RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

First draft genome of an iconic clownfish species (*Amphiprion frenatus*)

Anna Marcionetti[1,2] (iD) | Victor Rossier[1,2] | Joris A. M. Bertrand[1,2] | Glenn Litsios[1,2] | Nicolas Salamin[1,2]

[1]Department of Computational Biology, Biophore, University of Lausanne, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Correspondence
Nicolas Salamin, Department of Computational Biology, Biophore, University of Lausanne, Lausanne, Switzerland.
Email: nicolas.salamin@unil.ch

Abstract

Clownfishes (or anemonefishes) form an iconic group of coral reef fishes, principally known for their mutualistic interaction with sea anemones. They are characterized by particular life history traits, such as a complex social structure and mating system involving sequential hermaphroditism, coupled with an exceptionally long lifespan. Additionally, clownfishes are considered to be one of the rare groups to have experienced an adaptive radiation in the marine environment. Here, we assembled and annotated the first genome of a clownfish species, the tomato clownfish (*Amphiprion frenatus*). We obtained 17,801 assembled scaffolds, containing a total of 26,917 genes. The completeness of the assembly and annotation was satisfying, with 96.5% of the Actinopterygii Benchmarking Universal Single-Copy Orthologs (BUSCOs) being retrieved in *A. frenatus* assembly. The quality of the resulting assembly is comparable to other bony fish assemblies. This resource is valuable for advancing studies of the particular life history traits of clownfishes, as well as being useful for population genetic studies and the development of new phylogenetic markers. It will also open the way to comparative genomics. Indeed, future genomic comparison among closely related fishes may provide means to identify genes related to the unique adaptations to different sea anemone hosts, as well as better characterize the genomic signatures of an adaptive radiation.

KEYWORDS
adaptation, anemonefish, fish, genomics, Indo-Pacific

1 | INTRODUCTION

Clownfishes (or anemonefishes; subfamily Amphiprioninae, genera *Amphiprion* and *Premnas*) are an iconic and highly diverse group of coral reef fishes. They are part of the damselfish family (Pomacentridae), and they include 28 described species (Ollerton, McCollin, Fautin, & Allen, 2007). Their distribution spans the whole tropical belt of the Indo-West Pacific Ocean, but their highest species richness is situated in the Coral Triangle region, where up to nine clownfish species have been observed in sympatry (Elliott & Mariscal, 2001).

One distinctive characteristic of this group is the mutualistic interaction they maintain with sea anemones (Fautin & Allen, 1997). While all species of the clade are associated with sea anemones, there is a large variability in host usage within the group. Indeed, some species are strictly specialist and can interact with a unique species of sea anemones, while others are generalists and can cooperate with a large number of hosts (Ollerton et al., 2007). Studies have been conducted to understand both the process of host selection used by clownfishes (e.g., Arvedlund, McCormick, Fautin, & Bildsøe, 1999; Elliott, Elliott, & Mariscal, 1995; Elliott & Mariscal, 2001; Huebner, Dailey, Titus,

Khalaf, & Chadwick, 2012) and the mechanisms granting them protection from sea anemones toxins (reviewed in Mebs, 2009). However, we do not have yet a full answer for these questions. In particular, the genomic bases of these mechanisms remain poorly understood.

Clownfishes are characterized also by particular life history traits and strategies compared to other damselfishes and most other coral reef fishes. Indeed, they display an outstanding lifespan, with around 30 years estimated for *A. percula*. This lifespan is twice as long as any other damselfishes and six times greater than the expected longevity for a fish of their size (Buston & García, 2007). Moreover, clownfishes live in complex social structures within the sea anemones and are protandrous hermaphrodites. Studies have been conducted to understand the maintenance of this social structure (e.g., Buston, 2003, 2004; Hattori, 2000; Mitchell, 2003) and the mechanisms involved in sex change (e.g., Casas et al., 2016; Kim, Jin, Lee, Kil, & Choi, 2010; Kim, Lee, Habibi, & Choi, 2013; Kim, Shin, Habibi, Lee, & Choi, 2012; Miura, Kobayashi, Bhandari, & Nakamura, 2013).

Litsios et al. (2012) proposed that the obligate mutualistic interaction of clownfishes with sea anemones acted as the main key innovation that triggered the adaptive radiation of the group. It was further shown that geographic isolation associated with a rather small dispersal capacity and hybridization played a role in driving the burst of diversification and the adaptive process of this group (Litsios, Pearman, Lanterbecq, Tolou, & Salamin, 2014; Litsios & Salamin, 2014). Thus, the clownfishes potentially represent a new and interesting model system for the study of adaptive radiations and could be employed to validate the theoretical findings on the dynamics of this process (Gavrilets & Losos, 2009; Gavrilets & Vose, 2005).

Despite the many and different aspects of clownfishes that are being studied in different fields, the knowledge on their long-term evolution and its underlying genetic bases remains scarce. Yet, advances in next-generation sequencing technologies allow now to obtain genomic information also for nonmodel organisms. More precisely, the widely used Illumina short reads can be complemented with Pacific Biosciences (PacBio) long reads for hybrid assemblies (Deshpande, Fung, Pham, & Bafna, 2013; Koren et al., 2012; Miller et al., 2017). This dual strategy is fruitful as it allows to overcome the errors due to both the repeated regions of the genome that cannot be unambiguously assembled with short reads and the relatively higher error rate of long reads. Indeed, Illumina technology tends to be particularly sensitive to the first kind of error, whereas PacBio technology is expected to be more affected by the former one. Additionally, the sequencing of RNA can be used to improve the gene annotation in newly assembled genomes (Denton et al., 2014).

