A chromosome-scale genome assembly of the false clownfish,

Amphiprion ocellaris

Taewoo Ryu^{1,a,*}, Marcela Herrera^{2,a}, Billy Moore¹, Michael Izumiyama¹, Erina Kawai¹,

Vincent Laudet^{2,3}, and Timothy Ravasi^{1,4,*}

¹Marine Climate Change Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa, 904-0495 Japan.

²Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa, 904-0495 Japan.

³Marine Research Station, Institute of Cellular and Organismic Biology, Academia Sinica, I¹ Lan, Taiwan.

⁴Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia.

^aThese authors contributed equally.

* Correspondence should be addressed: taewoo.ryu@oist.jp; timothy.ravasi@oist.jp

ABSTRACT

Background: 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*. **Results:** 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. Transcriptomic 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. **Conclusions:** 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

DATA DESCRIPTION

Context

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 [1]. Yet, together with its sister species, the orange clownfish *Amphiprion percula*, it is one of the most recognizable fish, especially among the non-scientific community, following the Disney movie “Finding Nemo” [2]. 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 [3,4]. 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 [1].

As for all anemonefishes, *A. ocellaris* is a protandrous hermaphrodite (i.e., male-to-female sex change) that lives in association with sea anemones [5–7]. It is a mutualistic relationship in which the host sea anemone provides food and shelter from predators [6,7], whilst the fish provides supplemental nutrition [8] and increased oxygen uptake by modulating water flow among the anemone’s tentacles [9,10]. The false clownfish is a generalist and can colonize up to four hosts: the magnificent sea anemone (*Heteractis magnifica*), the leathery sea anemone (*Heteractis crispa*), the giant carpet anemone (*Stichodactyla gigantea*), and Mertens' carpet sea anemone (*Stichodactyla. mertensii*) [7]. Social groups typically consist of an adult breeding pair and several smaller (immature) juveniles ranked by size [11]. Within this social group a large dominant female is followed by a sub-dominant male, so that if the female is removed, the male changes sex and the largest non-breeder matures into a breeding male (sequential hermaphroditism) [12]. Within the Amphiprioninae subfamily, the two clownfish, *A. ocellaris* and *A. percula* form a separate clade that diverged early in the phylogeny, thus

making them attractive for studies of the evolutionary history and adaptive radiation of anemonefishes [13,14]. Both species share behavioral and ecological traits and have similar body coloration patterns (i.e., bright orange with three vertical white bars) that set them apart from other anemonefishes. The major difference between these two species of clownfish is their differing geographic distributions, with *A. percula* restricted to the Great Barrier Reef, Papua New Guinea, and the Solomon Islands, whilst *A. ocellaris* can be found across the Indo-Malaysian region, from the Ryukyu Islands of Japan, throughout South East Asia, and south to north western Australia [7,13,15–18].

Until now, genome assemblies of at least ten anemonefish species including *A. ocellaris* have been published [19–22]. 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 mis-assemblies. However, third-generation sequencing platforms such as Pacific Biosciences, produce longer reads (5 to 60 kb) that enhance the continuous assembly of genome sequences [23,24]. 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 [25]. For example, 29 % of N-gaps in the human reference genome (GRCh38) could be filled with PacBio long reads [26]. 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-fold increase in scaffold N50 (401.72 kb), and a 16% improvement in genome completeness [22]. 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) [19].

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. Moreover, transcriptomic analysis detected 1,237 genes with absolute tissue-specific expression. A further 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. Access to a high-quality chromosome-scale genome for *A. ocellaris* will not only unlock the potential to gain new insight into the evolutionary processes underlying the radiation of anemonefishes, but it will also facilitate our understanding of how the genomic architecture of this diverse group of reef fishes evolved. 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.

