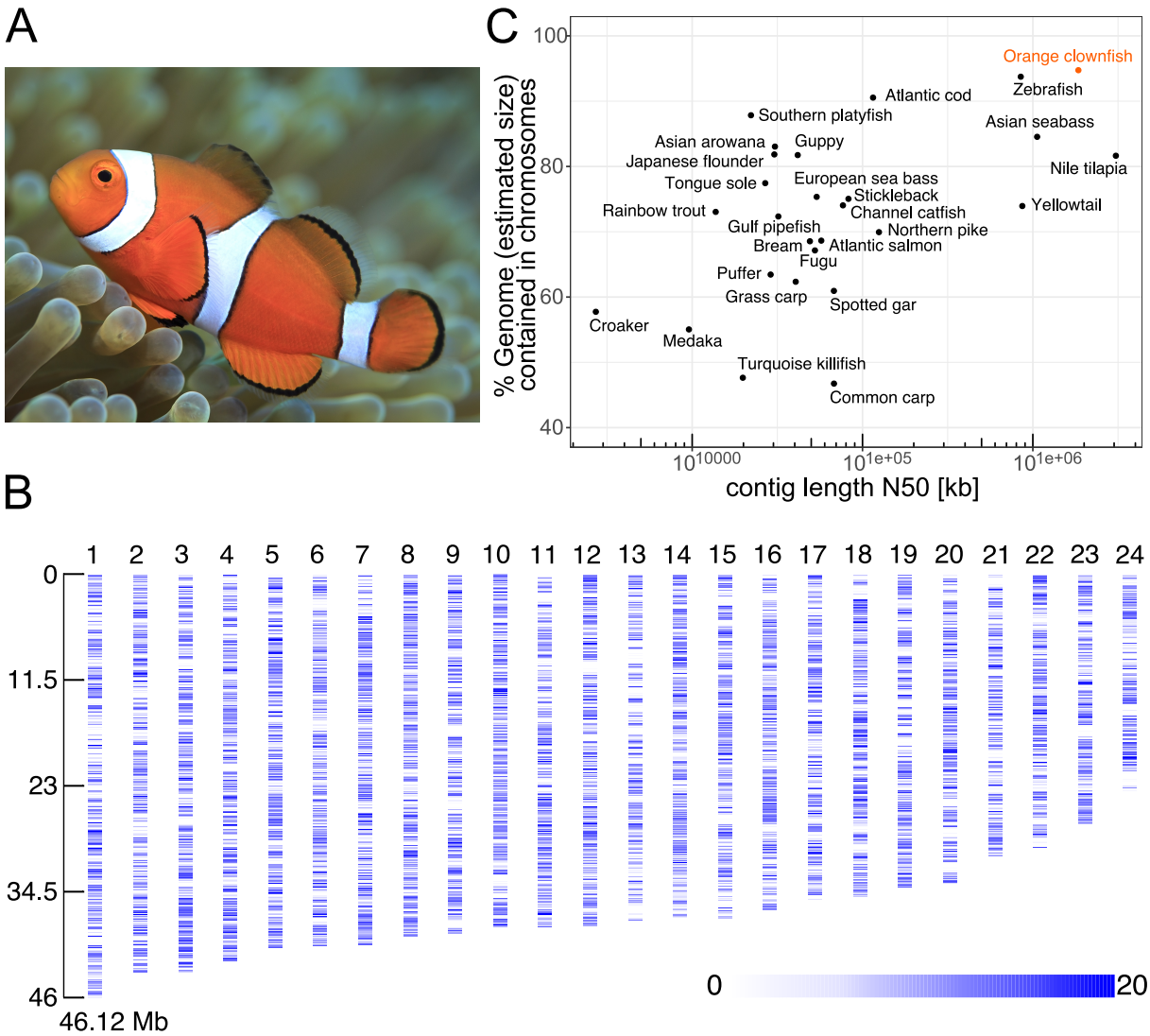

Fig. 1 **(A)** The iconic orange clownfish (*A. percula*). **(B)** Gene density on the 24 chromosomes, plotted in 100 kb windows. Chromosomes are ordered by size, as indicated on the left axis in Mb. **(C)** Contiguity (x-axis) and genome assembly completeness (y-axis) of the orange clownfish, and the 26 previously published, chromosome-scale fish genome assemblies. Details and statistics of the 27 assemblies are presented in Table S3.

