A chromosome-scale genome assembly of the false clownfish, *Amphiprion ocellaris*

Taewoo Ryu, 1,*,† Marcela Herrera, 2,† Billy Moore, 1 Michael Izumiya, 1, Erina Kawai, 1, Vincent Laudet, 2,3 Timothy Ravasi, 1,4,*

1Marine Climate Change Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495 Japan
2Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495 Japan
3Marine Research Station, Institute of Cellular and Organismic Biology, Academia Sinica, I-Lan, Taiwan
4Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD 4811, Australia

*Corresponding author: Marine Climate Change Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa 904-0495 Japan. Email: taewoo.ryu@oist.jp; †Corresponding author: Marine Climate Change Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa 904-0495 Japan. Email: timothy.ravasi@oist.jp

These authors contributed equally to this work.

**Abstract**

The false clownfish *Amphiprion ocellaris* is a popular fish species and an emerging model organism for studying the ecology, evolution, adaptation, and developmental biology of reef fishes. Despite this, high-quality genomic resources for this species are scarce, hindering advanced genomic analyses. Leveraging the power of PacBio long-read sequencing and Hi-C chromosome conformation capture techniques, we constructed a high-quality chromosome-scale genome assembly for the clownfish *A. ocellaris*. The initial genome assembly comprised of 1,551 contigs of 861.42 Mb, with an N50 of 863.85 kb. Hi-C scaffolding of the genome resulted in 24 chromosomes containing 856.61 Mb. The genome was annotated with 26,797 protein-coding genes and had 96.62% completeness of conserved actinopterygian genes, making this genome the most complete and high quality among published anemonefish genomes. Transcriptional analysis identified tissue-specific gene expression patterns, with the brain and optic lobe having the largest number of expressed genes. Further, comparative genomic analysis revealed 91 genome elements conserved only in *A. ocellaris* and its sister species *Amphiprion percula*, and not in other anemonefish species. These elements are close to genes that are involved in various nervous system functions and exhibited distinct expression patterns in brain tissue, potentially highlighting the genetic toolkits involved in lineage-specific divergence and behaviors of the clownfish branch. Overall, our study provides the highest quality *A. ocellaris* genome assembly and annotation to date, whilst also providing a valuable resource for understanding the ecology and evolution of reef fishes.

**Keywords:** *Amphiprion ocellaris*; anemonefish; clownfish; genome; chromosome-scale assembly

**Introduction**

The false clownfish *Amphiprion ocellaris* is one of 28 anemonefishes (from the subfamily Amphiprioninae in the family Pomacentridae) among thousands of tropical marine fish species (Roux et al. 2020). Yet, together with its sister species, the orange clownfish *Amphiprion percula*, it is one of the most recognizable fish, especially among the nonscientific community, following the Disney movie “Finding Nemo” (Militz and Foale 2017). Even before the release of this film more than 15 years ago, the visual appeal and ability to complete their life cycle in captivity made clownfish a highly desired species in the marine aquarium trade (Rhyne et al. 2017; Militz et al. 2018). For biologists, on the other hand, anemonefishes offer a unique opportunity to answer complex research questions about symbiosis, social dynamics, sex change, speciation, and phenotypic plasticity (Roux et al. 2020).

Until now, genome assemblies of at least 10 anemonefish species including *A. ocellaris* have been published (Marcionetti et al. 2018, 2019; Tan et al. 2018; Lehmann et al. 2019). Yet, except for *A. percula*, these genomes are mainly based on Illumina short-read technology and are therefore highly fragmented, resulting in multiple gaps and misassemblies. However, third-generation sequencing platforms such as Pacific Biosciences, produce longer reads (5–60 kb) that enhance the continuous assembly of genome sequences (van Dijk et al. 2018; Logsdon et al. 2020). This makes it possible to assemble complex regions of genomes, thus improving our ability to decipher genomic structures (such as chromosome rearrangements) and long-range regulatory analysis (Rhee et al. 2021). For example, 29% of N-gaps in the human reference genome (GRCh38) could be filled with PacBio long reads (Shi et al. 2016). In the case of *A. ocellaris*, the inclusion of Nanopore long reads together with Illumina data led to a 94% decrease in the number of scaffolds, an 18 times increase in scaffold N50 (401.72 kb), and a 16% improvement in genome completeness (Tan et al. 2018). The PacBio long-read assembly of the *A. percula* genome further emphasized the power of long-read technology, with an initial contig assembly N50 of 1.86 Mb, further anchored...
into the chromosome-scale assembly (scaffold N50 of 38.4 Mb) (Lehmann et al. 2019).

Here, we constructed a high-quality chromosome-scale genome assembly and gene annotation for the false clownfish _A. ocellaris_. Using a combination of PacBio and Hi-C sequencing, we produced a de novo assembly comprised of 1,551 contigs with an N50 length of 863,854 bp that were successfully anchored into 24 chromosomes of 856,612,077 bp. We annotated 26,797 protein-coding gene models with the proportion of conserved actinopterygian genes reaching 96.62%, making the quality and completeness of our genome better than previously published anemonefish genomes. A comparative genomic approach identified genomic elements conserved only in the _A. ocellaris_/ _A. percula_ branch but not in other anemonefishes, many of which are associated with genes involved in nervous system functioning and were differentially expressed in brain tissues. Ultimately, our work adds to the growing body of high-quality fish genomes critical to study genetic, ecological, evolutionary, and developmental aspects of marine fishes in general.