METHODS

Specimen collection and nucleic acid sequencing

Three adult *A. ocellaris* clownfish (one female and two males) were collected from 5 m depth in Motobu, Okinawa (26° 71' 29.83" N, 127° 91' 57.51" E) on March 25, 2020. Fish were kept under natural conditions at the OIST Marine Science Station in a 270 L (60 × 90 × 50 cm) tank until May 19, 2020. 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 two different platforms: PacBio Sequel II and Illumina NovaSeq6000 (Table S1). 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 hr 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 (Table S1). 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 minute, 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 two individuals (one male and one 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 integrity number (RIN) > 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 [27] with *k*-mer = 17 and default parameters, and GenomeScope v1.0 [28] with default parameters. Quality-trimmed Illumina short reads obtained from Trimmomatic v0.39 [29] using the parameter set 'ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:keepBothReads LEADING:3 TRAILING:3 MINLEN:36' were used as input for Jellyfish. Genomic contigs were assembled using the FALCON software version as of September 28, 2020. For chromosome-scale assembly,

initial contigs obtained from FALCON-phase were scaffolded with Phase Genomics' Proximo algorithm based on Hi-C chromatin contact maps. In brief, the processed Hi-C sequencing reads were aligned to the Falcon assembly with BWA-MEM [30] using the -5 SP and -t 8 options. PCR duplicates were flagged with SAMBLASTER v0.1.26 [31] and subsequently removed from all following analyses. Non-primary and secondary alignments were filtered using SAMtools v1.10 [32] with the -F 2304 flag. FALCON-Phase [33] was then used to correct phase switching errors in the scaffolds obtained from FALCON-Unzip [34].

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 [35]. A total of 40,000 individual Proximo runs were performed to optimize chromosome construction. Juicebox v1.13.01 [36] 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 Pilon v1.23 [37]. Quality-trimmed Illumina short reads obtained from Trimmomatic v0.39 [29] using the same parameter set described above were aligned to the proximo-assembled chromosome-scale genome with Bowtie2 v2.4.1 [38] using the default settings. SAM files were converted to BAM files with SAMtools v1.10 [32] and then used as input for Pilon. Error correction was completed using five iterations of Pilon.

To calculate the overall mean genome-wide base level coverage, PacBio reads were aligned to the assembled chromosome sequences using Pbbmm2 v1.4.0

(<https://github.com/PacificBiosciences/pbbmm2>). Per-base coverage of aligned reads across

entire chromosomal sequences was obtained using the BEDTools v2.30.0 [39] genomeCoverageBed function. Finally, we compared the quality of our genome assembly to three other published *A. ocellaris* genome sequences [21,22] using Quast v5.0.2 [40].

Prediction of gene models in *A. ocellaris*

Repetitive elements in the *A. ocellaris* genome were identified *de novo* using RepeatModeler v2.0.1 [41] with the parameter -LTRStruct. RepeatMasker v4.1.1 [42] was then used to screen known repetitive elements with two separate inputs: the RepeatModeler output and the vertebrata library of Dfam v3.3 [43]. The two output files were validated, merged, and redundancy was removed using GenomeTools v1.6.1 [44].

BRAKER v2.1.6 [45] was then used to annotate candidate gene models of *A. ocellaris*. For mRNA evidence for gene annotation, transcriptomic reads (Table S1) were trimmed with Trimmomatic v0.39 [29] using the parameter set mentioned above and mapped to the chromosome sequences with HISAT2 v2.2.1 [46] using the ‘-dta’ option. SAM files were then converted to BAM format using SAMtools v1.10 [32]. For protein evidence, manually annotated and reviewed protein records from UniProtKB/Swiss-Prot (UniProt Consortium, 2021) as of January 11, 2021 (563,972 sequences) in addition to the proteomes of the false clownfish (*A. ocellaris*: 48,668), zebrafish (*Danio rerio*: 88,631), spiny chromis damselfish (*Acanthochromis polyacanthus*: 36,648), Nile tilapia (*Oreochromis niloticus*: 63,760), Japanese rice fish (*Oryzias latipes*: 47,623), rainbow fish (*Poecilia reticulata*: 45,692), bicolor damselfish (*Stegastes partitus*: 31,760), tiger puffer (*Takifugu rubripes*: 49,529), and Atlantic salmon (*Salmo salar*: 112,302) from the NCBI protein database (<https://www.ncbi.nlm.nih.gov/protein>) were used. Only gene models with evidence support (mRNA or protein hints) or with homology to the Swiss-Prot protein database [47] or Pfam

domains [48] identified by Diamond v2.0.9 [49] and InterProScan v5.48.83.0 [50], respectively, were added to the final gene models. Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.4 [51] with the Actinopterygii-lineage dataset (actinopterygii_odb10) was used for quality assessment of gene annotation. Finally, for functional annotation of predicted gene models, NCBI BLAST v2.10.0 [52] was used with the NCBI non-redundant protein database (*nr*) as the target database. Gene Ontology (GO) terms were assigned to *A. ocellaris* genes using the ‘gene2go.gz’ and ‘gene2accession.gz’ files downloaded from the NCBI ftp site (<https://ftp.ncbi.nlm.nih.gov/gene/DATA/>) and the BLAST output.