In this study, we aimed at obtaining the first draft genome of a clownfish species: the tomato clownfish (*Amphiprion frenatus*). This resource will provide new tools for future investigation of clownfish life history traits and the study of their mutualism with sea anemones. Additionally, new markers for phylogenetic and population genetics studies can be developed thanks to this draft reference genome. This resource also opens the way to comparative genomics among closely related fishes to identify genes related to the unique adaptations of clownfishes to their different sea anemone hosts. This

genomic resource will provide the possibility to link these different fields of research and make a step forward in the understanding of clownfish ecology and evolution.

2 | MATERIALS AND METHODS

2.1 | *Amphiprion frenatus* samples

Samples from three individuals of *Amphiprion frenatus* were obtained from a local aquarium shop. The three individuals were not from the same breeding line, and because they were acquired from an aquarium shop, the exact origin of the individuals was not available. The individuals of *A. frenatus* passed away beforehand at the aquarium shop, and samples from deceased fish were received. Thus, the three individuals did not undergo any manipulation or experimentation in the laboratory. The three individuals were used for short-reads Illumina sequencing, long-reads PacBio sequencing and RNA sequencing, respectively. The full liver sample obtained from one individual was used for RNA extraction, while the full muscle sample obtained from the second individual was used for long-reads library preparation. Fin tissue sample obtained from the third individual was used for short-reads library preparation, and the remaining sample is stored at the Department of Ecology and Evolution, University of Lausanne (sample ID: F4.6.1.3,7).

Although the use of the same individual for the generation of different sequencing data is normally preferred, the small amount of genomic DNA obtained for each individual did not allow us to use the same individual for the preparation of all the libraries. To overcome this issue, we corrected the obtained long reads with the short Illumina reads, to account for both sequencing errors and intraspecific variation (see Section 2.4).

2.2 | DNA extraction, library construction and Illumina sequencing

Genomic DNA (gDNA) was extracted from about 50 mg of fin tissue from sample F4.6.1.3,7 using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The total amount of gDNA was measured using Qubit dsDNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific, Waltham, USA). The integrity of the gDNA was verified with Fragment Analyzer Automated CE System (Advanced Analytical Technologies, Fiorenzuola d'Arda, Italy). A total of 100 ng and 4 μg of gDNA were used for paired-end (PE) and mate-pair (MP) library preparation, respectively.

Short-insert (350 bp) PE library was prepared at the Lausanne Genomic Technologies Facility (LGTF, Switzerland) using TruSeq Nano DNA LT Library Preparation Kit (Illumina). Long-insert (3 kb) MP library was prepared at Fasteris SA (Geneva, Switzerland) using the Nextera Mate Pair Library Preparation Kit from Illumina. The concentration, purity and size of the libraries obtained were verified using Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). The PE library was sequenced on two lanes of Illumina HiSeq2000 at the LGTF (run type: paired-end reads, read length of

100). The MP library was sequenced on half lane of Illumina HiSeq2500 at Fasteris (run type: paired-end reads, read length of 125 bp).

2.3 | DNA extraction, library construction and Pacific Biosciences (PacBio) sequencing

High-molecular-weight gDNA was extracted from a second individual of *A. frenatus*, from about 100 mg of muscle tissue using QIAGEN Genomic-tip 100/G (Qiagen, Hilden, Germany) following the manufacturer's instructions. The total amount of gDNA was measured using Qubit dsDNA HS Assay Kit, and the integrity of the gDNA was verified with Fragment Analyzer Automated CE System. The construction of the SMRTbell sequencing libraries and the sequencing were performed at the LGTF from a starting material of 10 μg of gDNA. The SMRTbell libraries were sequenced on eight SMRT cells (Pacific Bioscience) using C2 chemistry on the PacBio RS (Pacific Biosciences) sequencing platform.

2.4 | Preprocessing of sequenced reads

Reads quality has a major impact on the quality of the resulting assembly, and the use of error-corrected reads increases dramatically the size of the contigs (Salzberg et al., 2012). Two different PE reads correction strategies were therefore performed. The first consisted in correcting raw reads, without prior processing, with ALLPATHS-LG module for fragment read error correction using default parameters (release 44837; Gnerre et al., 2011).

The second strategy consisted of three steps. We removed PE reads that failed the chastity filtering of the CASAVA pipeline with casava_filter_se.pl (version 0.1-1, from http://brianknaus.com/software/srtoolbox/). Remaining PE reads were trimmed using SICKLE (version 1.29; Joshi & Fass, 2011), with the following parameters: –qualthreshold 30 –length-threshold 80. Substitutions due to sequencing errors in the trimmed PE reads were corrected with QUAKE (version 0.3.5; Kelley, Schatz, & Salzberg, 2010). The k-mers frequency needed by QUAKE was obtained with JELLYFISH (version 1.1; Marçais & Kingsford, 2011). A k-mer size of $k = 18$ was selected according to QUAKE documentation, which suggests the use of $k = \log(200*\text{GenomeSize})/\log(4)$. The genome size for the calculation was obtained from the ANIMAL GENOME SIZE database (Gregory, 2017), in which the reported genome sizes for the *Amphiprion* genus ranged from 792 to 1,200 Mb. The genome size of the *A. frenatus* individual was subsequently estimated from the genomic data, by dividing the number of error-free 18-mers by their peak coverage depth. The expected number of chromosomes in clownfish was also reported in the ANIMAL GENOME SIZE database (Gregory, 2017), and it is of $2n = 48$ for *A. clarkii*.