**Materials and methods**

**Specimen collection and nucleic acid sequencing**

Three adult _A. ocellaris_ clownfish (1 female and 2 males) were collected from 5 m depth in Motobu, Okinawa (26°7′29.83″N, 127°9′57.51″E) on 2020 March 25. Fish were kept under natural conditions at the OIST Marine Science Station in a 270-l (60 × 90 × 50 cm) tank until 2020 May 19. Individuals were euthanized following the guidelines for animal use issued by the Animal Resources Section of OIST Graduate University. Tissues for genome sequencing were snap frozen in liquid nitrogen and then stored at −80°C until further processing. Genomic DNA was extracted from a male clownfish using a Qiagen tissue genomic DNA extraction kit (Hilden, Germany) and sequenced at Macrogen (Tokyo, Japan). For genome assembly, we sequenced genomic DNA from the brain tissue of the same male fish using 2 different platforms: PacBio Sequel II and Illumina NovaSeq6000 (Supplementary Table 1). For long-read sequencing, 8 μg of genomic DNA was used to generate a 20 kb SMRTbell library according to the manufacturer’s instructions (Pacific Biosciences, CA, USA). Briefly, a 10-μl SMRTbell library was prepared using a SMRTbell Express Template Prep Kit 2.0 and the resulting templates were bound to DNA polymerases with a Sequel II Binding Kit 2.0 and Internal Control Kit 1.0. Sequencing on the PacBio Sequel II platform was performed using a Sequel II Sequencing Kit 2.0 and a SMRT cells 8M Tray. SMRT cells using 15 h movies were captured. For short-read sequencing, a library was prepared from 1 μg of genomic DNA and a TruSeq DNA PCR-free Sample Preparation Kit (Illumina, CA, USA). Paired-end (151 bp per read) sequencing was conducted using a NovaSeq6000 platform (Illumina, CA, USA).

Hi-C reads were also sequenced to capture chromatin conformation for chromosome assembly. Liver (>100 mg) tissue from another male fish was snap frozen and stored at −80°C (Supplementary Table 1). The tissue was sliced into small pieces using a razor blade (to increase the surface area for efficient cross-linking), resuspended in 15 ml of 1% formaldehyde solution, and then incubated at room temperature for 20 min with periodic mixing. Glycine powder was added to the solution for a final concentration of 125 mM followed by a 15-min incubation at room temperature with periodic mixing. Samples were spun down at 1,000 g for 1 min, the supernatant was removed, and the tissue was rinsed with Milli-Q water. Tissues were then ground into a fine powder using a liquid nitrogen-chilled mortar and pestle. Powdered samples were collected and stored at −80°C. Chromatin isolation, library preparation, and Hi-C sequencing was performed by Phase Genomics (WA, USA). Following the manufacturer’s instructions, a Proximo Hi-C 2.0 Kit (Phase Genomics, WA, USA) was used to prepare the proximity ligation library and process it into an Illumina-compatible sequencing library. Hi-C reads were sequenced on an Illumina NovaSeq6000 platform to generate 150 bp paired-end reads.

Tissues for transcriptome sequencing were dissected from 2 individuals (1 male and 1 female) and stored in RNAlater stabilization solution (Sigma Life Science, MO, USA) at −80°C. Transcriptome library preparation and sequencing were performed by Macrogen (Tokyo, Japan). Briefly, mRNA was extracted from brain optic lobe, caudal fin, eye, gill, gonads (from male and female fish), intestine, kidney, liver, the rest of the brain, skin (from orange and white bands), and stomach tissues using a Qiagen RNeasy Mini Kit (Hilden, Germany). Only high-quality RNA samples with an RNA integration number >7.0 were used for library construction. Libraries were prepared with 1 μg of total RNA for each sample using a TruSeq Stranded mRNA Sample Prep Kit (Illumina, CA, USA). Paired-end sequencing (151 bp) was conducted on a NovaSeq6000 machine.

**Chromosome-scale genome assembly of _A. ocellaris_**

Prior to de novo assembly genome size was estimated using Jellyfish v2.3.0 (Marçais and Kingsford 2011) with k-mer = 17 and default parameters, and GenomeScope v1.0 (Vurture et al. 2017) with default parameters. Quality-trimmed Illumina short reads obtained from Trimmomatic v0.39 (Bolger et al. 2014) using the strategy "ILLUMINACLIP: TruSeq3-PE.fa: 2:30:10:8: [[Y/N]:1] NCM: 1 NHT: 2 [[N/1] NCM:1 NHT:2}". PCR duplicates were flagged using SAMBLASTER v0.1.26 (Faust and Hall 2014) and subsequently removed from all following analyses. Nonprimary and secondary alignments were filtered using SAMtools v1.10 (Li et al. 2009) with the -f 2304 flag. FALCON-Phase (Kronenberg et al. 2018) was then used to correct phase switching errors in the scaffolds obtained from FALCON-Unzip (Chin et al. 2016). A genome-wide contact frequency matrix was built from the aligned Hi-C read pairs and normalized by the number of DPNII restriction sites (GATC) on the scaffolds, as previously described (Bickhart et al. 2017). A total of 40,000 individual FALCON runs were performed to optimize chromosome construction. Juicebox v1.13.01 (Durand et al. 2016) was used to correct scaffolding errors and FALCON-Phase was again used to correct phase switching errors detectable at the chromosome level but not at the scaffold level. Local base accuracy in the long read-based draft assembly was improved with Illumina short reads using Plion v1.23 (Walker et al. 2014). Quality-trimmed Illumina short reads obtained from Trimmomatic v0.39 (Bolger et al. 2014) using the same parameter set described above were aligned to the proximo-assembled chromosome-scale genome with Bowtie2 v2.4.1 (Langmead and Salzberg 2012) using the default settings. SAM files were converted to BAM files with SAMtools v1.10 (Li et al. 2016).
et al. 2009) and then used as input for Pilon. Error correction was completed using 5 iterations of Pilon.

To calculate the overall mean genome-wide base level coverage, PacBio reads were aligned to the assembled chromosome sequences using Pbbm2 v1.4.0 (https://github.com/PacificBiosciences/pbbm2). Per-base coverage of aligned reads across entire chromosomal sequences was obtained using the BEDTools v2.30.0 (Quinlan 2014) genomeCoverageBed function. Finally, we compared the quality of our genome assembly to 3 other published A. ocellaris genome sequences (Tan et al. 2018; Marcionetti et al. 2019) using Quast v5.0.2 (Mikheenko et al. 2018).

Prediction of gene models in A. ocellaris
Repetitive elements in the A. ocellaris genome were identified de novo using RepeatModeler v2.0.1 (Flynn et al. 2020) with the parameter -LTRStruct. RepeatMasker v4.1.1 (Tempel 2012) was then used to screen known repetitive elements with 2 separate inputs: the RepeatModeler output and the vertebrata library of Dfam v3.3 (Storer et al. 2021). The 2 output files were validated, merged, and redundancy was removed using GenomeTools v1.6.1 (Gremme et al. 2013).