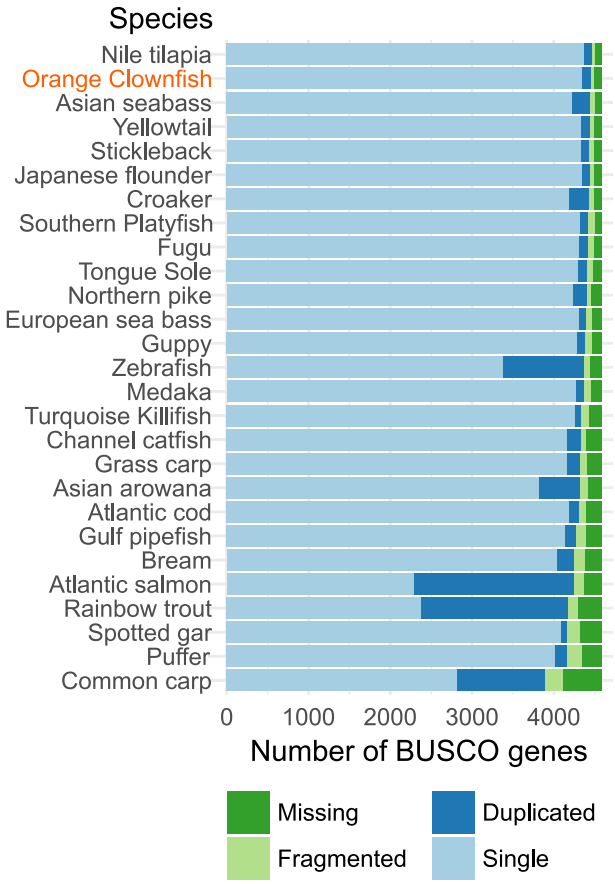

Fig. 2 Genome assembly completeness of all published chromosome-scale fish genome assemblies, as measured by the proportion of the BUSCO set of core genes detected in each assembly. Genome assemblies on the y-axis are sorted by the sum of single copy and duplicated BUSCO genes.