The MP reads were processed at Fasteris SA (Geneva). Because MP libraries can have a relatively low total diversity, the data set was screened for paired-end reads sharing the exact same sequences on the first 30 bases of both ends. This can be expected due to PCR duplicates, and only one of the copies was kept to obtain unique pairs. The data set was additionally screened to remove reads containing empty inserts. The linker sequences were searched and trimmed in the unique and non-empty pairs. The software SICKLE was used to remove the remaining low-quality bases (parameters: –qualthreshold 25 –length-threshold 80). The quality of the resulting MP was verified with FASTQC (version 0.11.2; Andrews, 2010).

PacBio long reads were corrected with PROOVREAD (version 2.12; Hackl, Hedrich, Schultz, & Förster, 2014) using trimmed and error-corrected PE reads. This method allows to increase SMRT sequencing accuracy, which is substantially lower compared to Illumina technologies (Goodwin, McPherson, & McCombie, 2016). Because two different individuals were sequenced, PROOVREAD also corrects the possible polymorphism based on the Illumina-sequenced individual. This will remove possible errors due to the sequencing of different individuals for the genome assembly (Zhu et al., 2016).

2.5 | Nuclear genome assembly

Trimmed MP and PE reads resulting from the two strategies of read correction were assembled using both PLATANUS (version 1.2.1; Kajitani et al., 2014) and SOAPDENOVO2 (version 2.04.240; Luo et al., 2012). One of the advantages of PLATANUS is its automatic optimization of all parameters, including k-mer size. In SOAPDENOVO2, assemblies were performed with a k-mer size of $k = 35$ and $k = 63$. The two values were chosen to span a large range, with the lower being comparable to the starting k-mer size of PLATANUS and the larger being close to the best k proposed by KMERGENIE (release 1.6982; Chikhi & Medvedev, 2013).

Scaffolding and gap-closing were performed within the PLATANUS or SOAPDENOVO2 pipelines. For scaffolding, both short-insert and long-insert libraries were used. The best genome assembly was selected by investigating assembly statistics (N50, maximum scaffold length, number of scaffolds, gap number). The best genome assembly was reached using the reads corrected with ALLPATHS-LG modules and assembled with PLATANUS. Because of the substantial better quality of PLATANUS assemblies over the SOAPDENOVO2 ones, we decided not to perform SOAPDENOVO2 assemblies by progressively increasing k-mer sizes.

We further closed gaps in the resulting best assembly using the corrected PacBio long reads with PBJELLY2 (version 14.1; English et al., 2012). The script FakeQuals.py was used to set a quality score of 40 to each base in each scaffold. The mapping of long reads on the genome in PBJELLY2 was performed by BLASR (version 1.3.1; Chaisson & Tesler, 2012), with the parameters set as following: -minPctIdentity 70 -SdpTupleSize 11 -nCandidates 20. The other parameters were left as default. Scaffolds smaller than 1 kb were removed from the final assembly.

2.6 | RNA extraction, library construction, sequencing and read processing

Liver sample from an additional individual of *A. frenatus* was obtained from a local aquarium shop for RNA sequencing to improve

gene annotation of genome assembly (Denton et al., 2014). RNA was extracted with RNeasy Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The total amount of RNA in each sample and its quality were measured using Fragment Analyzer Automated CE System.

A strand-specific cDNA library was prepared using TruSeq Stranded mRNA Sample Prep Kit (Illumina) from an initial amount of total RNA of 1 μg and following the manufacturer's instruction. The concentration, purity and size of the library were tested using Fragment Analyzer Automated CE System. The library was sequenced on one lane of Illumina HiSeq2000 at the LGTF (run type: paired-end reads, read length of 100). Obtained PE reads were trimmed with SICKLE, with the following parameters: –qual-threshold 20 –length-threshold 20.

2.7 | Nuclear genome validation

We investigated the quality of the assembled genome by evaluating the mapping rates of the PE and MP libraries using BWA (version 0.7.12; Li & Durbin, 2009), with default parameters. PE reads were subsampled, and only the reads from a single Illumina lane were used. Prior to BWA mapping, MP reads were reversed to obtain the forward–reverse orientation with a homemade script. The RNA-Seq reads from *A. frenatus* were mapped to the genome with HISAT2 (version 2.0.2, default parameters; Kim, Langmead, & Salzberg, 2015). Mapping statistics were summarized with BAMTOOLS STATS (version 2.3.0; Barnett, Garrison, Quinlan, Strömberg, & Marth, 2011). Insert sizes and read orientation were checked with PICARD (version 2.2.1, "CollectInsertSizeMetrics" tool, http://picard.sourceforge.net).

The composition of the short scaffolds (<1 kb) removed from the final assembly was assessed using BLASTN (version 2.3.30, https://blast.ncbi.nlm.nih.gov/Blast.cgi) against REFSEQ database (Release 80, E-value cut-off of 10^{-4}).

To further assess the quality of the assembly, a microsynteny analysis against *Oreochromis niloticus* genome (GCA_000188235.2) (Brawand et al., 2014) was performed with SYNCHRO (Drillon, Carbone, & Fischer, 2014). We allowed for 5 to 10 intervening genes between gene pairs, as performed in DiBattista et al. (2016). Finally, the completeness of the genome assembly was assessed with CEGMA (version 2.3) (Parra, Bradnam, & Korf, 2007).

