












The chromosome-scale genome assembly of the yellowtail clownfish *Amphiprion clarkii* provides insights into the melanic pigmentation of anemonefish

Billy Moore ^{1,†} Marcela Herrera ^{2,†} Emma Gairin ² Chengze Li¹ Saori Miura ² Jeffrey Jolly ¹
Manon Mercader ² Michael Izumiyama ¹ Erina Kawai ¹ Timothy Ravasi ^{1,3} Vincent Laudet ^{2,4,*}
Taewoo Ryu ^{1,*}

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

³Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD 4811, Australia

⁴Marine Research Station, Institute of Cellular and Organismic Biology, Academia Sinica, I-Lan 262, Taiwan

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

Email: taewoo.ryu@oist.jp; *Corresponding author: Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan and Marine Research Station, Institute of Cellular and Organismic Biology, Academia Sinica, I-Lan 262, Taiwan. Email: vincent.laudet@oist.jp

[†]These authors contributed equally to this work.

Abstract

Anemonefish are an emerging group of model organisms for studying genetic, ecological, evolutionary, and developmental traits of coral reef fish. The yellowtail clownfish *Amphiprion clarkii* possesses species-specific characteristics such as inter-species co-habitation, high intra-species color variation, no anemone specificity, and a broad geographic distribution, that can increase our understanding of anemonefish evolutionary history, behavioral strategies, fish-anemone symbiosis, and color pattern evolution. Despite its position as an emerging model species, the genome of *A. clarkii* is yet to be published. Using PacBio long-read sequencing and Hi-C chromatin capture technology, we generated a high-quality chromosome-scale genome assembly initially comprised of 1,840 contigs with an N50 of 1,203,211 bp. These contigs were successfully anchored into 24 chromosomes of 843,582,782 bp and annotated with 25,050 protein-coding genes encompassing 97.0% of conserved actinopterygian genes, making the quality and completeness of this genome the highest among all published anemonefish genomes to date. Transcriptomic analysis identified tissue-specific gene expression patterns, with the brain and optic lobe having the largest number of expressed genes. Further analyses revealed higher copy numbers of *erbb3b* (a gene involved in melanocyte development) in *A. clarkii* compared with other anemonefish, thus suggesting a possible link between *erbb3b* and the natural melanism polymorphism observed in *A. clarkii*. The publication of this high-quality genome, along with *A. clarkii*'s many unique traits, position this species as an ideal model organism for addressing scientific questions across a range of disciplines.

Keywords: anemonefish, *Amphiprion clarkii*, chromosome-scale assembly, genome, *erbb3b*, melanism, pigmentation

Introduction

Anemonefish are a group of 28 species that belong to the Pomacentridae family (Fautin and Allen 1992). They are social fish that undergo sex change and live in association with sea anemones (Fautin 1991; Fautin and Allen 1997). Anemonefish have recently gained interest from the scientific community as an emerging model species (Roux et al. 2020), providing an alternative to freshwater teleost models such as zebrafish. This interest has arisen as anemonefish have multiple unique traits, including their community social dynamics, phenotypic plasticity, and ability to complete their life cycle in captivity that make them attractive future model species for exploring scientific questions across ecological, evolutionary, and developmental fields (reviewed in Roux et al. 2020; Laudet and Ravasi 2022).

Amphiprion clarkii's (Bennett 1830) unique features make it arguably the most interesting model species within the Amphiprioninae subfamily (Fig. 1, a and b). These features include the following: (1) Co-habitation of anemones with other species of anemonefish (Hattori 2002; Camp et al. 2016; De Brauwert et al. 2016). (2) *A. clarkii* is the least host-specific anemonefish and only inhabitant of *Cryptodendrum adhaesivum* and *Heteractis malu* (Fautin and Allen 1992, 1997). (3) The broad distribution of *A. clarkii*, and its wide temperature tolerance (Moyer 1980) make it a robust and accessible study organism. (4) *A. clarkii* displays the greatest intra-species variation in melanism (a darkening of body pigmentation) (Fig. 1b; Militz et al. 2016; Salis et al. 2018) among anemonefish. With melanin-based coloration in *A. clarkii* observed to vary with social rank (Moyer 1976, 1980), environmental conditions (Bell et al. 1982), and host anemone (Fautin and Allen 1997; Militz et al. 2016). Despite these trends, the extent to which these

Received: August 25, 2022. Accepted: December 12, 2022

© The Author(s) 2023. Published by Oxford University Press on behalf of the Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