Assembly and annotation of the mitochondrial genome

The mitochondrial genome of *A. ocellaris* was assembled using Norgal v1.0.0 [53] with quality trimmed Illumina genomic reads. MitoAnnotator v3.67 [54] was used to annotate the organelle genes. Annotated genes in this study were compared to previously published *A. ocellaris* genes using BLASTn v2.10.0 [52] 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 Trimmomatic v0.39 [29] using the parameter set mentioned above and mapped to the genome using HISAT2 v2.2.1 [46]. SAM files were then converted to BAM files using SAMtools v1.10 [32]. Expression levels were quantified and TPM (transcripts per million) was normalized with StringTie v2.1.4 [55]. Tissue-specify index (τ) was calculated for each gene using the R package tispec v0.99 [56], with the relationship between τ and TPM expression values visualized on a two dimensional

histogram with ggplot2 v3.3.5 [57]. TPM expression values per tissue were visualized in an Upset plot with the UpSetR v1.4.0 package [58].

Gene orthology and phylogenetic analyses

To identify evolutionary relationships between *A. ocellaris* and other Amphiprioninae species, two species combinations were used: (1) a dataset that includes 11 anemonefish proteomes, i.e. our *A. ocellaris* proteome and ten other anemonefishes (*Amphiprion akallopisos*, *Amphiprion bicinctus*, *Amphiprion frenatus*, *Amphiprion melanopus*, *Amphiprion nigripes*, *Amphiprion percula*, *Amphiprion perideraion*, *Amphiprion polymnus*, *Amphiprion sebae*, and *Premnas biaculeatus*) [19–21], and *A. polyacanthus* as a single outgroup species, and (2) a dataset comprised of all 11 anemonefishes, *A. polyacanthus*, and five additional outgroup species across the teleost phylogenetic tree: zebrafish (*D. rerio*), bicolor damselfish (*S. partitus*), Asian seabass (*Lates calcarifer*), Nile tilapia (*O. niloticus*), and southern platyfish (*Xiphophorus maculatus*). The proteomes of outgroup species were obtained as previously described [19]. 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 [59]. 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 multi-alignments of one-to-one orthologs. In brief, sequences of single-copy orthologs present in all species were first aligned using MAFFT v7.130 [60] using the options ‘-localpair -maxiterate 1,000 -leavegappyregion’, then trimmed with trimAl v1.2 [61] using the ‘-gappyout’ flag, and finally concatenated with FASconCAT-G [62].

Phylogenetic trees were first constructed based on maximum-likelihood criteria using the two datasets described above. The MPI version of RAxML v8.2.9 (raxmlHPC-MPI-AVX) [63] 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 [51] 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 [64]. 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 re-rooted using iTOL v6.4 [65]. 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 [66].

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 [67]. Briefly, an all-versus-all BLASTp search (using the parameters ‘-evaluate 10^{-10} -max_target_seqs 5’) was first performed to identify gene pairs among species. Synteny blocks between two species were then calculated using the following parameters: ‘-k 50 -g -1 -s 10 -e 1e-05 -u 10,000 -m 25 -b 2’. This approach identified collinear blocks that had at least ten genes with an alignment significance < 10^{-5} in a maximum range of 10,000 nucleotides between genes. Results were visualized using

SynVisio [68]. The divergence time between two species were obtained from the TimeTree database [69].

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 [41] and RepeatMasker v4.1.1 [42] as described above and then soft-masked using BEDTools v2.30.0 [39]. Repeat-masked genome sequences and phylogenetic trees constructed with RAxML v8.2.9 [63] were used as input for whole genome alignment with Cactus multiple genome aligner v1.3.0 [70]. Resulting HAL databases were converted to MAF format using hal2maf v2.1 [71] with the *A. ocellaris* genome as a reference. MAFFILTER v1.3.1 [72] was then used to exclude repetitive regions and short alignments (< 100 bp). RPHAST v1.6.11 [73] was used to identify conserved genomic elements in the *A. ocellaris/A. percula* branch from the alignment. Adjusted p-values of significant conservation for genomic elements were computed using the Benjamini & Hochberg method with the p.adjust function implemented in the R package stats v4.1.0 [74]. Genomic elements with an adjusted p-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 [39]. Conserved elements were visualized using Circos v0.69-8 [75]. 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 [76].