2.8 | Nuclear genome annotation

Interspersed repeats and low-complexity DNA sequences in the genome were identified with REPEATMODELER (version 1.08, engine ncbi) and soft-masked with REPEATMASKER (version 4.0.5; Smit, Hubley, & Green, 2015). Ab initio gene prediction was carried out with BRAKER1 (version 1.9; Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2015). RNA-Seq data of *A. frenatus* previously mapped with HISAT2 (see Genome Validation) were used within BRAKER1 to improve ab initio gene prediction. RNA-Seq data were subsequently assembled into transcripts with CUFFLINK (version 2.2.1, default parameters; Trapnell et al., 2010). These transcripts, together with the ab initio gene

predictions and the proteomes of *Danio rerio* (GCA_000002035.3), *O. niloticus* (GCA_000188235.2) and *Stegastes partitus* (GCA_000690725.1), were used to provide evidence for the inference of gene structures. The different evidences were aligned on each genome and synthesized into coherent gene models with MAKER2 (version 2.31.8; Holt & Yandell, 2011). The quality of the annotation was assessed by investigating the annotation edit distance (AED), which is calculated by MAKER2.

The completeness of the resulting gene models was assessed by comparing the length of the predicted proteins with the *O. niloticus* protein length. We performed BLASTP searches against *O. niloticus* proteome (total of 47,713 proteins, E-value cut-off of 10^{-6}). We assumed that the best blast hit was orthologous and calculated the difference in protein length. We also calculated the "query" (*A. frenatus*) and "target" (*O. niloticus*) coverage, as defined in https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oreochromis_niloticus/102/ (see also Figure S1).

Functional annotation was performed with BLASTP searches against the SWISSPROT database (subset: metazoans proteins, downloaded on June 2016, total of 104,439 proteins), with an E-value cut-off of 10^{-6}. We also blasted (BLASTP, E-value cut-off of 10^{-6}) *A. frenatus* proteins against REFSEQ database (subset Actinopterygii sequences, downloaded on June 2016, total of 175,995 sequences), which is less accurate than SWISSPROT but more comprehensive. To provide further functional annotations, we used INTERPROSCAN (version 5.16.55.0; Jones et al., 2014) to predict protein domains based on homologies with the PFAM database (release 28, 16,230 families; Finn et al., 2016). Gene ontologies (GO) were annotated to each predicted protein by retrieving the GO associated with its best SWISSPROT hit. Additionally, GO associated with protein domains were annotated in the INTERPROSCAN pipeline (option -goterms).

The completeness of the genome annotation was investigated with BUSCO (version 2, data sets: Metazoan and Actinopterygii, mode: proteins; Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). Additionally, we calculated the "query" and "target" coverage for *A. frenatus* proteins and their SWISSPROT hits (Figure S1). "Query" coverage and "target" coverage were compared to *O. niloticus*, *Maylandia zebra* and *D. rerio* coverages retrieved from https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Oreochromis_niloticus/102/. The *Chaetodon austriacus* proteome was downloaded from http://caus.reefgenomics.org on January 2017. As for *A. frenatus*, protein sequences of *C. austriacus* were blasted against SWISSPROT metazoan, and "query" and "target" coverages were calculated.

2.9 | Mitochondrial genome reconstruction and annotation

We reconstructed the entire mitochondrial genome from a random subsample of 20 millions of the PE reads filtered with ALLPATH-LG. To do so, we followed the baiting and iterative mapping approach implemented in MITOBIM (version 1.9; Hahn, Bachmann, & Chevreux, 2013). We checked for the consistency of the outputs of two

reconstruction methods. First, we used as reference a previously published complete mitochondrial genome of *A. frenatus* that we retrieved from GENBANK (GB KJ833752; Li, Chen, Kang, & Liu, 2015). Alternatively, we also worked using a conspecific barcode sequence (i.e., COI gene) as a seed to initiate the process (GB FJ582759; Steinke, Zemlak, & Hebert, 2009). The circularity of the sequence was manually inferred, and the reads of the pool were mapped back onto the resulting mitochondrial genome to check for the reconstruction success and to assess coverage using GENEIOUS (version 10.2.2; Kearse et al., 2012). We used MITOANNOTATOR and the MITOFISH online database to annotate the inferred mitochondrial genome (Iwasaki et al., 2013).

3 | RESULTS AND DISCUSSION

3.1 | Nuclear genome sequencing and assembly

We obtained 534.9 million raw PE reads, corresponding to 108 Gb and 126X coverage. The ALLPATH-LG error correction led to a total of around 105 Gb, while QUAKE strategy led to a smaller number of bases in total (76 Gb). This difference is due to the strict Phred score threshold that was set to 30 during the trimming, which caused the removal of most of the reads from the data set. For MP data, we obtained 123.4 million raw pairs, which decreased to 57.1 million reads after filtering (9.6X). For PacBio long reads, we obtained 552,529 reads after correction with Illumina PE covering 1.8 Gb (2.2X). A summary of the sequencing results is provided in Table 1.

The frequency of *k*-mers in ALLPATH-LG module estimated a genome size of 857 Mb, while in QUAKE strategy, the estimated genome size was 820 Mb. The *C*-values for *Amphiprion frenatus* are not known, but available *C*-values for *A. perideraion* range from 0.81 (792 Mb) to 1.22 (1.2 Gb; from ANIMAL GENOME SIZE database, Gregory, 2017).