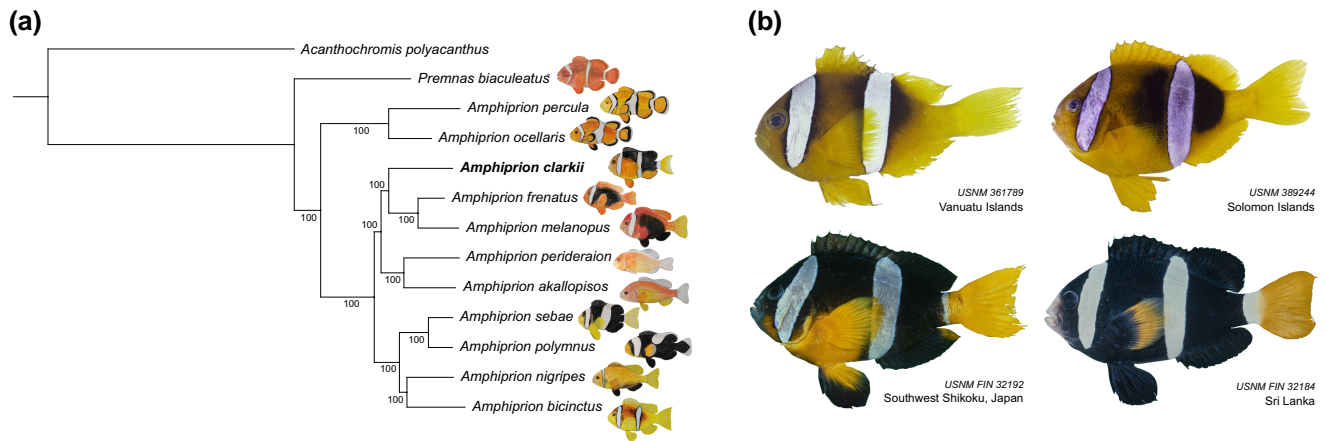


Fig. 1. a) Phylogenetic reconstruction of the Amphiprioninae species tree using a maximum-likelihood approach. Bootstrap support values (%) are shown in each branching node. b) Melanistic polymorphism in *Amphiprion clarkii*. Images taken from the Division of Fishes Collections of the Smithsonian National Museum of Natural History (<https://collections.nmnh.si.edu/search/fishes/>). Catalog number and sampling location are indicated for each specimen.

variables influence anemonefish melanism, particularly at the molecular level, remains uncertain.

For all model species, a high-quality genome is an essential resource, required for many advanced genomic approaches. Yet, the genome of *A. clarkii* is yet to be published, resulting in previous genomic studies of *A. clarkii* using suboptimal de novo assembly-based approaches during analysis (Catalano et al. 2021). Thus, the availability of a high-quality genome will enhance the appeal and quality of future genetic studies of *A. clarkii*. Here, we present the first genome assembly for the yellowtail clownfish *A. clarkii* from Okinawa, Japan. We generated a de novo assembly consisting of 1,840 contigs with an N50 of 1,203,211 bp that were successfully anchored into 24 chromosomes of 843,582,782 bp. We annotated 25,050 protein-coding genes encompassing 97.0% of conserved actinopterygian genes, making the quality and completeness of this *A. clarkii* genome the best of all published anemonefish genomes to date: *Amphiprion percula* (Lehmann et al. 2019), *Amphiprion frenatus* (Marcionetti et al. 2018), *Amphiprion akallopisos*, *Amphiprion bicinctus*, *Amphiprion melanopus*, *Amphiprion nigripes*, *Amphiprion perideraion*, *Amphiprion polymnus*, *Amphiprion sebae*, and *Premnas biaculeatus* (Marcionetti et al. 2019), *Amphiprion ocellaris* (Tan et al. 2018; Marcionetti et al. 2019; Ryu et al. 2022). Using a comparative genomic approach, we also studied genes involved in pigmentation and identified higher copy numbers of the *erbb3b* gene, suggesting a possible link between this gene and the natural melanism polymorphism in *A. clarkii*. Ultimately, the publication of this genome provides a high-quality resource that will enhance the use of *A. clarkii* as a model species, thus facilitating scientific research that spans a wide range of biological disciplines.