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 (Table S1). Prior to the *de novo* draft genome assembly, we investigated the global properties of the genome with Illumina short reads using Jellyfish v2.3.0 [27] and GenomeScope v1.0 [28]. 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 non-repetitive 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 [34], we obtained the primary (1,551 sequences, 861,420,186 bp, N50: 863,854bp) 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 (Table S1), 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 [37]. 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 bp to 43,941,765 bp, totaling 856,612,077 bp (Figure 1). Overall GC content of the *A. ocellaris* genome was 39.58 %. The mean base-level coverage

of the assembled chromosomes was $103.89\times$. Completeness of the genome assembly was assessed with BUSCO v4.1.4 [51] 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 to three other *A. ocellaris* draft genomes that have been previously published [21,22] (Table S2). In addition to large differences in size, which ranged from 744,831,443 to 880,704,246 bp, many misassembly events such as relocations, translocations, and inversions were observed 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 two approaches: (1) pattern matching using previously cataloged repetitive elements and (2) *de novo*. First the vertebrata repeat library from DFAM [43] was queried against the *A. ocellaris* genome sequences using RepeatMasker v4.1.1 [42]. We then identified 2,301 *de novo* repetitive elements using RepeatModeler v2.0.1 [41] 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 (LINEs, 7.67 %), long terminal repeats (LTRs, 3.77 %), and rolling-circle transposons (RCs, 1.55 %) (Figure 2 and Table S3). 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 [28] 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 (Figure 2).

We then annotated the genome using BRAKER v2.1.6 [45] with mRNA and protein evidence. This evidence consisted of mapped transcriptomic reads sequenced from 13 tissues (Table S1), protein datasets from UniProtKB/Swiss-Prot [47], 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 non-support 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 one associated GO term. The completeness of our gene annotation was assessed using BUSCO v4.1.4 [51]. 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 [19–22], 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 *A. ocellaris* genomes [21,22]. 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 (Table S4).

Assembly and annotation of mitochondrial genome

We constructed the mitochondrial genome of *A. ocellaris* using Norgal v1.0.0 [53]. 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 two mitochondrial genomes showed 99.83 % sequence identity (16,621 of 16,649 bp) as calculated by BLASTn v2.10.0 [52]. MitoAnnotator v3.67 [54] was used to annotate the 37 organelle genes including 22 tRNA (Figure S1).

Analysis of gene expression patterns across tissues

Gene expression levels of *A. ocellaris* genes were quantified using 13 tissue transcriptomes (Table S1). We investigated tissue-specificity of gene expression levels using the tau (τ) index as it is the most robust metric for identifying tissue-specific genes [77]. Given the range (0 ~ 1) of the τ index, we obtained 1,237 (4.62 %) absolutely specific genes ($\tau = 1$; genes expressed only in one tissue), 5,302 (19.79 %) highly specific genes ($0.85 \leq \tau < 1$; genes highly expressed in a few tissues), and 3,431 (12.8 %) housekeeping genes ($\tau \leq 0.2$; genes expressed in nearly all tissues without biased expression) as defined by the R package *tispec* v0.99 [56]. Tissue-specificity of gene expression showed a negative correlation (Pearson's correlation coefficient between τ and $\log_{10}$ maximum TPM value per gene = -0.46) with expression levels (Figure 3A), which is consistent with previous observations that highly tissue-specific genes tend to have lower expression levels [77,78].