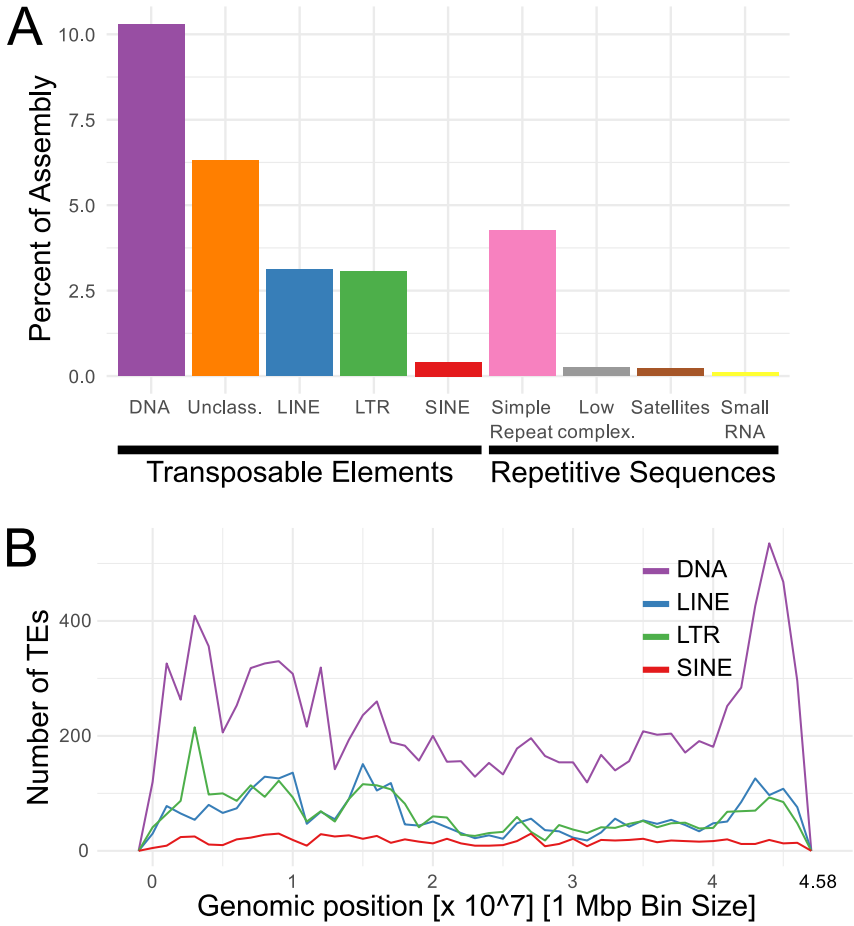

Fig. 3 Repeat content of the orange clownfish genome assembly. **(A)** Repeat content of the whole genome as classified into transposable elements and repetitive sequences. **(B)** Spatial distribution of the four main identified classes of transposable elements on chromosome 1. Transposable element spatial distribution for chromosomes 2-24 is show in Fig S2. Detailed transposable element content is shown in Table S4.

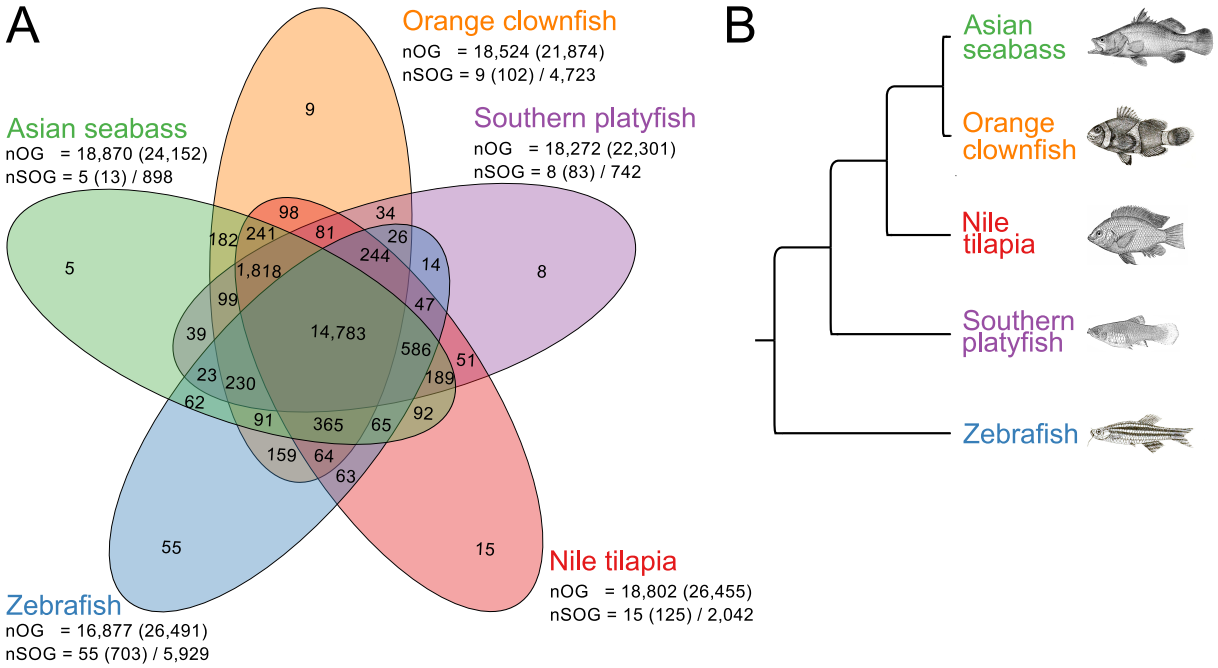

Fig. 4 (A) The overlap of orthologous gene families of the orange clownfish, southern platyfish, Nile tilapia, zebrafish and Asian seabass. The total number of orthogroups (nOG) followed by the number of genes assigned to these groups is provided below the species name. The number of species-specific orthogroups (nSOG) and the respective number of genes is also indicated, followed by the number of genes not assigned to any orthogroups. **(B)** The inferred phylogenetic tree based on the ortholog groups that contain a single gene from each species.