Materials and methods

Fish collection and nucleic acid sequencing

Two adult *A. clarkii* (one male and one female) anemonefish were collected for genome and transcriptome sequencing from Tancha Bay, Okinawa (26.4736 N, 127.8278 E) on the 18th of August 2020. These two fish resided together in the anemone *Heteractis crispa* at a depth of 7 m. Following collection, the fish were transferred to Okinawa Institute of Science and Technology (OIST) Marine Science Station where they remained under natural conditions in a 270 L flow through outdoor tank overnight until they were euthanized the day after. Additionally, ten *A. clarkii* juveniles of

different color morphs (orange and black) were collected for quantitative real-time PCR (qPCR) assays from various shallow sites around Okinawa (2–11 m deep) between August 2021 and June 2022 (Supplementary Table 1). Five orange and black juveniles were collected from *Heteractis* sp. and *Stichodactyla* sp. host anemones, respectively. The fish were collected using SCUBA and hand nets, before being euthanized in a 200 mg/L Tricaine Methanesulfonate (MS222) solution and preserved in RNAlater. Samples were placed in 4°C for 48 h and then transferred to a –30°C freezer until RNA extractions were performed. All fish were euthanized following the guidelines outlined by the Animal Resources Section of OIST Graduate University.

For genome sequencing with PacBio (Pacific Biosciences, CA, USA), the liver of the adult female was extracted, snap frozen in liquid nitrogen, and stored at –80°C. Liver tissue from the adult male, on the other hand, was used for Hi-C sequencing. Thirteen tissues from the same (male) individual were also used for transcriptome sequencing. Finally, total RNA from the whole body of the juveniles was extracted to perform qPCR. Details on extractions and library preparation are provided in Supplementary Methods 1.

Chromosome-scale genome assembly

Raw PacBio long reads were assembled de novo using Flye v2.9 (Kolmogorov et al. 2019) with the “keep haplotypes” option. Assessment of the resulting genomic contigs with Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.4 (Simão et al. 2015) and the Actinopterygii-lineage dataset (actinopterygii_odb10) identified high levels of gene duplication. Therefore, duplicates were removed from the initial Flye assembly using purge_dups v0.03 (Guan et al. 2020). The chromosome-scale genome assembly was generated by Phase Genomics using the de novo assembly, FALCON-phase (Kronenberg et al. 2018), Hi-C sequencing reads, and Phase Genomics’ Proximo algorithm based on Hi-C chromatin contact maps (as described in Bickhart et al. 2017). Error correction of this chromosome-scale assembly was conducted with Illumina short reads and Pilon v1.23 (Walker et al. 2014). Quality-trimmed Illumina short reads (Trimmomatic v0.39) (Bolger et al. 2014) using the parameters “ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:keepBothReads LEADING:3 TRAILING:3 MINLEN:36” were aligned to the genome using Bowtie2 v2.4.1 (Langmead and Salzberg 2012) with the default parameters, and the resulting SAM files were converted to BAM

format using SAMtools v1.10 (Li et al. 2009). BAM files were then used as input for error correction with Pilon. The quality and completeness of the final assembly was assessed using Quast v5.0.2 (Mikheenko et al. 2018) and BUSCO v4.1.4 (actinopterygii_odb10) (Simão et al. 2015), and base-level accuracy (QV) was assessed using trimmed Illumina short reads, Merqury v1.3 (Rhie et al. 2020), and a k-mer value of 20.

Genome size and coverage estimation

Genome size and heterozygosity were estimated using quality-trimmed Illumina short reads (as described above), Jellyfish v2.3.0 (Marçais and Kingsford 2011) with k-mer = 17, and GenomeScope v1.0 (Vurture et al. 2017) with default parameters. Additionally, the overall mean genome-wide base-level coverage of the final assembly was calculated by aligning the raw PacBio reads to the assembled chromosome sequences using Pbbmm2 v1.4.0 (<https://github.com/PacificBiosciences/pbbmm2>). The genomeCoverageBed function of BEDTools v2.30.0 (Quinlan 2014) was then used to calculate the per-base coverage of aligned reads across all chromosomal sequences.