BRAKER v2.1.6 (Brüna et al. 2021) was then used to annotate candidate gene models of A. ocellaris. For mRNA evidence for gene annotation, transcriptomic reads (Supplementary Table 1) were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) using the parameter set mentioned above and mapped to the chromosome sequences with HISAT2 v2.2.1 (Kim et al. 2019) using the “-du” option. SAM files were then converted to BAM format using SAMTools v1.10 (Li et al. 2009). For protein evidence, manually annotated and reviewed protein records from UniProtKB/Swiss-Prot (UniProt Consortium 2021) as of 2021 January 11 (563,972 sequenced and annotated and reviewed protein records from UniProtKB/Swiss-Prot) were used to screen known repetitive elements with 2 separate inputs: the RepeatModeler output and the vertebrata library of Dfam v3.3 (Storer et al. 2021). The 2 output files were validated, merged, and redundancy was removed using GenomeTools v1.6.1 (Gremme et al. 2013).

BRAKER v2.1.6 (Brüna et al. 2021) was then used to annotate candidate gene models of A. ocellaris. For mRNA evidence for gene annotation, transcriptomic reads (Supplementary Table 1) were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) using the parameter set mentioned above and mapped to the chromosome sequences with HISAT2 v2.2.1 (Kim et al. 2019) using the “-du” option. SAM files were then converted to BAM format using SAMTools v1.10 (Li et al. 2009). For protein evidence, manually annotated and reviewed protein records from UniProtKB/Swiss-Prot (UniProt Consortium 2021) as of 2021 January 11 (563,972 sequenced and annotated and reviewed protein records from UniProtKB/Swiss-Prot) were used to screen known repetitive elements with 2 separate inputs: the RepeatModeler output and the vertebrata library of Dfam v3.3 (Storer et al. 2021). The 2 output files were validated, merged, and redundancy was removed using GenomeTools v1.6.1 (Gremme et al. 2013).

BRAKER v2.1.6 (Brüna et al. 2021) was then used to annotate candidate gene models of A. ocellaris. For mRNA evidence for gene annotation, transcriptomic reads (Supplementary Table 1) were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) using the parameter set mentioned above and mapped to the chromosome sequences with HISAT2 v2.2.1 (Kim et al. 2019) using the “-du” option. SAM files were then converted to BAM format using SAMTools v1.10 (Li et al. 2009). For protein evidence, manually annotated and reviewed protein records from UniProtKB/Swiss-Prot (UniProt Consortium 2021) as of 2021 January 11 (563,972 sequenced and annotated and reviewed protein records from UniProtKB/Swiss-Prot) were used to screen known repetitive elements with 2 separate inputs: the RepeatModeler output and the vertebrata library of Dfam v3.3 (Storer et al. 2021). The 2 output files were validated, merged, and redundancy was removed using GenomeTools v1.6.1 (Gremme et al. 2013).

Assembly and annotation of the mitochondrial genome
The mitochondrial genome of A. ocellaris was assembled using Norgal v1.0.0 (Al-Nakeeb et al. 2017) with quality trimmed Illumina genomic reads. MitoAnnotator v3.67 (Sato et al. 2018) was used to annotate the organelle genes. Annotated genes in this study were compared with previously published A. ocellaris genes using BLASTn v2.10.0 (Altschul et al. 1990) with e-value 10^-4 as a threshold to predict homology. Only the longest isoform of each gene model was used for the homology search.

Analysis of gene expression
Transcriptomic reads from each tissue were processed with Trimomatic v0.39 (Bolger et al. 2014) using the parameter set mentioned above and mapped to the genome using HISAT2 v2.2.1 (Kim et al. 2019). SAM files were then converted to BAM files using SAMTools v1.10 (Li et al. 2009). Expression levels were quantified and TPM (transcripts per million) was normalized with StringTie v2.1.4 (Pertea et al. 2016). Tissue-specific index (τ) was calculated for each gene using the R package tissue v0.99 (Condron 2020), with the relationship between τ and TPM expression values visualized on a 2D histogram with ggplot2 v3.3.5 (Wickham 2009). TPM expression values per tissue were visualized in an UpSet plot with the UpSetR v1.4.0 package (Conway et al. 2017).

Gene orthology and phylogenetic analyses
To identify evolutionary relationships between A. ocellaris and other Amphiprioninae species, 2 species combinations were used: (1) a dataset that includes 11 anemonefish proteomes, i.e. our A. ocellaris proteome and 10 other anemonefishes (Amphiprion akallopisos, Amphiprion bicinctus, Amphiprion frenatus, Amphiprion melanopus, Amphiprion nigriceps, Amphiprion percula, Amphiprion perideraion, Amphiprion polymnus, Amphiprion sebae, and Premnas biaculeatus) (Marcionetti et al. 2018, 2019; Lehmann et al. 2019), and A. polycanthus as a single outgroup species, and (2) a dataset comprised of all 11 anemonefishes, A. polycanthus, and 5 additional outgroup species across the teleost phylogenetic tree: zebrafish (D. rerio), bicolor damselfish (S. partitus), Asian seahass (Lates calcarifer), Nile tilapia (O. niloticus), and southern platyfish (Xiphophorus maculatus). The proteomes of outgroup species were obtained as previously described (Lehmann et al. 2019). In all cases, only the longest isoform of each gene model was utilized. Ortholog gene relationships between all taxa were investigated using OrthoFinder v2.5.2 (Emms and Kelly 2019). Proteins were reciprocally blasted against each other, and clusters of orthologous genes (i.e. genes descended from a single gene in the last common ancestor) were defined using the default settings. Phylogenetic relationships of fish species were then assessed based on concatenated multialignments of one-to-one orthologs. In brief, sequences of single-copy orthologs present in all species were first aligned using MAFFT v7.130 (Katoh and Standley 2013) using the options “-localpair -maxiterate 1000 -leavesgappregion,” then trimmed with trimAl v1.2 (Capella-Gutiérrez et al. 2009) using the “-gappyout” flag, and finally concatenated with FASconCAT-G (Kück and Longo 2014).