We selected the best genome assembly by investigating assembly statistics (N50, maximum scaffold length, number of scaffolds, gap number). The best assembly was achieved with ALLPATH-LG corrected PE and assembled with PLATANUS (Table S1). After further gap-closing with PacBio long reads, the final assembly included 17,801 scaffolds (>1 kb), which covered a total length of 803.3 Mb (Table 2). Although the number of scaffolds is still important, 95% of the assembly is contained in less than 5,000 scaffolds (Figure 1), and the N50 and N90 statistics are 244.5 and 48.1 kb, respectively. The longest scaffold measures 1.7 Mb, and the assembly contains 1.5% of gaps (Table 2). These values are comparable to other published bony fish genomes (Austin, Tan, Croft, Hammer, & Gan, 2015; DiBattista et al., 2016; Nakamura et al., 2013). For example, the genome of the Pacific bluefin tuna (*Thunnus orientalis*) is composed of 16,802 scaffolds (>2 kb), with a N50 of 137 kb and the longest scaffold of 1 Mb (Nakamura et al., 2013). Similarly, the draft genome assembly of the blacktail butterflyfish (*Chaetodon austriacus*) is composed of 13,967 scaffolds (>200 bp), with a N50 of 150.2 kb and 6.85% of gaps (DiBattista et al., 2016).

3.2 | Nuclear genome validation

The overall mapping rates for PE, MP and RNA-Seq PE data were of 99.4%, 98.2% and 92.3%, respectively (Table 3). The distribution of insert sizes estimated from the mapping was similar to the distribution obtained during the library preparation (Table 3 and Figure S2). Some larger inserts were estimated for RNA-Seq data and are explained by the presence of introns in the genome. The high mapping rates and expected insert sizes reflect an overall good assembly. This is especially true for RNA-Seq data, which was obtained independently and was not used during genome assembly.

We omitted around 1.5 million scaffolds of small size (<1 kb; 21% of the assembly) from the final assembly, the majority of which (89.5%) had no matches in the REFSEQ database.

TABLE 1 *Amphiprion frenatus* genome sequencing statistics

|	# Pairs	# Orphan	Mean length (bp)	Total # bases	Coverage[a]
Paired-end Library (350-bp insert)					
Raw data	534,892,676	0	101.0	108,048,320 552	126.1X
ALLPATH-LG	507,172,071	22,941,362	101.0	104,765,835 904	122.2X
Quake	335,811,384	88,162,294	99.9	75,932,521,405	88.6X
Mate-pair library (3-kb insert)					
Raw data	123,437,124	0	125.0	30,859,281,000	36.0X
Unique reads	96,476,505	0	125.0	24,119,126,250	28.1X
Final	57,140,128	3,245,451	70.3	8,260,918,652	9.6X
	# Reads	Mean length (bp)	Median length (bp)	Total # bases	Coverage[a]
PacBio reads					
Raw data	660,691	3,480	2,751	2,299,043,897	2.7X
Corrected reads	552,529	3,634	3,009	1,898,588,929	2.2X

Statistics are reported for both raw and preprocessed data.
[a]Coverage is calculated with an estimate genome size of *A. frenatus* of 857 Mb.

TABLE 2 *Amphiprion frenatus* genome assembly statistics

	Contigs	Scaffolds
Total assembly size	790,913,538	803,326,750
# Sequences	102,763	17,801
Longest sequence (bp)	152,672	1,727,223
Average length (bp)	7,696	45,128
GC content (%)	39.6	39.6
Non-ATGC characters (%)	0	1.5
Number of gaps	0	84,962
Sequences ≥10 Kb	25,615	5,646
Sequences ≥1 Mb	0	10
N50 index (count)[a]	14,928 (15,134)	244,530 (1,001)
N90 index (count)[a]	3,644 (55,579)	48,151 (3,637)

[a]The N50 or N90 index indicates the shortest sequence length (contig and scaffold of the final genome) above which 50% or 90% of the genome, respectively, are assembled.
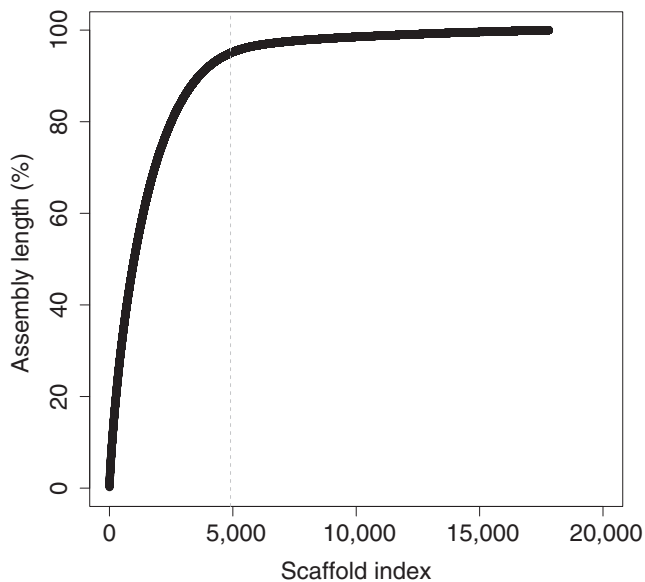
FIGURE 1 Cumulative length of *Amphiprion frenatus* assembly. Scaffolds are sorted from the longest to the shortest along the horizontal axis. The vertical dotted line indicates the number of scaffolds containing 95% of the assembly

We used CEGMA to assess the completeness of the assembly, which resulted in 99% of the core genes being either completely or partially represented in our assembly (Table S2). The microsynteny analysis of *A. frenatus* and *O. niloticus* genome gave 2,383 syntenic blocks containing a total of 13,821 (5 intervening genes) and 13,847 (10 intervening genes) genes (Table S3).

3.3 | Nuclear genome annotation

The amount of repetitive elements in our *A. frenatus* genome was 27.83%. With a combined approach of ab initio gene prediction and evidence-based homology, we identified 26,917 genes coding for 31,054 predicted proteins (Table S4). All the genes were predicted in a total of 6,497 scaffolds composing the 93% of the total assembly length. The quality of the models is satisfying, with an average and median annotation edit distance (AED) of 0.19 and 0.14, respectively (Figure S3).