Prediction of gene models in *A. clarkii*

Repetitive elements were identified de novo using RepeatModeler v2.0.1 (Flynn et al. 2020) with the “LTRStruct” option. RepeatMasker v4.1.1 (Tempel 2012) was used to screen known repetitive elements with two inputs: (1) the RepeatModeler output and (2) the vertebrata library of Dfam v3.3 (Storer et al. 2021). The resulting output files were validated and merged before redundancy was removed using GenomeTools v1.6.1 (Gremme et al. 2013). To identify and annotate candidate gene models, BRAKER v2.1.6 (Brůna et al. 2021) was used with mRNA and protein evidence. For annotation with BRAKER, the chromosome sequences were soft masked using the maskfasta function of BEDTools v2.30.0 (Quinlan 2014) with the “soft” option. Protein evidence consisted of protein records from UniProtKB/Swiss-Prot (UniProt Consortium 2021) as of 2021 January 11 (563,972 sequences) as well as selected fish proteomes from the NCBI database (*A. ocellaris*: 48,668, *Danio rerio*: 88,631, *Acanthochromis polyacanthus*: 36,648, *Oreochromis niloticus*: 63,760, *Oryzias latipes*: 47,623, *Poecilia reticulata*: 45,692, *Stegastes partitus*: 31,760, *Takifugu rubripes*: 49,529, and *Salmo salar*: 112,302). Transcriptomic reads from 13 tissues were used as mRNA evidence. These Illumina short reads were trimmed with Trimmomatic v0.39 (Bolger et al. 2014) as described above and mapped to the chromosome sequences with HISAT2 v2.2.1 (Kim et al. 2019). The resulting SAM files were converted to BAM format with SAMtools v1.10 (Li et al. 2009) and used as input for BRAKER. Of the resulting gene models, only those with supporting evidence (mRNA or protein hints) or with homology to the Swiss-Prot protein database (UniProt Consortium 2021) or Pfam domains (Mistry et al. 2021) were selected as final gene models. Homology to Swiss-Prot protein database and Pfam domains was identified using Diamond v2.0.9 (Buchfink et al. 2015) or InterProScan v5.48.83.0 (Zdobnov and Apweiler 2001), respectively. Functional annotation of the final gene models was completed using NCBI BLAST v2.10.0 (Altschul et al. 1990) with the NCBI non-redundant (nr) protein database. Gene Ontology (GO) terms were assigned to *A. clarkii* genes using the BLAST output and the “gene2go” and “gene2accession” files from the NCBI ftp site (<https://ftp.ncbi.nlm.nih.gov/gene/DATA/>). Completeness of the gene annotation was assessed with BUSCO v4.1.4 (actinopterygii_odb10) (Simão et al. 2015).

Mitochondrial genome assembly and annotation

Quality-trimmed Illumina reads were used as input for GetOrganelle v1.7.0 (Jin et al. 2020) which was used to assemble the mitochondrial genome of *A. clarkii*. Mitochondrial genes

were then annotated with MitoAnnotator v3.67 (Sato et al. 2018). The mitochondrial genome assembled here was compared with two previously published mitochondrial genomes of *A. clarkii* (Tao et al. 2016; Thongtam Na Ayudhaya et al. 2019) (NCBI accessions: NC_023967.1 and AB979449.1) using BLASTn v2.10.0 (Altschul et al. 1990) with an *e*-value 10^{-4} as a threshold to predict overall sequence identity.

Analysis of tissue-specific gene expression

As for gene annotation, quality-trimmed transcriptomic reads from 13 tissues were mapped to the chromosome assembly with HISAT2 v2.2.1 (Kim et al. 2019) and the resulting SAM files were converted to BAM format using SAMtools v1.10 (Li et al. 2009). The resulting BAM files and final gene annotation file were used as input into StringTie v2.1.4 (Pertea et al. 2016) to quantify expression levels and normalize TPM (transcripts per million). The tissue specificity index (τ) of each gene was calculated using the R package tispec v0.99 (Condon 2020) and a two-dimensional histogram was used to display the relationship between τ and expression level (TPM). The number of genes expressed in each tissue and different combinations of tissues were displayed in an Upset plot generated with the UpSetR v1.4.0 R package (Conway et al. 2017).

Gene orthology and phylogenetic analyses

Orthologous relationships between *A. clarkii* and the other anemonefish were investigated using OrthoFinder v2.5.2 (Emms and Kelly 2019). Briefly, protein sequences of *A. clarkii*, *A. akallopisos*, *A. bicinctus*, *A. frenatus*, *A. melanopus*, *A. nigripes*, *A. ocellaris*, *A. percula*, *A. perideraion*, *A. polymnus*, *A. sebae*, and *P. biaculeatus* (Marcionetti et al. 2018, 2019; Lehmann et al. 2019; Ryu et al. 2022), and the spiny chromis *A. polyacanthus* (used as the outgroup species), were reciprocally blasted against each other and clusters of orthologous genes were defined using the default settings. In all cases, only the longest isoform of each gene model was used. Sequences of single-copy orthologs present in all species were aligned using MAFFT v7.130 (Katoh and Standley 2013) using the options “local pair”, “maxiterate 1,000”, and “leavegappyregion”, trimmed with trimAl v1.2 (Capella-Gutiérrez et al. 2009) using the “strict” flag, and then concatenated with FASconCAT-G (Kück and Longo 2014). Maximum-likelihood phylogenetic trees were then constructed with RAXML v8.2.9 (Stamatakis 2014). The MPI version (raxmlHPC-MPI-AVX) was executed using a LG substitution matrix, heterogeneity model GAMMA, and 1,000 bootstrap inferences. Trees were visualized using iTOL v6.4 (Letunic and Bork 2021). Branch supports in the trees were evaluated with the standard bootstrap values from RaxML.