A

B

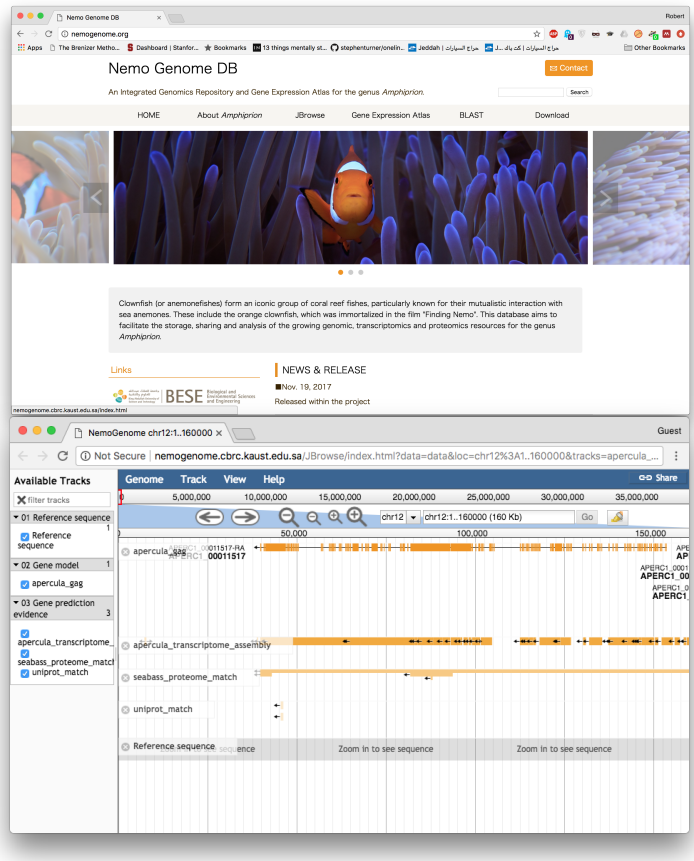

Fig. 5 (A) Front page of the Nemo Genome DB database, which is a portal to access the data described in this manuscript and is accessible at www.nemogenome.org. **(B)** Genome viewer representation of the Titin gene.

917 **Table 1** Contig statistics for the preliminary candidate-assemblies

Assembly	Length (Mb)	Number	N50 (Mb)	L50	Longest (Mb)	Missing genes (number (%))
A1	950.4	4,874	1.024	254	9.59	148 (3.23)
A2	945.4	4,374	1.040	251	6.67	156 (3.30)
A3	926.5	3,629	1.070	236	7.21	140 (3.05)
A4	921.8	2,829	1.380	184	8.16	134 (2.92)
A5	883.9	1,017	1.469	167	10.24	146 (3.18)
A6	902.2	2,204	1.401	174	12.38	134 (2.92)
A7	920.7	2,473	1.801	138	15.84	132 (2.88)
A8	924.6	2,629	1.742	143	16.51	139 (3.03)
A9	924.9	2,638	1.742	143	16.51	140 (3.05)
A10	917.1	2,368	1.648	140	10.21	146 (3.18)
A11	899.9	2,049	1.571	160	9.07	151 (3.29)
A12	908.8	2,086	1.602	142	10.21	143 (3.12)

918
919
920
921
922
923
924
925

926 **Table 2** Assembly statistics of the orange clownfish genome assemblies
927