Next, we checked gene expression patterns across tissues. After filtering for $\text{TPM} \geq 10$, brain was the tissue with the highest number of expressed genes ($n = 13,283$) followed by optic lobe ($n = 12,547$) and eye ($n = 11,809$) (Figure 3B). This high number of genes expressed in the brain has also been reported in other vertebrates [78–80], 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 (Figure 3B). Considering the high quality and similar numbers of transcriptomic reads per tissue generated in this study (Table S1), 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 [59] with *A. polyacanthus* as an outgroup species (Table S5). 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 (Table S6). Fifty percent of all proteins were in orthogroups consisting of 12 or more genes ($G_{50} = 12$) 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 (Table S6). Interestingly, all trees (Figure 4 and Figure S2) obtained here using maximum likelihood or Bayesian inference approaches had the same topology and shared some similarities with aspects of earlier work [13,14,21]: 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 (Figure 4 and Figure S2). This topology has only been reported in two other studies that used mitochondrial genes to reconstruct anemonefish phylogenies [81,82]. Pomacentrids (and anemonefishes in particular) have long been a challenge in systematics due to their high diversity and intraspecific variation [83], 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 [13,14,82]. This topology could also be the result of gene choice, as incongruences between trees based on mitochondrial and nuclear data has previously been observed [14]. Yet, here, we used an alignment matrix consisting of more than twelve thousand single-copy genes (182,497 parsimony informative sites and 2.8% gaps) and still obtained this topology (Figure 4). This was further confirmed using BUSCO genes (Figure S2), predefined sets of reliable markers for phylogenetic inference [84].

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. biaculeatus* despite having 100 % bootstrap support (Figure 4 and Figure S2), 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 divergent species (i.e., *P. biaculeatus*, the *A. ocellaris*/*A. percula* clade, and *A. latezonatus*), 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 macro-scale 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 [67,85]. Thus, here we used MCScanX [67] 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 (Table S7 and Figure S3), ultimately reflecting how gene gains or losses and sequence divergence increase proportionally with evolutionary time [85]. 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 (Table S7 and Figure S3). This analysis identified 175 syntenic blocks of 19,872 genes (Table S7) 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 [67]. Teleost fish genomes have been dynamically shaped by several forces (such as WGD and transposon 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 [86]. Some chromosomal regions are translocated to new positions whereas others are inversed (Figure S4). 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 [86]. 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 PFAST program (see Methods) [73]. 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 two 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 (Table S8). 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 [87]. Furthermore, a recent study also showed that mice lacking one copy of this gene have reduced social approach behavior [88]. 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 mechano-sensing the environment [89]. The neuronal pentraxin 2 protein, encoded by the *nptx2* gene, plays a role in the alteration of cellular activities for long-term neuroplasticity [90]. The *tafa5* gene encodes a neurokinine involved in behavior related to spatial memory in mice [91] and peripheral nociception in zebrafish [92].

Additionally, these 62 genes also showed distinct expression patterns in *A. ocellaris* brain tissues (optic lobe and the rest of the brain) (Figure 5). Mean TPM expression level 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 suggests 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 two sister species. Certainly, anemonefish societies are highly species-specific [17,93]. 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 [94–96]. 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 two 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 STATEMENT

The sequencing reads generated in this study have been deposited in GenBank database under the BioProject ID PRJNA787397. The genome assembly and annotation are available in Dryad repository (https://datadryad.org/stash/share/UpzvIVKZOj21CcO38uwnPMgF1_ONpMM_LqX_S_0pSoE).

COMPETING INTERESTS

Authors declared no competing interests.

FUNDING

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

AUTHORS' CONTRIBUTIONS

TR1 (Timothy Ravasi) and VL conceived the genome sequencing project. TR2 (Taewoo Ryu) designed the overall analysis. EK collected fish samples. BM and MI dissected tissues for sequencing. TR2 performed the overall computational analysis. MH performed the orthology, phylogenetic, and synteny analyses. TR2 and MH wrote the manuscript with help from all authors. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

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 Figures 1 and 4.

REFERENCES

1. Roux N, Salis P, Lee S-H, Besseau L, Laudet V. Anemonefish, a model for Eco-Evo-Devo. *EvoDevo*. 2020; doi: 10.1186/s13227-020-00166-7.
2. Militz TA, Foale S. The “Nemo Effect”: perception and reality of Finding Nemo’s impact on marine aquarium fisheries. *Fish and Fisheries*. Wiley Online Library; 18:596–6062017;
3. Militz TA, Foale S, Kinch J, Southgate PC. Natural rarity places clownfish colour morphs at risk of targeted and opportunistic exploitation in a marine aquarium fishery. *Aquatic Living Resources*. EDP Sciences; 31:182018;
4. Rhyne AL, Tlusty MF, Szczebak JT, Holmberg RJ. Expanding our understanding of the trade in marine aquarium animals. *PeerJ*. PeerJ Inc.; 5:e29492017;
5. Da Silva KB, Nedosyko A. Sea anemones and anemonefish: a match made in heaven. *The Cnidaria, past, present and future*. Springer; p. 425–38.