Identification of pigmentation genes

Based on Lorin et al. (2018) and Salis et al. (2021), a list of 211 genes known to be involved in pigmentation were identified for this study (Supplementary Table 2). For each of these genes, the related protein sequence of *A. ocellaris* (or, if not available, the closest related species) was retrieved from the Ensembl genome database (<https://www.ensembl.org>, last accessed on February 2022). Next a BLASTp search (using the parameters “-evalue 10^{-10} -max_target_seqs 5”) of these 211 protein sequences was performed against *A. clarkii* gene models. The 211 *A. clarkii* gene models this identified were then confirmed to be the correct pigmentation genes by checking the previously completed *A. clarkii* gene annotation.

Confirmation of *erbb3b* genes in *A. clarkii*

The presence of three *erbb3b* genes identified in the *A. clarkii* genome was validated using polymerase chain reaction (PCR) and DNA from the same individual used for whole-genome sequencing. Additionally, PCR was used to investigate the presence of *erbb3b* genes in other species of anemonefish (*A. ocellaris*, *A. frenatus*, *A. polymnus*, *A. perideraion*, *A. sandaracinos*, and *Amphiprion akindynos*). For these species, DNA was extracted from a piece of caudal fin using a Maxwell RSC Blood DNA Kit (Promega, Madison, WI, USA). Extractions were performed following the manufacturer's instructions with the exception of a longer two-hour lysis step. DNA was quantified using a Qubit dsDNA BR (Broad Range) Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). DNA was then diluted to a working concentration of 20 ng/μl and stored at -30°C.

Primers targeting a conserved region of intron 8 of all three *erbb3b* genes in *A. clarkii* were designed using Geneious v2022.1 (Kearse et al. 2012). Gaps of different lengths were present across the three genes (Supplementary Fig. 1), thus making it easy to amplify them using only one pair of primers. PCRs were performed using the forward 5' TGTCCACTTCCAGGATGAGAC 3' and reverse 3' ACCCCTCGATCTCATCTCTGT 5' primers. Each PCR run used 12.5 μl Q5 High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, MA, USA), 2.5 μl template DNA, 1.25 μl 10 μM forward and reverse primer, and 7.5 μl nuclease-free water for a final reaction volume of 25 μl. The thermal cycling conditions used were 30 s at 98°C, followed by 35 cycles of 10 s at 98°C, 30 s at 67°C, and 30 s at 72°C, followed by a final extension step of 2 min at 72°C. For each sample, PCR products were visualized using 2% agarose gel electrophoresis (Supplementary Fig. 2), excised from the gel and purified using a QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany). PCR amplicons were then bidirectionally sequenced by the company FASMAC, which uses Applied Biosystems Big Dye Terminator v3.1 technology and an Applied Biosystems 3130xl Genetic Analyzer (Applied Biosystems, Waltham, MA, USA). Sequence analysis was performed using the software Geneious v2022.1 (Kearse et al. 2012).

qPCR assays to measure *erbb3b* gene expression in *A. clarkii*

In total, ten juveniles (five orange and five black) were assayed to measure gene expression of the three *erbb3b* genes identified in *A. clarkii*. Specific primers for each *A. clarkii* *erbb3b* gene (two short genes containing 1,911 bp and one long gene of 4,275 bp) were designed manually based on their genomic sequence (Supplementary Table 3). Primers previously used by Roux et al. (2022) with *A. ocellaris* were used to target the housekeeping genes ribosomal protein L7 (*rpl7*) and ribosomal protein L32 (*rpl32*). Extracted RNA from each juvenile (as described in Supplementary Methods 1) was converted to cDNA using PrimeScript RT-PCR Kit (Takara Bio, Shiga, Japan). The efficiency and specificity of the designed primers was tested through PCR using the GoTaq Green Master kit (Promega, Madison, USA) with thermal cycling conditions of 2 min at 95°C, followed by 30 cycles of 45 s at 95°C, 45 s at 60/63/65°C, and 30 s 72°C, a final extension step of 5 min at 72°C, preservation at 4°C, and subsequent agarose gel electrophoresis (Supplementary Fig. 3). The specificity was also tested through direct forward and reverse Sanger sequencing by aligning the forward and reverse outputs and blasting the obtained amplicons against the reference genomic sequences (Supplementary Fig. 4).