The lengths of *A. frenatus* predicted proteins were compared with the corresponding *O. niloticus* best hits. A total of 28,964 predicted proteins aligned with 22,110 *O. niloticus* targets, and around half (56.3%) had less than 50 amino acids length differences with the target proteins (Figure S4, left panel). Additionally, for 20,411 *A. frenatus* proteins, the "query" coverage was higher than 90%. Similarly, the "target" coverage was higher than 90% in 17,419 cases (Figure S4, right panel).

The majority of the genes (86.5%) returned a match to SWISSPROT metazoan proteins. This number further increased to 94.9% when we blasted our data against REFSEQ database. Protein domain annotation was possible for 25,002 genes with 5,397 domains and 2,999 gene ontologies associated with these domains. A total of 17,788 gene ontologies were also mapped to 25,862 proteins (Table S5).

The largest number of genes annotated with REFSEQ is explained by a lower divergence between *A. frenatus* and the Actinopterygii species selected from the REFSEQ database. Indeed, most of the best SWISSPROT database hits were obtained with human sequences (Figure S5). This lower divergence also explains the higher identity for the matches obtained with REFSEQ database (82.1% of average identity) compared with the SWISSPROT database (61.5% of average identity). Similarly, only 4,607 proteins had identity higher than 80% with proteins from SWISSPROT, while this number increased to 19,322 for REFSEQ (Figure S6). When comparing the completeness of *A. frenatus* gene models with other Actinopterygii species, we obtained results

TABLE 3 Mapping rates for paired-end (PE), mate-pair (MP) and RNA-Seq data

	# Reads	# Mapped reads	# Mapped reads (with pair)[a]	# Concordantly mapped[b]	Mapping rate (%)	Average insert size (bp)
PE	508,471,016	505,167,148	503,236,407	472,532,563	99.4	395.9
MP	114,789,629	112,791,964	111,019,771	88,384,076	98.3	3,163.0
RNA-Seq	377,879,448	348,711,457	332,467,926	328,485,994	92.3	278.9

PE and MP reads were mapped with BWA. RNA-Seq data were mapped with HiSat.
[a]Number of reads were both pairs mapped.
[b]Concordantly mapped: pairs mapping at the expect insert size and with the right orientation.
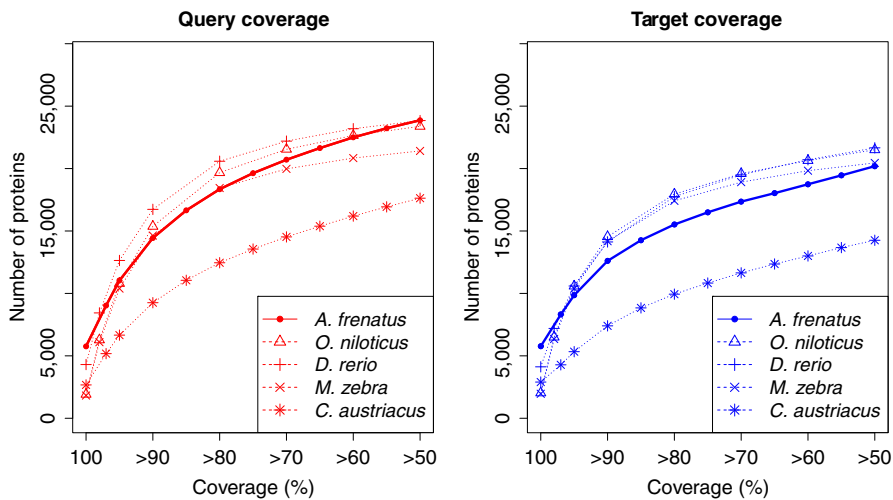
FIGURE 2 Query (left panel) and target (right panel) coverage for *Amphiprion frenatus*, *Oreochromis niloticus*, *Danio rerio*, *Maylandia zebra* and *Chaetodon austriacus* proteins and their best SWISSPROT hit proteins [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 BUSCO results for completeness of *Amphiprion frenatus* genome assembly and annotation

	Actinopterygii		Metazoan	
	Counts	Percentage	Counts	Percentage
Complete, single copy	3,542	77.3	821	83.9
Complete, duplicated	737	16.1	104	10.6
Total complete	4,279	93.3	925	94.6
Fragmented	150	3.3	19	1.9
Missing	155	3.4	34	3.5

similar to *O. niloticus*, *M. zebra* and *D. rerio*, with 14,441 *A. frenatus* proteins having a "query" coverage larger than 90% and 12,605 proteins having a "target" coverage higher than 90%. Similar results were obtained for *C. austriacus* genome (Figure 2).

BUSCO analyses were performed to assess the completeness of *A. frenatus* assembly and annotation. For metazoan BUSCOs, 3.4% of the genes were missing, while for Actinopterygii BUSCOs, 3.5% of the genes were missing (Table 4).