Phylogenetic trees were first constructed based on maximum-likelihood criteria using the 2 datasets described above. The MPI version of RAXML v8.2.9 (raxmlHPC-MPI-AVX) (Stamatakis 2014) was executed using a LG substitution matrix, heterogeneity model GAMMA, and 1,000 bootstrap inferences. Next, a subset of proteins for each species that has a complete match to the Actinopterygii-lineage (actinopterygii_odb10) identified by BUSCO v4.1.4 (Simão et al. 2015) were selected, concatenated, and used to construct new maximum-likelihood and Bayesian trees. Bayesian tree reconstructions were conducted under the CAT-GTR model as implemented in PhyloBayes MPI v1.8 (Lartillot et al. 2013). Two independent chains were run for at least 5,000 cycles and sampled every 10 trees. The first 2,000 trees were removed as burn-in. Chain convergence was evaluated so that the maximum and average differences observed at the end of each run were < 0.01 in
all cases. Trees were visualized and rerooted using iTOL v6.4 (Letunic and Bork 2021). Branch supports in the phylogenetic trees were evaluated with the standard bootstrap values from RAxML and PhyloBayes for maximum-likelihood and Bayesian trees, respectively. Site concordance factors (i.e. the proportion of alignment sites that support each branch) were also evaluated using IQ-TREE v2.1.3 (Minh et al. 2020).

Interspecies synteny
Patterns of synteny (i.e. the degree to which genes remain on corresponding chromosomes) and collinearity (i.e. in corresponding order) across all anemonefish genomes were investigated using the MCScanX toolkit (Wang et al. 2012). Briefly, an all-vs-all BLASTp search (using the parameters “-eval 1e-10 -max_target_seqs 5”) was first performed to identify gene pairs among species. Synteny blocks between 2 species were then calculated using the following parameters: “-k 50 -g -i -s 10 -e 1e-05 -u 10,000 -m 25 -b 2.” This approach identified collinear blocks that had at least 10 genes with an alignment significance <10^-5 in a maximum range of 10,000 nucleotides between genes. Results were visualized using SynVisio (Bandi and Gutwin 2020). The divergence time between 2 species were obtained from the TimeTree database (Kumar et al. 2017).

Identification of conserved genomic elements
For whole-genome alignment analysis, genome sequences and gene annotations of the previously selected 11 anemonefish species and A. polyacanthus were used. Repeat elements were identified using RepeatModeler v2.0.1 (Flynn et al. 2020) and RepeatMasker v4.1.1 (Tempel 2012) as described above and then soft-masked using BEDTools v2.30.0 (Quinlan 2014). Repeat-masked genome sequences and phylogenetic trees constructed with RAxML v8.2.9 (Stamatakis 2014) were used as input for whole genome alignment with Cactus multiple genome aligner v1.3.0 (Armstrong et al. 2020). Resulting HAL databases were converted to MAF format using hal2maf v2.1 (Hickey et al. 2013) with the A. ocellaris genome as a reference. MAFFILTER v1.3.1 (Dutheil et al. 2014) was then used to exclude repetitive regions and short alignments (<100 bp). RPHAST v1.6.11 (Hubisz et al. 2011) was used to identify conserved genomic elements in the A. ocellaris/A. percula branch from the alignment. Adjusted F-values of significant conservation for genomic elements were computed using the Benjamini and Hochberg method with the p.adjust function implemented in the R package stats v4.1.0 (R Core Team 2019). Genomic elements with an adjusted F-value <0.05 were considered as significantly conserved. Genes close to these genomic elements were identified with the closestBed function of BEDTools v2.30.0 (Quinlan 2014). Conserved elements were visualized using Circos v0.69-8 (Krzywinski et al. 2009). Expression values of genes close to conserved elements in the A. ocellaris/A. percula branch were visualized with a heatmap using the heatmap.2 function in gplots v3.1.1 (Warnes 2015).

Results and discussion
Chromosome-scale genome assembly of A. ocellaris
To construct high-quality chromosomes of A. ocellaris, we first generated 12,376,320 PacBio long-reads (average read length 10,239 bp) and 672,631,646 Illumina short reads (read length 151 bp) from brain tissue of an adult A. ocellaris individual (Supplementary Table 1). Prior to the de novo draft genome assembly, we investigated the global properties of the genome using IQ-TREE v2.1.3 (Minh et al. 2020), and GenomeScope v1.0 (Vurture et al. 2017). At k-mer = 17, the heterozygosity of A. ocellaris genome inferred from short reads was 0.26% and the estimated haploid genome size was 805,385,376 bp. The repetitive and nonrepetitive regions of the genome were estimated to be 343,219,574 bp (42.62%) and 462,165,802 bp (57.38%), respectively.

After the phased FALCON assembly with PacBio long reads (Chin et al. 2016), we obtained the primary (1,551 sequences, 861,420,186 bp, N50: 863,854 bp) and alternate (8,604 sequences, 679,345,988 bp, N50: 116,448 bp) haplotigs. To build the chromosome-scale assembly, 145,019,677 Hi-C read pairs (150 bp) were generated from liver tissue (Supplementary Table 1), and the Proximo scaffolding platform (Phase Genomics, WA, USA) was employed to orient de novo contigs into the chromosomes. This resulted in 353 sequences (865,612,980 bp) that consisted of 24 chromosome sequences (856,672,469 bp) and 329 short scaffolds that were not placed into chromosomes (8,940,511 bp). To improve the quality of the chromosome assembly, we performed iterative error-correction on the 24 chromosome sequences with Illumina short reads using Pilon v1.23 (Walker et al. 2014). At the 5th iterative run, 97.94% of the reads were aligned to the 24 chromosome sequences. Finally, we obtained 24 chromosomes, with a length ranging from 21,987,767 to 43,941,765 bp, totaling 856,612,077 bp (Fig. 1). Overall GC content of the A. ocellaris genome was 39.58%. The mean base-level coverage of the assembled chromosomes was 103.89×. Completeness of the genome assembly was assessed with BUSCO v4.1.4 (Simão et al. 2015) using the Actinopterygii lineage dataset (actinopterygii_odb10). The overall BUSCO score was 97.01% (complete and single-copy BUSCOs: 96.21%; complete and duplicated BUSCOs: 0.8%; fragmented BUSCOs: 0.52%; missing BUSCOs: 2.47%) (Table 1). Finally, we compared our chromosome-scale assembly with 3 other A. ocellaris draft genomes that have been previously published (Tan et al. 2018; Marcionetti et al. 2019) (Supplementary Table 2). In addition to large differences in size, which ranged from 744,831,443 to 880,704,246 bp, many misassemblies exist in these other A. ocellaris genomes. This is likely due to the limitations of the short-read sequencing technologies upon which these assemblies were constructed.