The expression of each *erbb3b* gene and the two housekeeping genes (*rpl7* and *rpl32*) was obtained by RT-qPCR at 65°C (PrimeScript transcriptase, Takara, SYBRgreen) and normalized with the Pfaffle equation (Ståhlberg et al. 2004):

$$RE = \frac{E(gi)^{Ct(gi)ctrl - Ct(gi)sample}}{\sqrt{E(rpl7)^{Ct(rpl7)ctrl - Ct(rpl7)sample} * E(rpl32)^{Ct(rpl32)ctrl - Ct(rpl32)sample}}}$$

where RE is the relative expression, $E(x)$ is the efficiency of the amplification for isoform x , and $Ct(x)$ is the quantification cycle of gene x .

Results and discussion

Chromosome-scale genome assembly of *A. clarkii*

We assembled the genome of the anemonefish *A. clarkii* by sampling two individuals from Okinawa and generating 19,675,845 PacBio reads with an average read length of 13,144 bp (Supplementary Table 4). These reads were assembled de novo using Flye v2.9 (Kolmogorov et al. 2019) with the initial assembly consisting of 2,635 contigs of 855,782,104 bp with an N50 of 1,187,902 bp. Following processing with Purge_Dups v0.0.3 (Guan et al. 2020), the final de novo assembly consisted of 1,840 contigs of 845,361,362 bp and had an N50 of 1,203,211 bp. Using 228,099,434 150 bp Hi-C reads from liver tissue and the ProximoTM scaffolding platform (Phase Genomics, WA, USA), we generated 24 chromosomes of 843,295,090 bp and 168 short scaffolds (2,826,673 bp) that were not placed into chromosomes. This chromosome-scale assembly was polished with Illumina short reads using Pilon (Walker et al. 2014) generating a final assembly of 843,582,782 bp. Chromosome lengths ranged from 42,519,526 bp to 20,115,265 bp (Fig. 2). The mean base-level coverage of these chromosomes was 250.4x and the overall base-level accuracy (QV) was 39.44 (Supplementary Table 5). The final *A. clarkii* genome contained 118,106 non-ATGC characters, a GC content of 39.71% and a repeat content of 44.26% (Table 1). The structure of our genomic assembly was compared with properties of the *A. clarkii* genome estimated by Jellyfish v2.3.0 (Marçais and Kingsford 2011) and GenomeScope v1.0 (Vurture et al. 2017) with Illumina short reads. At k-mer=17, genome size was estimated at 793,832,155 bp, repeat content was estimated at 42.33%, and heterozygosity was estimated at 0.51%. Repeat content was identified using RepeatMasker v4.1.1 (Tempel 2012) by querying repetitive elements from the Dfam (Storer et al. 2021) vertebrata library and repetitive elements identified de novo using RepeatModeler v2.0.1 (Flynn et al. 2020) against the *A. clarkii* genome. This approach identified repeat content of 373,358,331 bp (Supplementary Fig. 5). Of the identified repetitive elements, DNA transposons were the most frequent, occupying 23.26% of the *A. clarkii* genome. Long interspersed nuclear elements (7.33%), long terminal repeats (3.75%), and simple repeats (1.73%) were the next most frequent in the genome. However, 27.93% of the *A. clarkii* genome is occupied by repetitive elements that could not be identified (Supplementary Fig. 5).

Comparison with the two other chromosome-scale anemonefish genomes revealed similar structures and assembly statistics between the *A. clarkii*, *A. percula*, and *A. ocellaris* genomes (Lehmann et al. 2019; Ryu et al. 2022). For example, with sizes of 890,200,000 bp and 856,612,077 bp, respectively, the *A. percula* and *A. ocellaris* genomes are only slightly larger than the 843,582,782 bp *A. clarkii* genome assembled here, while the GC content of all three genomes is between 39.55% and 39.71%. The