3.4 | Mitochondrial genome reconstruction

We successfully reconstructed the complete mitochondrial genome of *A. frenatus*. The two methods used gave highly congruent results with each other. The mapping of the reads onto the inferred sequences led to a mean coverage of 20X (4X to 35X) and confirmed that the sequence could be unambiguously reconstructed. The inferred consensus sequence had a total length of 16,740 bp, which is slightly shorter than the 16,774 bp of the two available *A. frenatus* mitochondrial sequences (GB KJ833752, Li et al., 2015; and GB LC089039). Its H-strand nucleotide composition is A: 29.6%, T: 25.7%, C: 29.3% and G: 15.4%, and its GC content is 44.7%. This circular genome has a structure that is typical of fish mitochondrial genomes. It contains 13 protein-coding genes, 22 transfer RNA (tRNAs) genes, 2 ribosomal RNAs (rRNAs), 1 control region (D-loop) plus another 33-bp short noncoding region (OL) located between the tRNA-Asn and the tRNA-Cys (see

Table S6 and Figure S7 for details). Pairwise differentiation between the three mitochondrial genomes available ranged from 0.77% to 2.0%, suggesting an interesting amount of intraspecific variation in *A. frenatus*.

4 | CONCLUSION

Here, we presented the first nuclear genomic resource for a clownfish species. Despite the fragmented nature of our assembly, the overall quality and completeness of the tomato clownfish nuclear genome are satisfying and comparable to other recent bony fish genome assemblies.

The genome that we present here, along with further sequencing of additional species and possible sequencing refinement, provides a new resource for future investigations of clownfish adaptive radiation and their particular life history traits. It will also enable a deeper understanding of the origin of the mutualistic interactions with sea anemones by opening the way for comparative genomic analyses, which could allow the identification of the genomic bases of clownfish adaptive radiation. Additionally, this resource will allow the design of new phylogenetic or population genomic markers that can be useful to study clownfish and damselfish evolution.

DATA ACCESSIBILITY

Raw Illumina and PacBio reads are available in the Sequence Read Archive (SRA), NCBI database (SRA Accession no.: SRP132439). The

assembled nuclear genome, mitogenome and their annotation are available in DRYAD Repository (https://doi.org/10.5061/dryad.nv1sv).

AUTHOR CONTRIBUTION

ORCID

Anna Marcionetti (iD) http://orcid.org/0000-0002-2450-2879

REFERENCES

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arvedlund, M., McCormick, M. I., Fautin, D. G., & Bildsøe, M. (1999). Host recognition and possible imprinting in the anemonefish *Amphiprion melanopus* (Pisces: Pomacentridae). In *Marine ecology progress series*, pp. 207–218.

Austin, C. M., Tan, M. H., Croft, L. J., Hammer, M. P., & Gan, H. M. (2015). Whole genome sequencing of the Asian Arowana (*Scleropages formosus*) provides insights into the evolution of ray-finned fishes. *Genome Biology and Evolution*, 7(10), 2885–2895. https://doi.org/10.1093/gbe/evv186

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., . . . Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518), 375–381. https://doi.org/10.1038/nature13726

Buston, P. (2003). Social hierarchies: Size and growth modification in clownfish. *Nature*, 424(6945), 145–146. https://doi.org/10.1038/424145a

Buston, P. M. (2004). Territory inheritance in clownfish. *Proceedings of the Royal Society B: Biological Sciences*, 271(Suppl. 4), S252–S254. https://doi.org/10.1098/rsbl.2003.0156

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. https://doi.org/10.1111/j.1095-8649.2007.01445.x

Casas, L., Saborido-Rey, F., Ryu, T., Michell, C., Ravasi, T., & Irigoien, X. (2016). Sex change in clownfish: Molecular insights from transcriptome analysis. *Scientific Reports*, 6, 35461. https://doi.org/10.1038/srep35461

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, 238. https://doi.org/10.1186/1471-2105-13-238

Chikhi, R., & Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1), 31–37.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology*, 10(12), e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Deshpande, V., Fung, E. D., Pham, S., & Bafna, V. (2013). Cerulean: A hybrid assembly using high throughput short and long reads. In A. Darling & J. Stoye (Eds.), *International workshop on algorithms in bioinformatics* (pp. 349–363). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-40453-5

DiBattista, J. D., Wang, X., Saenz-Agudelo, P., Piatek, M. J., Aranda, M., & Berumen, M. L. (2016). Draft genome of an iconic Red Sea reef fish, the blacktail butterflyfish (Chaetodon austriacus): Current status and its characteristics. *Molecular Ecology Resources*, 18(2), 347–355. https://doi.org/10.1111/1755-0998.12588

Drillon, G., Carbone, A., & Fischer, G. (2014). SynChro: A fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS ONE*, 9(3), e92621. https://doi.org/10.1371/journal.pone.0092621

Elliott, J. K., Elliott, J. M., & Mariscal, R. N. (1995). Host selection, location, and association behaviors of anemonefishes in field settlement experiments. *Marine Biology*, 122(3), 377–389. https://doi.org/10.1007/BF00350870

Elliott, J. K., & Mariscal, R. N. (2001). Coexistence of nine anemonefish species: Differential host and habitat utilization, size and recruitment. *Marine Biology*, 138(1), 23–36. https://doi.org/10.1007/s002270000441

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., . . . Gibbs, R. A. (2012). Mind the Gap: Upgrading genomes with pacific biosciences RS long-read sequencing technology. *PLoS ONE*, 7(11), e47768. https://doi.org/10.1371/journal.pone.0047768

Fautin, D., & Allen, G. R. (1997). Anemone fishes and their host sea anemones. Western Australian Museum, Perth, 159 pp.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(Database issue), D279–D285. https://doi.org/10.1093/nar/gkv1344

Gavrilets, S., & Losos, J. B. (2009). Adaptive radiation: Contrasting theory with data. *Science*, 323(5915), 732–737. https://doi.org/10.1126/science.1157966

Gavrilets, S., & Vose, A. (2005). Dynamic patterns of adaptive radiation. *Proceedings of the National academy of Sciences of the United States of America*, 102(50), 18040–18045. https://doi.org/10.1073/pnas.0506330102

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., . . . Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4), 1513–1518. https://doi.org/10.1073/pnas.1017351108

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gregory, T. R. (2017). Animal genome size database. http://www.genomesize.com.