Prediction of A. ocellaris gene models
Repetitive elements in the A. ocellaris genome were examined by 2 approaches: (1) pattern matching using previously cataloged repetitive elements and (2) de novo. First the vertebrata repeat library from DFAM (Storer et al. 2021) was queried against the A. ocellaris genome sequences using RepeatMasker v4.1.1 (Tempel 2012). We then identified 2,301 de novo repetitive elements using RepeatModeler v2.0.1 (Flynn et al. 2020) and again searched for them in the A. ocellaris genome using RepeatMasker. A large fraction of the genome consisted of DNA transposons (24.11%), long-interspersed nuclear elements (7.67%), long-terminal repeats (LTRs, 3.77%), and rolling-circle transposons (1.55%) (Fig. 2; Supplementary Table 3). In total, 44.7% (382,912,159 bp) of the whole genome was identified as repetitive elements. This is similar to the repeat content estimated from the unassembled short reads (42.62%) using GenomeScope v1.0 (Vurture et al. 2017) as described above. It should be noted though, that the sum of occupied percentages in the genome per repeat group is larger than the actual percentage (44.7%) in the genome due to nested and overlapping repetitive elements (Fig. 2).
We then annotated the genome using BRAKER v2.1.6 (Brüna et al. 2021) with mRNA and protein evidence. This evidence consisted of mapped transcriptomic reads sequenced from 13 tissues (Supplementary Table 1), protein datasets from UniProtKB/Swiss-Prot (UniProt Consortium 2021), and the proteomes of 9 fish species. BRAKER predicted 26,433 gene models that were supported by either mRNA or protein hints, and 12,333 gene models with no evidence support. To account for the incompleteness of the evidence provided here and the gene annotation algorithm, we further added 364 nonsupport genes that have homology to the Swiss-Prot protein database and/or Pfam domains to the final gene models. This led to 26,797 final gene models from which 26,498 genes (98.88%) had significant homology to the NCBI nr database (bit-score ≥ 50) and 21,230 genes (79.23%) had at least 1 associated GO term. The completeness of our gene annotation was assessed using BUSCO v4.1.4 (Simão et al. 2015). We obtained 96.62% of completeness using the Actinopterygii-lineage dataset (complete and single-copy BUSCOs: 95.52%, complete and duplicated BUSCOs: 1.1%; fragmented BUSCOs: 1.04%, missing BUSCOs: 2.34%) (Table 1). This is higher than all other A. ocellaris and anemonefish gene annotations (Marcionetti et al. 2018, 2019; Tan et al. 2018; Lehmann et al. 2019), in both the overall completeness and duplicated ratio, thus suggesting that the genome we present here is currently the best anemonefish genome annotation. Furthermore, our gene models include the majority (93.97–97.63%) of gene models reported in previously published

![Amphiprion ocellaris](image)

**Fig. 1.** Chromosome architecture of the *Amphiprion ocellaris* genome. From the outermost layer inwards, each layer represents (1) chromosomes indicated by lines and ordered by size; (2) genic regions; (3) the number of repeats per 100 kb; (4) the PhyloP score calculated from the whole-genome alignment of anemonefishes with *Acanthochromis polyacanthus* as an outgroup species; and (5) tissue-specificity index (*t*) of each gene.
A. ocellaris genomes (Tan et al. 2018; Marcionetti et al. 2019). However, gene models from these studies include fewer gene models from this study (87.93–91.43%), again indicating that our gene models are the most comprehensive published to date (Supplementary Table 4).

**Assembly and annotation of mitochondrial genome**

We constructed the mitochondrial genome of A. ocellaris using Norgal v1.0.0 (Al-Nakeeb et al. 2017). This resulted in a 16,649 bp circular mitogenome, which has the same length as another previously sequenced A. ocellaris mitochondrial genome (NCBI accession number: NC_009065.1). These 2 mitochondrial genomes showed 99.83% sequence identity (16,621 of 16,649 bp) as calculated by BLASTn v2.10.0 (Altschul et al. 1990). MitoAnnotator v3.67 (Sato et al. 2018) was used to annotate the 37 organelle genes including 22 tRNA (Supplementary Fig 1).

**Analysis of gene expression patterns across tissues**

Gene expression levels of A. ocellaris genes were quantified using 13 tissue transcriptomes (Supplementary Table 1). We investigated tissue-specificity of gene expression levels using the tau (τ) index as it is the most robust metric for identifying tissuespecific genes (Kryuchkova-Mostacci and Robinson-Rechavi 2017). Given the range (0–1) of the τ index, we obtained 1,237 (4.62%) absolutely specific genes (τ = 1; genes expressed only in 1 tissue), 5,302 (19.79%) highly specific genes (0.85 ≤ τ < 1; genes highly expressed in a few tissues), and 3,431 (12.8%) housekeeping genes (τ ≤ 0.2; genes expressed in nearly all tissues without biased expression) as defined by the R package tispec v0.99 (Condron 2020). Tissue-specificity of gene expression showed a negative correlation (Pearson's correlation coefficient between τ and log10 maximum TPM value per gene = −0.46) with expression levels (Fig. 3a), which is consistent with previous observations that highly tissue-specific genes tend to have lower expression levels (Kryuchkova-Mostacci and Robinson-Rechavi 2017, Bentz et al. 2019).

Next, we checked gene expression patterns across tissues. After filtering for TPM ≥ 10, brain was the tissue with the highest number of expressed genes (n = 13,283) followed by optic lobe (n = 12,547) and eye (n = 11,809) (Fig. 3b). This high number of genes expressed in the brain has also been reported in other vertebrates (Lein et al. 2007; Hawrylycz et al. 2012; Bentz et al. 2019), and is most likely due to the complex role the brain has as the bodies control center. Furthermore, we observed that 1,957 genes were expressed in all 13 tissues sequenced here, and only 438 and 255 genes were exclusively expressed in the brain and optic lobe, respectively (Fig. 3b). Considering the high quality and similar numbers of transcriptomic reads per tissue generated in this study (Supplementary Table 1), we are confident that these results represent the most accurate transcriptomic atlas for A. ocellaris to date.