6. Fautin DG. Review article The Anemonefish Symbiosis: What is Known and What is Not. *Symbiosis*. Balaban Publishers; 1991;
7. Fautin DG, Allen GR. Anemone fishes and their host sea anemones: a guide for aquarists and divers. Sea Challengers;
8. Holbrook SJ, Schmitt RJ. Growth, reproduction and survival of a tropical sea anemone (Actiniaria): benefits of hosting anemonefish. *Coral reefs*. Springer; 24:67–732005;
9. Herbert N, Bröhl S, Springer K, Kunzmann A. Clownfish in hypoxic anemones replenish host O₂ at only localised scales. *Scientific reports*. Nature Publishing Group; 7:1–102017;
10. Szczebak JT, Henry RP, Al-Horani FA, Chadwick NE. Anemonefish oxygenate their anemone hosts at night. *Journal of Experimental Biology*. Company of Biologists; 216:970–62013;
11. Buston P. Mortality is associated with social rank in the clown anemonefish (*Amphiprion percula*). *Marine Biology*. Springer; 143:811–52003;
12. Casas L, Saborido-Rey F, Ryu T, Michell C, Ravasi T, Irigoien X. Sex change in clownfish: molecular insights from transcriptome analysis. *Scientific reports*. Nature Publishing Group; 6:1–192016;
13. Litsios G, Pearman PB, Lanterbecq D, Tolou N, Salamin N. The radiation of the clownfishes has two geographical replicates. *Journal of Biogeography*. John Wiley & Sons, Ltd; 2014; doi: 10.1111/jbi.12370.
14. Litsios G, Salamin N. Hybridisation and diversification in the adaptive radiation of clownfishes. *BMC Evolutionary Biology*. BioMed Central; 14:1–92014;
15. Camp EF, Hobbs J-PA, De Brauwier M, Dumbrell AJ, Smith DJ. Cohabitation promotes high diversity of clownfishes in the Coral Triangle. *Proceedings of the Royal Society B: Biological Sciences*. The Royal Society; 283:201602772016;
16. Fautin DG, Allen GR, Allen GR, Naturalist A, Allen GR, Naturaliste A. Field guide to anemonefishes and their host sea anemones. Western Australian Museum Perth; 1992;
17. Litsios G, Sims CA, Wüest RO, Pearman PB, Zimmermann NE, Salamin N. Mutualism with sea anemones triggered the adaptive radiation of clownfishes. *BMC Evolutionary Biology*. 2012; doi: 10.1186/1471-2148-12-212.
18. Timm J, Figiel M, Kochzius M. Contrasting patterns in species boundaries and evolution of anemonefishes (Amphiprioninae, Pomacentridae) in the centre of marine biodiversity. *Molecular Phylogenetics and Evolution*. Elsevier; 49:268–762008;
19. Lehmann R, Lightfoot DJ, Schunter C, Michell CT, Ohyanagi H, Mineta K, et al.. Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish *Amphiprion percula*. *Molecular Ecology Resources*. John Wiley & Sons, Ltd; 2019; doi: 10.1111/1755-0998.12939.

20. Marcionetti A, Rossier V, Bertrand JAM, Litsios G, Salamin N. First draft genome of an iconic clownfish species (*Amphiprion frenatus*). *Molecular Ecology Resources*. John Wiley & Sons, Ltd; 2018; doi: 10.1111/1755-0998.12772.
21. Marcionetti A, Rossier V, Roux N, Salis P, Laudet V, Salamin N. Insights into the genomics of clownfish adaptive radiation: genetic basis of the mutualism with sea anemones. *Genome biology and evolution*. Oxford University Press; 11:869–822019;
22. Tan MH, Austin CM, Hammer MP, Lee YP, Croft LJ, Gan HM. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly. *GigaScience*. Oxford University Press; 7:gix1372018;
23. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*. Nature Publishing Group; 21:597–6142020;
24. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. *Trends in Genetics*. Elsevier; 34:666–812018;
25. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al.. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. Nature Publishing Group; 592:737–462021;
26. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al.. Long-read sequencing and de novo assembly of a Chinese genome. *Nature communications*. Nature Publishing Group; 7:1–102016;
27. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. Oxford University Press; 27:764–702011;
28. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. Oxford University Press; 33:2202–42017;
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. Oxford University Press; 30:2114–202014;
30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*. 2013;
31. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. Oxford University Press; 30:2503–52014;
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.. The sequence alignment/map format and SAMtools. *Bioinformatics*. Oxford University Press; 25:2078–92009;
33. Kronenberg ZN, Hall RJ, Hiendleder S, Smith TP, Sullivan ST, Williams JL, et al.. FALCON-Phase: integrating PacBio and Hi-C data for phased diploid genomes. *BioRxiv*. Cold Spring Harbor Laboratory; :327064 2018;

34. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*. Nature Publishing Group; 13:1050–42016;
35. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al.. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature genetics*. Nature Publishing Group; 49:643–502017;
36. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al.. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems*. Elsevier; 3:99–1012016;
37. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al.. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*. Public Library of Science San Francisco, USA; 9:e1129632014;
38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; doi: 10.1038/nmeth.1923.
39. Quinlan AR. BEDTools: the Swiss army tool for genome feature analysis. *Current protocols in bioinformatics*. Wiley Online Library; 47:11–22014;
40. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*. Oxford University Press; 34:i142–502018;
41. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al.. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*. National Acad Sciences; 117:9451–72020;
42. Tempel S. Using and understanding RepeatMasker. *Mobile genetic elements*. Springer; p. 29–51.
43. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*. BioMed Central; 12:1–142021;
44. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM transactions on computational biology and bioinformatics*. IEEE; 10:645–562013;
45. Brůna 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 genomics and bioinformatics*. Oxford University Press; 3:lqaa1082021;
46. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*. Nature Publishing Group; 37:907–152019;
47. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. Oxford University Press; 49:D480–92021;

48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al.. Pfam: The protein families database in 2021. *Nucleic Acids Research*. Oxford University Press; 49:D412–92021;
49. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature methods*. Nature Publishing Group; 12:59–602015;
50. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. Oxford University Press; 17:847–82001;
51. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. Oxford University Press; 31:3210–22015;
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. Elsevier; 215:403–101990;
53. Al-Nakeeb K, Petersen TN, Sicheritz-Pontén T. Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC bioinformatics*. Springer; 18:1–72017;
54. Sato Y, Miya M, Fukunaga T, Sado T, Iwasaki W. MitoFish and MiFish pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular biology and evolution*. Oxford University Press; 35:1553–52018;
55. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols*. Nature Publishing Group; 11:1650–672016;
56. Condon K. tispec: calculates tissue specificity from RNA-seq data.
57. Wickham H. Elegant graphics for data analysis (ggplot2). *Media*. Springer; 35:10–10072009;
58. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017;
59. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*. Springer; 20:1–142019;
60. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*. Society for Molecular Biology and Evolution; 30:772–802013;
61. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. Oxford University Press; 25:1972–32009;
62. Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in zoology*. Springer; 11:1–82014;

63. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. Oxford University Press; 30:1312–32014;
64. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*. Oxford University Press; 62:611–52013;
65. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*. Oxford University Press; 49:W293–62021;
66. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al.. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*. Oxford University Press; 37:1530–42020;
67. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al.. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research*. Oxford University Press; 40:e49–e492012;
68. 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;
69. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution*. Oxford University Press; 34:1812–92017;
70. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al.. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*. Nature Publishing Group; 587:246–512020;
71. Hickey G, Paten B, Earl D, Zerbino D, Haussler D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*. Oxford University Press; 29:1341–22013;
72. Duthiel JY, Gaillard S, Stukenbrock EH. MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC genomics*. Springer; 15:1–102014;
73. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in bioinformatics*. Oxford University Press; 12:41–512011;
74. R Core Team. R: A language and environment for statistical computing. Vienna, Austria; 2013;
75. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al.. Circos: an information aesthetic for comparative genomics. *Genome research*. Cold Spring Harbor Lab; 19:1639–452009;
76. Warnes GR. gplots: Various R Programming Tools for Plotting Data.
77. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Briefings in bioinformatics*. Oxford University Press; 18:205–142017;