Hackl, T., Hedrich, R., Schultz, J., & Förster, F. (2014). *proovread*: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, 30(21), 3004–3011. https://doi.org/10.1093/bioinformatics/btu392

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, 41(13), e129. https://doi.org/10.1093/nar/gkt371

Hattori, A. (2000). Social and mating systems of the protandrous anemonefish Amphiprion perideraion under the influence of a larger congener. *Austral Ecology*, *25*(2), 187–192. https://doi.org/10.1046/j.1442-9993.2000.01035.x

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2015). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769.

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

Huebner, L. K., Dailey, B., Titus, B. M., Khalaf, M., & Chadwick, N. E. (2012). Host preference and habitat segregation among Red Sea anemonefish: Effects of sea anemone traits and fish life stages. *Marine Ecology Progress Series*, *464*, 1–15. https://doi.org/10.3354/meps09964

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., ... Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, *30*(11), 2531–2540. https://doi.org/10.1093/molbev/mst141

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... Pesseat, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software].

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... Kohara, Y. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*(8), 1384–1395. https://doi.org/10.1101/gr.170720.113

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: Quality-aware detection and correction of sequencing errors. *Genome Biology*, *11*(11), R116. https://doi.org/10.1186/gb-2010-11-11-r116

Kim, N. N., Jin, D. H., Lee, J., Kil, G. S., & Choi, C. Y. (2010). Upregulation of estrogen receptor subtypes and vitellogenin mRNA in cinnamon clownfish *Amphiprion melanopus* during the sex change process: Profiles on effects of 17β-estradiol. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *157*(2), 198–204. https://doi.org/10.1016/j.cbpb.2010.06.003

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, N. N., Lee, J., Habibi, H. R., & Choi, C. Y. (2013). Molecular cloning and expression of caspase-3 in the protandrous cinnamon clownfish, *Amphiprion melanopus*, during sex change. *Fish Physiology and Biochemistry*, *39*(3), 417–429. https://doi.org/10.1007/s10695-012-9709-y

Kim, N. N., Shin, H. S., Habibi, H. R., Lee, J., & Choi, C. Y. (2012). Expression profiles of three types of GnRH during sex-change in the protandrous cinnamon clownfish, *Amphiprion melanopus*: Effects of exogenous GnRHs. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *161*(2), 124–133. https://doi.org/10.1016/j.cbpb.2011.10.003

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J., Ganapathy, G., ... Phillippy, A. M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

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. https://doi.org/10.1371/journal.pone.0123894

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

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. https://doi.org/10.1111/jbi.12370

Litsios, G., & Salamin, N. (2014). Hybridisation and diversification in the adaptive radiation of clownfishes. *BMC Evolutionary Biology*, *14*, 245. https://doi.org/10.1186/s12862-014-0245-5

Litsios, G., Sims, C. A., Wüest, R. O., Pearman, P. B., Zimmermann, N. E., & Salamin, N. (2012). Mutualism with sea anemones triggered the adaptive radiation of clownfishes. *BMC Evolutionary Biology*, *12*, 212. https://doi.org/10.1186/1471-2148-12-212

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., ... Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, *1*, 18. https://doi.org/10.1186/2047-217X-1-18

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Mebs, D. (2009). Chemical biology of the mutualistic relationships of sea anemones with fish and crustaceans. *Toxicon*, *54*(8), 1071–1074. https://doi.org/10.1016/j.toxicon.2009.02.027

Miller, J. R., Zhou, P., Mudge, J., Gurtowski, J., Lee, H., Ramaraj, T., ... Silverstein, K. A. T. (2017). Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics*, *18*, 541. https://doi.org/10.1186/s12864-017-3927-8

Mitchell, J. S. (2003). Social correlates of reproductive success in false clown anemonefish: Subordinate group members do not pay-to-stay. *Evolutionary Ecology Research*, *5*(1), 89–104.

Miura, S., Kobayashi, Y., Bhandari, R. K., & Nakamura, M. (2013). Estrogen favors the differentiation of ovarian tissues in the ambisexual gonads of anemonefish Amphiprion clarkii. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, *319*(10), 560–568. https://doi.org/10.1002/jez.1818

Nakamura, Y., Mori, K., Saitoh, K., Oshima, K., Mekuchi, M., Sugaya, T., ... Inouye, K. (2013). Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(27), 11061–11066. https://doi.org/10.1073/pnas.1302051110

Ollerton, J., McCollin, D., Fautin, D. G., & Allen, G. R. (2007). Finding NEMO: Nestedness engendered by mutualistic organization in anemonefish and their hosts. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1609), 591–598. https://doi.org/10.1098/rspb.2006.3758

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., ... Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*(3), 557–567. https://doi.org/10.1101/gr.131383.111

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. https://doi.org/10.1093/bioinformatics/btv351

Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology. http://repeatmasker.org.

Steinke, D., Zemlak, T. S., & Hebert, P. D. N. (2009). Barcoding Nemo: DNA-Based Identifications for the Ornamental Fish Trade. *PLoS ONE*, *4*(7), e6300. https://doi.org/10.1371/journal.pone.0006300

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., ... Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Zhu, W., Wang, L., Dong, Z., Chen, X., Song, F., Liu, N., ... Fu, J. (2016). Comparative transcriptome analysis identifies candidate genes related to skin color differentiation in red tilapia. *Scientific Reports*, *6*, 31347. https://doi.org/10.1038/srep31347

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.