**Phylogenetic analysis**

Comparative analyses investigating the diversity and abundance of A. ocellaris gene families relative to other anemonefishes were performed using OrthoFinder v2.5.2 (Emms and Kelly 2019) with A. polyacanthus as an outgroup species (Supplementary Table 5). Overall, most sequences (96.7%) could be assigned to one of 29,111 orthogroups, with the remainder identified as “unassigned genes” with no clear orthologs (Supplementary Table 6). Fifty

---

**Table 1.** Statistics of the *Amphiprion ocellaris* chromosome-scale genome assembly and gene annotation.

| Chromosome assembly size | 24 sequences (856,612,077 bp) |
|--------------------------|-------------------------------|
| Non-ATGC characters      | 136,641 bp (0.02%)            |
| GC contents              | 39.58%                        |
| Mean base-level coverage | 103.89 x                      |
| Repeat contents          | 44.7%                         |
| BUSCO genome completeness| 3,531 (97.01%)                |
| Complete and single copy | 3,502 (96.21%)                |
| Complete and duplicated  | 29 (0.8%)                     |
| Fragmented               | 19 (0.52%)                    |
| Missing                   | 90 (2.47%)                    |
| Number of protein-coding genes | 26,797                   |
| BUSCO gene annotation completeness | 3,517 (96.62%)           |
| Complete and single copy | 3,477 (95.52%)                |
| Complete and duplicated  | 40 (1.1%)                     |
| Fragmented               | 38 (1.04%)                    |
| Missing                   | 85 (2.34%)                    |

---

**Fig. 2.** Repeat composition of the *Amphiprion ocellaris* genome. a) The number of occurrences per repeat group classified by the DFAM database. b) Occupied length in the genome per repeat group.
percentage of all proteins were in orthogroups consisting of 12 or more genes ($G_{120} = 12$) and were contained in the largest 10,672 orthogroups. Further, 15,899 orthogroups were shared amongst all the species examined here, and from these, 12,765 consisted entirely of single-copy genes (Supplementary Table 6).

Interestingly, all trees (Fig. 4; Supplementary Fig. 2) obtained here using maximum likelihood or Bayesian inference approaches had the same topology and shared some similarities with aspects of earlier work (Litsios and Salamin 2014; Litsios et al. 2014b; Marcionetti et al. 2019): a monophyletic *A. polymnus* and *A. sebae* group clustering with an Indian Ocean clade represented by *A. bicinctus* and *A. nigripes*, the skunk anemonefishes *A. akallopisos* and *A. perideraion*, an "ephippium complex" comprising *A. frenatus* and *A. melanopus*, and the monophyletic *A. ocellaris/A. percula* sister-species.

Our tree diverges most dramatically from previous analyses in that *P. biaculeatus* was not positioned within the *A. ocellaris/A. percula* clade but became the root of all other anemonefishes with 100% bootstrap support (Fig. 4; Supplementary Fig. 2). This topology has only been reported in 2 other studies that used mitochondrial genes to reconstruct anemonefish phylogenies (Santini and Polacco 2006; Nguyen et al. 2020). Pomacentrids (and anemonefishes in particular) have long been a challenge in systematics due to their high diversity and intraspecific variation (Tang et al. 2021), therefore discordances in our tree may also stem from insufficient information (i.e. only 11 out of the 28 described species were used). Specifically, the inclusion of *Amphiprion* latezonatus, the sister group of all *Amphiprion* except for the *A. ocellaris/A. percula* clade, could be essential to resolve the molecular phylogeny of anemonefishes (Santini and Polacco 2006; Litsios and Salamin 2014; Litsios et al. 2014b). This topology could also be the result of gene choice, as incongruences between trees based on mitochondrial and nuclear data has previously been observed (Litsios and Salamin 2014). Yet, here, we used an alignment matrix consisting

![Figure 3](image_url)  
**Fig. 3.** a) Tissue-specificity of *Amphiprion ocellaris* gene expression. The maximum TPM (transcripts per million) values across tissues and tissue-specificity index ($r$) was plotted in the 2D histogram. Trendline was fit using the linear model. b) Upset plot for the number of unique and shared genes expressed in different combinations of tissues. TPM values $>$10 were used as the threshold for gene expression in the specific tissue. Intersection size represents the number of expressed genes in the designated sets.

![Figure 4](image_url)  
**Fig. 4.** Phylogenetic reconstruction of the Amphiprioninae species tree using a maximum-likelihood approach. Numbers on each branching node are the bootstrap support (%) and the site concordance factor (%). These values were calculated only for nonoutgroup species using the IQ-TREE algorithm.
of more than 12,000 single-copy genes (182,497 parsimony informative sites and 2.8% gaps) and still obtained this topology (Fig. 4). This was further confirmed using BUSCO genes (Supplementary Fig. 2), predefined sets of reliable markers for phylogenetic inference (Waterhouse et al. 2018).

We also observed weak support values using site concordance factors (i.e. the percentage of sites supporting a specific branch over 1,000 randomly sampled quartets) in some branches. For example, a support value of 45.89% was recovered at the branching node of P. biauculeatus despite having 100% bootstrap support (Fig. 4; Supplementary Fig. 2), thus suggesting high uncertainty. Still, despite the incongruence observed here, our phylogenetic reconstructions are based on large-scale genomic evidence, whilst other studies have used only a few genes. Although we are confident that our trees have a good resolution and represent one of the most enriched phylogenies for anemonefishes in terms of supporting genomic loci and reduced stochastic error, we are nonetheless cautious in our interpretation of the phylogenetic delimitation of species presented here. Certainly, establishing a well-resolved phylogeny of anemonefish, particularly the early delimitation of species presented here. Certainly, establishing a well-resolved phylogeny of anemonefish, particularly the early divergent species (i.e. P. biauculeatus, the A. ocellaris/A. percula clade, and A. latzonatus), is critically important to understanding the evolution, genomic underpinning of their lifestyle (e.g. symbiosis with sea anemones, complex social structure) and fascinating biological features (e.g. pigmentation, sex change, aging).