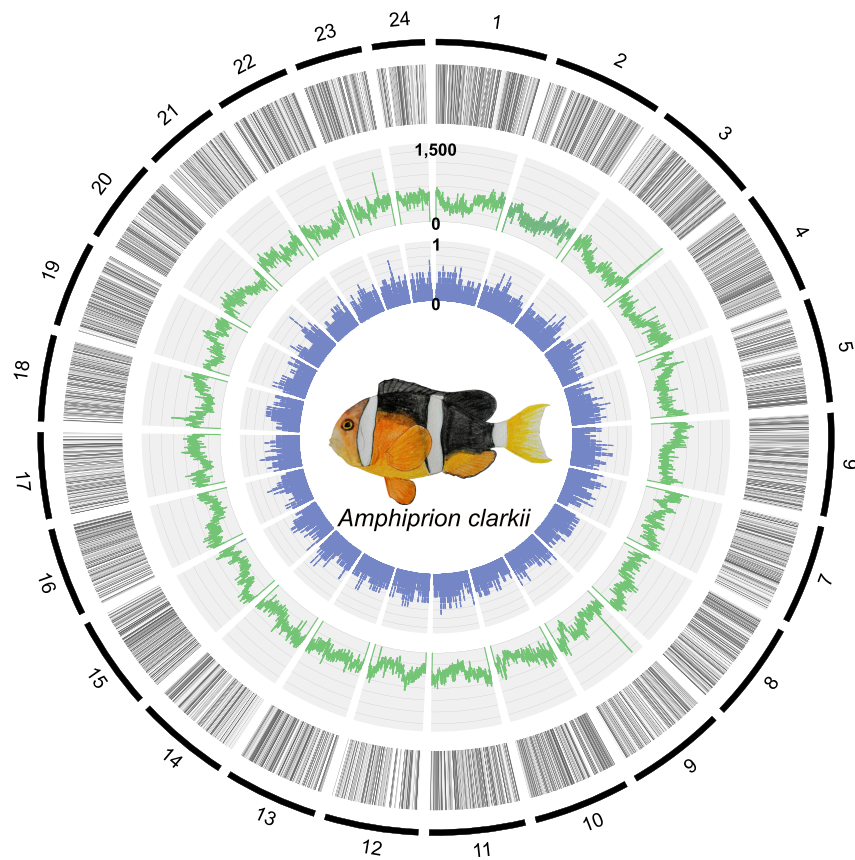


Fig. 2. The genome structure of *Amphiprion clarkii*. From the outside in, the circo plot layers display: (1) reference chromosomes 1 to 24 in clockwise orientation, (2) black bars showing genic regions of the genome, (3) green bars showing the number of repeats per 100 kb, and (4) blue bars showing tissue specificity index of genes displayed in layer 1. Drawing of *A. clarkii* is displayed in the middle of the circo plot.

repeat content of the *A. clarkii* genome closely matched that of *A. ocellaris* (44.7% vs 44.26%), yet it was greater than that of the *A. percula* genome (28%). This is likely due to the different repeat annotation methods used by Lehmann et al. (2019) compared with those used here and by Ryu et al. (2022). As similarities in repeat content of the sister species *A. ocellaris* and *A. percula* (Litsios et al. 2014) would be expected to match the similarity between *A. ocellaris* and *A. clarkii*. Furthermore, the characteristics of these high-quality anemonefish genomes match that of closely related

species within the Pomacentridae family such as *A. polyacanthus* (991,600,000 bp) (ASM210954v1, GCF_002109545.1, NCBI).

Genome completeness was assessed using BUSCO v4.1.4 (Simão et al. 2015) and the Actinopterygii-lineage dataset. The *A. clarkii* genome contained 3,593 conserved actinopterygian benchmark genes giving a BUSCO score of 98.7% (Complete and single copy: 97.8%; Complete and duplicated: 0.9%; Fragmented: 0.4%; Missing: 0.9%) (Table 1). Previous non-chromosome-scale anemonefish genomes (Marcionetti et al. 2018, 2019; Tan et al. 2018) are much less contiguous (contig numbers of 17,801 and 6,404, respectively) and have a maximum BUSCO score of 96.5% (Actinopterygii). Although contiguity is important, the genic completeness of an assembly is vital for its future use by the research community. With BUSCO scores of <97.1%, previous chromosome-scale anemonefish genome assemblies are less complete than the assembly presented here, highlighting this *A. clarkii* assembly as the best quality for anemonefish to date.

Table 1. Genome assembly statistics, gene annotation statistics, and BUSCO completeness.

Chromosome assembly size	843,582,782 bp
Non-ATGC characters	118,106
GC content	39.7%
Mean base-level coverage	250.4x
Repeat content	44.3%
Chromosome-scale N50	26,694,648 bp
Contig N50	1,203,211 bp
BUSCO genome completeness	3,593 (98.7%)
Complete and single copy	3,560 (97.8%)
Complete and duplicated	33 (0.9%)
Fragmented	13 (0.4%)
Missing	34 (0.9%)
Number of protein-coding genes	25,050
BUSCO gene annotation completeness	3,532 (97.0%)
Complete and single copy	3,498 (96.1%)
Complete and duplicated	34 (0.9%)
Fragmented	41 (1.1%)
Missing	67 (1.9%)