78. Bentz 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. *Scientific reports*. Nature Publishing Group; 9:1–122019;
79. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al.. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*. Nature Publishing Group; 489:391–92012;
80. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al.. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. Nature Publishing Group; 445:168–762007;
81. Nguyen H-TT, Dang BT, Glenner H, Geffen AJ. Cophylogenetic analysis of the relationship between anemonefish Amphiprion (Perciformes: Pomacentridae) and their symbiotic host anemones (Anthozoa: Actiniaria). *null*. Taylor & Francis; 2020; doi: 10.1080/17451000.2020.1711952.
82. Santini S, Polacco G. Finding Nemo: Molecular phylogeny and evolution of the unusual life style of anemonefish. *Gene*. 2006; doi: 10.1016/j.gene.2006.03.028.
83. Tang KL, Stiassny MLJ, Mayden RL, DeSalle R. Systematics of Damselfishes. *Ichthyology & Herpetology*. 2021; doi: 10.1643/i2020105.
84. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al.. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*. 2018; doi: 10.1093/molbev/msx319.
85. Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: a systematic evaluation. *BMC bioinformatics*. BioMed Central; 19:1–132018;
86. Volff J. Genome evolution and biodiversity in teleost fish. *Heredity*. Nature Publishing Group; 94:280–942005;
87. 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*. Multidisciplinary Digital Publishing Institute; 9:27112020;
88. Schoch H, Kreibich AS, Ferri SL, White RS, Bohorquez D, Banerjee A, et al.. Sociability deficits and altered amygdala circuits in mice lacking *Pcdh10*, an autism associated gene. *Biological psychiatry*. Elsevier; 81:193–2022017;
89. Cheng Y-R, Jiang B-Y, Chen C-C. Acid-sensing ion channels: dual function proteins for chemo-sensing and mechano-sensing. *Journal of biomedical science*. Springer; 25:1–142018;
90. Chapman G, Shanmugalingam U, Smith PD. The role of neuronal pentraxin 2 (NP2) in regulating glutamatergic signaling and neuropathology. *Frontiers in cellular neuroscience*. Frontiers; 13:5752020;
91. Huang S, Zheng C, Xie G, Song Z, Wang P, Bai Y, et al.. FAM19A5/TAF5, a novel neurokinin, plays a crucial role in depressive-like and spatial memory-related behaviors in mice. *Molecular psychiatry*. Nature Publishing Group; 26:2363–792021;

92. Jeong I, Yun S, Shahapal A, Cho EB, Hwang SW, Seong JY, et al.. FAM19A5l affects mustard oil-induced peripheral nociception in zebrafish. *bioRxiv*. Cold Spring Harbor Laboratory; 2020;
93. Litsios G, Kostikova A, Salamin N. Host specialist clownfishes are environmental niche generalists. *Proceedings of the Royal Society B: Biological Sciences*. Royal Society; 2014; doi: 10.1098/rspb.2013.3220.
94. 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. *Marine Biology*. 2011; doi: 10.1007/s00227-010-1583-5.
95. Schmiede PFP, D'Aloia CC, Buston PM. Anemonefish personalities influence the strength of mutualistic interactions with host sea anemones. *Marine Biology*. 2016; doi: 10.1007/s00227-016-3053-1.
96. 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. *Marine biology*. Springer Nature BV; 162:24092015;

TABLES AND FIGURES

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 ×
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 %)

Figure 1. Chromosome architecture of the *Amphiprion ocellaris* genome. From the outermost layer inwards, each layer represents 1) chromosomes indicated by black lines and ordered by size, 2) genic regions (black colored), 3) the number of repeats per 100 kb (orange), 4) the PhyloP score calculated from the whole genome alignment of anemonefishes with

Acanthochromis polyacanthus as an outgroup species (green), and 5) Tissue-specificity index (τ) of each gene (light blue).

Figure 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.

Figure 3. (A) Tissue-specificity of *Amphiprion ocellaris* gene expression. The maximum TPM (transcripts per million) values across tissues and tissue-specificity index (τ) was plotted in the two-dimensional histogram. Trendline (blue) 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. 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 non-outgroup species using the IQ-TREE algorithm.

Figure 5. Gene expression levels of the 62 *Amphiprion ocellaris* genes nearest to the conserved elements in the *Amphiprion ocellaris*/*Amphiprion percula* branch across 13 tissues are shown. Color key indicates $\log_2$ -transformed TPM values.

A

B

A

B

A. polyacanthus

A. ocellaris

A. percula

100/81.5

P. biaculeatus

100/59.6

A. perideraion

100/42.2

A. akallopisos

100/45.1

A. melanopus

100/76.2

A. frenatus

100/81.5

A. sebae

100/78.1

A. polymnus

100/67.4

A. nigripes

100/45.2

A. bicinctus