**Whole-genome synteny of anemonefishes**

Syntenic blocks are often used to evaluate micro- and macroscale patterns of evolutionary conservation and divergence among related species. Identifying conserved gene order at the chromosomal level among species furthers our understanding of the molecular processes that led to the evolution of chromosome structure across species (Wang et al. 2012; Liu et al. 2018). Thus, here we used MCScanX (Wang et al. 2012) to investigate whole-genome synteny among all species present. Overall, synteny patterns were consistent with the phylogenetic tree, in that closely related species had a higher number of conserved blocks than distant species (Supplementary Table 7 and Supplementary Fig. 3), ultimately reflecting how gene gains or losses and sequence divergence increase proportionally with evolutionary time (Liu et al. 2018). Yet, since all species studied here are still closely related, shared synteny among species pairs is considerably high. As expected, synteny between A. ocellaris and A. percula was much higher than comparisons to other anemonefishes (Supplementary Table 7 and Fig. 3). This analysis identified 175 syntenic blocks of 19,872 genes (Supplementary Table 7) ranging from 11 to 1,010 gene pairs with 76.2% of these being collinear (i.e. conserved order).

Although studying pairwise collinear relationships among chromosomal regions allows for the elucidation of gene family evolution, the alignment of multiple regions is even more important as it can reveal complex chromosomal duplication and/or rearrangement relationships (Wang et al. 2012). Teleost fish genomes have been dynamically shaped by several forces (such as WGD and transpon activity). These in turn, led to various types of chromosome rearrangements either through differential loss of genes or formation of deletions, duplications, inversions, and translocations, which together contribute to reproductive isolation and therefore might promote the formation of a new species (Vollff 2005). Some chromosomal regions are translocated to new positions whereas others are inverses (Supplementary Fig 4). While the information shown here is merely an initial overview of the large-scale synteny of the false clownfish and other anemonefishes, it is still an important first step in obtaining evolutionary insights into the Amphiprioninae lineage.

**Lineage-specific conserved genomic elements in the A. ocellaris/A. percula branch**

Conserved genomic elements are relatively unchanged sequences across species. They are often parts of essential proteins or regulatory units and can be related to characteristics of specific lineages (Vollff 2005). To identify the signature of such conserved elements in the A. ocellaris genome, we first attempted to identify conserved elements in A. ocellaris but not in other anemonefish using the PHAST program (see Materials and Methods) (Hubisz et al. 2011). However, we were unable to identify genomic elements that were only conserved in this species, therefore we next sought to identify conserved elements shared by the 2 sister species of A. ocellaris and A. percula. We identified 91 conserved genome elements that showed significant conservation (adjusted P-value < 0.05).

To understand the possible role of these conserved elements, we investigated the function of 62 genes located around these elements (Supplementary Table 8). It is interesting to note that at least 21 out of these 62 genes could be involved in neurological functions. For example, the pcdh10 gene, encoding the protocadherin-10 protein, has been shown to be expressed in the olfactory and visual system of vertebrates as well as being involved in synapse and axon formation in the central nervous system (Mancini et al. 2020). Furthermore, a recent study also showed that mice lacking 1 copy of this gene have reduced social approach behavior (Schoch et al. 2017). Similarly, the asic2 gene is expressed in the central and peripheral nervous system of vertebrates and its encoded protein, acid-sensing ion channel 2, is vital for chemo- and mechno-sensing the environment (Cheng et al. 2018). The neuronal pentraxin 2 protein, encoded by the nptx2 gene, plays a role in the alteration of cellular activities for long-term neuroplasticity (Chapman et al. 2019). The tufa5 gene encodes a neurokine involved in behavior related to spatial memory in mice (Huang et al. 2021) and peripheral nociception in zebrafish (Jeong et al. 2020).

Additionally, these 62 genes also showed distinct expression patterns in A. ocellaris brain tissues (optic lobe and the rest of the brain) (Fig. 5). Mean TPM expression levels of these genes were 29.8 and 29.71 for the optic lobe and the other part of brain, respectively, whereas mean TPM for other tissues was 14.17, potentially indicating different, yet unknown roles of these genes in the brain. Although further investigation is required, this data suggest that neuronal genes located around specifically conserved elements in the A. ocellaris/A. percula branch could represent genomic signatures related to the distinct ecology of these 2 sister species. Certainly, anemonefish societies are highly species-specific (Litsios et al. 2012, 2014a). For example, while A. percula have a reduced range of movement, spending more time inside of their anemones and thus have a lower probability of social rank being usurped by outsiders, other species like Amphiprion clarkii have more opportunity for movement (due to their higher swimming abilities), increasing the probability of being taken over by outsiders so that the dominant individuals must display constant aggression to maintain control of their territory (Cleveland et al. 2011; Verde et al. 2015; Schmiege et al. 2017). Future research should endeavor to better characterize these differences and investigate whether they are linked to the genes we have identified here.
Conclusions

Here, we assembled the highly contiguous and complete chromosome-scale genome of the false clownfish *A. ocellaris* by de novo assembly using PacBio long reads and Hi-C chromatin conformation capture technologies. We annotated 26,797 protein-coding genes with 96.62% completeness of conserved actinopterygian genes, the highest level among anemonefish genomes available so far. We also identified tissue-specific gene expression patterns in *A. ocellaris*. Finally, we identified genomic elements conserved only in *A. ocellaris*/*A. percula*, which might underpin lineage-specific characteristics of these 2 species when compared to other anemonefishes. The high-quality of our genome and annotation will not only serve as a resource to better understand the genomic architecture of anemonefishes, but it will further strengthen the false clownfish as an emerging model organism for molecular, ecological, developmental, and environmental studies of reef fishes.

Data availability

The genomic and transcriptomic sequencing reads generated in this study have been deposited in NCBI GenBank database under the BioProject ID PRJNA787397. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JAJUWX000000000. The genome annotation is available in Dryad repository (https://datadryad.org/stash/share/UpzvIvKZQj21CcO38uwnPMgF1tQnPMMLqX_S_QpSoE).

Supplemental material is available at G3 online.