	A7	A7-phased	A7-phased -polished	A7-PGA	Nemo v1
Technology					
Falcon	✓	-	-	-	-
Falcon_Unzip	-	✓	✓	✓	✓
PacBio	✓	✓	✓	✓	✓
Quiver	-	-	✓	✓	✓
Hi-C maps	-	-	-	✓	✓
PBJelly	-	-	-	-	✓
Contigs					
Length (Mb)	920.7	905.0	903.6	903.6	908.9
Number	2,473	1,505	1,414	1,414	1,045
N50 length (Mb)	1.80	1.85	1.86	1.86	3.12
L50 count	138	135	134	134	84
Longest (Mb)	15.84	15.83	15.85	15.9	16.6
No. Scaffolded	-	-	-	1,073	704
Scaffolds					
Length (Mb)	-	-	-	903.7	908.9
Number	-	-	-	365	365
N50 length (Mb)	-	-	-	38.1	38.4
L50 count	-	-	-	12	12
Longest (Mb)	-	-	-	45.8	46.1
Ns	-	-	-	104,900	32,395
Number of gaps	-	-	-	1,049	680
Chromosomes					
Length in chr (Mb)	-	-	-	885.4	890.2
% assembly in chr	-	-	-	98.0%	97.9%
% assembly not in chr	-	-	-	2.0%	2.1%
% of predicted genome size in chr	-	-	-	94.3%	94.8%

928
929 * Predicted genome size is 938.88 Mb (Hardie & Hebert, 2004).

930

931 **Table 3** Chromosome metrics before and after polishing of the final assembly

932

Chromosome	A7-PGA assembly		Nemo v1 assembly			
	Contigs	Length (Mb)	Contigs	Length (Mb)	Genes	Gene density (genes/Mb)
1	57	45.8	31	46.1	1,091	23.8
2	41	43.3	31	43.4	1,132	26.1
3	55	43.2	28	43.4	1,395	32.3
4	47	42.0	29	42.2	1,259	30.0
5	32	40.5	31	40.6	1,303	32.2
6	44	40.4	24	40.6	1,337	33.1
7	37	40.2	32	40.4	1,324	32.9
8	42	39.3	26	39.4	1,276	32.5
9	47	39.0	25	39.2	1,083	27.8
10	55	38.3	38	38.6	1,339	35.0
11	40	38.3	23	38.5	1,037	27.1
12	48	38.1	23	38.4	1,067	28.0
13	30	37.6	20	37.7	1,014	27.0
14	33	37.3	33	37.4	1,362	36.5
15	45	37.3	22	37.4	1,091	29.2
16	77	36.3	50	36.6	1,018	28.0
17	35	35.2	23	35.4	987	28.0
18	40	34.9	32	35.1	1,126	32.3
19	53	34.0	35	34.2	1,062	31.2
20	46	33.4	31	33.7	1,132	33.9
21	40	30.7	21	30.8	725	23.6
22	29	29.6	20	29.8	786	26.6
23	32	27.2	23	27.4	904	33.2
24	68	23.4	53	23.7	723	30.9
In chr:	1,073	885.4	704	890.2	26,309	Ave: 29.7
Not in chr:	341	18.4	341	18.8	288	15.3
Total:	1,414	903.7	1,045	908.8	26,597	Ave: 29.3

933

934 **Table 4** Gene annotation statistics

	Initial BRAKER1	Final MAKER2
Genes	49,881	26,597
mRNAs	55,273	35,478
Exons	391,637	463,688
Introns	336,364	428,210
CDSs	55,273	35,478
Overlapping genes	2,407	1,852
Contained genes	744	463
Longest gene	264,684	264,684
Longest mRNA	264,684	264,684
Mean gene length	8,097	13,049
Mean mRNA length	9,841	17,727
% of genome covered by genes	44.4	38.2
% of genome covered by CDS	7.5	8.1
Exons per mRNA	7	13
Introns per mRNA	6	12
BUSCO		
Completeness	95.94%	96.25%
Complete	4,398	4,412
Single-Copy	3,588	3,888
Duplicated	810	524
Fragmented	138	96
Missing	48	76
Total	4,584	4,584

935

936

937

938