---

**Fig. 5.** Gene expression levels of the 62 *Amphirion ocellaris* genes nearest to the conserved elements in the *Amphirion ocellaris*/*Amphirion percula* branch across 13 tissues are shown. Color key indicates log2-transformed TPM values.
Acknowledgments

We thank Mr Hidenori Kinjo for collecting the A. ocellaris samples used in this study and Dr Konstantin Khalturin (OIST) for his help with the phylogenetic analysis. We also thank Mr Lilian Carlu for the anemonefish drawings shown in Figs. 1 and 4.

Funding

Research reported in this publication was supported by funding from the Okinawa Institute of Science and Technology Graduate University.

Conflicts of interest

None declared.

Literature cited

Al-Nakeeb K, Petersen TN, Sicheritz-Pontén T. Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. BMC Bioinformatics. 2017;18(1):1–7.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410.

Arms特朗 strm Jr, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587(7833):246–251.

Bandi V, Gutwin C. Interactive exploration of genomic conservation. In: Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020, Waterloo, Canada; 2020.

Benzt AB, Thomas GW, Rusch DB, Rosvall KA. Tissue-specific expression profiles and positive selection analysis in the tree swallow (Tachycineta bicolor) using a de novo transcriptome assembly. Sci Rep. 2019;9(1):1–12.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49(4):643–650.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–2120.

Bámia T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3(1):lgaa108.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1914–2120.

Chapman G, Shanmugalingam U, Smith PD. The role of neuronal function proteins for chemo-sensing and mechano-sensing. J Biomed Sci. 2018;25(1):14.

Chin C-S, Peluso P, Sedláček FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O’Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13(12):1050–1054.

Cleveland A, Verde EA, Lee RW. Nutritional exchange in a tropical tripartite symbiosis: direct evidence for the transfer of nutrients from anemonefish to host anemone and zooxanthellae. Mar Biol. 2011;158(3):589–602.

Condon K. tispec: Calculates Tissue Specificity from RNA-Seq Data; 2020 (https://github.com/roonysgalbi/tispec).

Conway JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. An atomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391–399.

Hawyrycz MJ, Lein ES, Guillotet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. An atomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391–399.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical model for multiple genome alignment. Bioinformatics. 2013;29(10):1341–1342.

Huang S, Zheng C, Xie G, Song Z, Wang P, Bai Y, Chen D, Zhang Y, Lv P, Liang W, et al. FAM19AS1/TAFA5, a novel neurokinin, plays a crucial role in depressive-like and spatial memory-related behaviors in mice. Mol Psychiatry. 2021;26(6):2363–2379.

Hubsz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12(1):41–51.

Jeong I, Yun S, Shahapal A, Cho EB, Hwang SW, Seong JY, Park HC. FAM19AS1 affects mustard oil-induced peripheral nociception in zebrafish. bioRxiv. 2020. https://doi.org/10.1101/2020.08.11.245738

Kath K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–915.

Kronenberg ZN, Hall RJ, Hiendleder S, Smith TP, Sullivan ST, et al. Hi-C data for phased diploid genomes. bioRxiv 327064; 2018. https://doi.org/10.1101/2018.06.01.327064

Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18(2):205–214.

Krzynski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–1645.

Kuck P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 2014;11(1):81.
Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34(7):1812–1819.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–359.

Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 2013;62(4):611–615.

Lehmann R, Lightfoot DJ, Schunter C, Micheli CT, Ohyanagi H, Mineta K, Fere T, Berumen ML, Miller DJ, Aranda M, et al. Finding Nemo’s Genes: a chromosome-scale reference assembly of the genome of the orange clownfish Amphiprion percula. Mol Ecol Resour. 2019;19(3):570–585.

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445(7124):168–176.

Letunic I, Bork P. Interactive Tree Of Life (iTOl) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49(W1):W293–W296.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint, arXiv:1303.3997; 2013. Doi: 10.6084/M9.FIGSHARE.963153.V1.

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

Litsios G, Kostikova A, Salamin N. Host specialist clownfishes are environmental niche generalists. Proc R Soc B. 2014a;281(1795):20133220.

Litsios G, Pearman PB, Lanterbecq D, Tolou N, Salamin N. The radiation of the clownfishes has two geographical replicates. J Biogeogr. 2014b;41(11):2140–2149.

Litsios G, Salamin N. Hybridisation and diversification in the adaptive radiation of clownfishes. BMC Evol Biol. 2014;14:245–249.

Litsios G, Sims CA, Wuest RO, Pearman PB, Zimmermann NE, Salamin N. Mutualism with sea anemones triggered the adaptive radiation of clownfishes. BMC Evol Biol. 2012;12:212.

Liu D, Hunt M, Tsai LJ. Inferring synteny between genome assemblies: a systematic evaluation. Bioinformatics. 2018;19(1):1–13.

Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614.

Mancini M, Bassani S, Passafaro M. Right place at the right time: how changes in protocadherins affect synaptic connections contributing to the etiology of neurodevelopmental disorders. Cells. 2020;9(12):2711.

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

Marcionetti A, Rossi V, Bertrand JAM, Litsios G, Salamin N. First draft genome of an iconic clownfish species (Amphiprion frenatus). Mol Ecol Resour. 2018;18(5):1092–1101.

Marcionetti A, Rossi V, Roux N, Salis P, Laudet V, Salamin N. Insights into the genomes of clownfish adaptive radiation: genetic basis of the mutualism with sea anemones. Genome Biol Evol. 2019;11(3):869–882.

Mikheenko A, Prijibelski A, Savilev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–i150.

Militz TA, Foale S. The “Nemo Effect”: perception and reality of Finding Nemo’s impact on marine aquarium fisheries. Fish Fish. 2017;18(3):596–606.
Tempel S. Using and understanding RepeatMasker. Methods Mol Biol. 2012;859:29–51.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–D489.

Verde EA, Cleveland A, Lee RW. Nutritional exchange in a tropical tripartite symbiosis II: direct evidence for the transfer of nutrients from host anemone and zooxanthellae to anemonefish. Mar Biol. 2015;162(12):2409–2429.

Volff J. Genome evolution and biodiversity in teleost fish. Heredity (Edinb). 2005;94(3):280–294.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202–2204.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11) e112963.

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.

Warnes GR. gplots: Various R Programming Tools for Plotting Data; 2015.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35(3):543–548.

Wickham H. Elegant graphics for data analysis (ggplot2). Media. 2009;35:10–1007.

Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–848.

Communicating editor: A. McCallion