A. clarkii gene annotation

The genome was annotated using BRAKER v2.1.6 (Brůna et al. 2021) with mRNA and protein evidence. This resulted in an initial 41,083 predicted gene models. These gene models included different isoforms from the same gene locus, therefore gene models were filtered to keep only the longest isoform of each gene. This resulted in 36,949 unique gene models. Only gene models with either mRNA or protein evidence support (24,571) or homology to the Swiss-Prot protein database or Pfam domains (479) were retained. This resulted in 25,050 final gene models. Of these 25,050

gene models, 23,700 (94.61%) had significant homology to the NCBI *nr* database (bit-score ≥ 50) and 19,982 genes (79.77%) had at least one associated GO term. The completeness of this set of annotated genes was assessed using BUSCO v4.1.4 (Simão et al. 2015) and the Actinopterygii-lineage dataset. The 25,050 gene models contained 3,532 conserved actinopterygian benchmark genes, giving a BUSCO score of 97.0% (Complete and single copy: 96.1%; Complete and duplicated: 0.9%; Fragmented: 1.1%; Missing: 1.9%) (Table 1). Of the previously reported anemonefish genome annotations, the annotation of the chromosome-scale *A. ocellaris* genome (Ryu et al. 2022) was the most complete with a BUSCO score of 96.62%. Thus, the annotation reported here represents the most complete genome annotation for an anemonefish to date. This high-quality annotation will facilitate genetic studies of *A. clarkii* that require an understanding of specific gene functions and locations.

Assembly and annotation of mitochondrial genome

The mitochondrial genome of *A. clarkii* was assembled using GetOrganelle v1.7.0 (Jin et al. 2020) and annotated with MitoAnnotator v3.67 (Sato et al. 2018). This resulted in a 16,812 bp circular mitogenome that contained 37 organelle genes consisting of 13 protein-coding genes, 22 tRNAs, and 2 rRNAs, as well as one control region (Supplementary Fig. 6, Supplementary Discussion 1).

Gene expression tissue specificity

The tissue specificity of the 25,050 *A. clarkii* genes identified here was investigated using the transcriptomes of 13 different tissues (Supplementary Table 4). The total number of genes and the unique number of genes expressed per tissue, as well as tau index (τ) (Kryuchkova-Mostacci and Robinson-Rechavi 2017) were used to quantify tissue specificity. A total of 1,814 genes were expressed in all tissues (Fig. 3a), which is similar to the 1,957 genes expressed in all tissues in *A. ocellaris* (Ryu et al. 2022), but less than the ~8,000 genes ubiquitously expressed in multiple human and mouse tissues (Ramsköld et al. 2009). Although only 1,814 genes were expressed in all tissues, 3,697 genes have a $\tau \leq 0.2$, indicating they are expressed in nearly all tissues without biased expression, and are therefore considered housekeeping genes. Thus, the number of housekeeping genes identified in *A. clarkii* is very similar to *A. ocellaris* (3,431 housekeeping genes) (Ryu et al. 2022). Genes with greater tissue specificity were more abundant than housekeeping genes as we identified 4,362 highly specific genes ($0.85 \leq \tau < 1$) as well as 1,068 absolutely tissue-specific genes ($\tau = 1$), that were only expressed in one tissue type. The eye expressed the highest number of these unique genes (563) with the brain (491) displaying the second highest (Fig. 3a). However, when considered together, the brain and optic lobe expressed 1,824 absolutely tissue-specific genes. The number of unique genes expressed in tissue types reflected the total number of genes expressed in different tissues, as the brain (13,714), optic lobe (13,138), and eye (12,003) expressed a high number of genes. The number of genes expressed per tissue are very similar to those observed for *A. ocellaris* (Ryu et al. 2022) and the corresponding human (Ramsköld et al. 2009) tissues, yet is slightly higher than the number expressed in corresponding rainbow trout tissues (Salem et al. 2015). Interestingly, the gall bladder expressed the lowest number of genes (3,252), with only 19 of these being absolutely tissue-specific. Across all tissues, tissue specificity of gene expression negatively correlated (Pearson's correlation coefficient between τ and \log_{10}) with expression levels (Fig. 3b), indicating that tissue-specific genes

have lower expression levels in general (Kryuchkova-Mostacci and Robinson-Rechavi 2017).

Ortholog identification and anemonefish phylogeny

We used OrthoFinder v2.5.2 (Emms and Kelly 2019) to identify orthologous relationships between the amino acid sequences of *A. clarkii* and 11 other anemonefish (and the spiny chromis *A. polyacanthus* as an outgroup species). Overall, 96.7% of the sequences could be assigned to one of 29,855 orthogroups, with the remainder identified as "unassigned genes" with no clear orthologs (Supplementary Table 6). Fifty percent of all proteins were in orthogroups consisting of ≥ 13 genes and were contained in the largest 10,641 orthogroups. 15,771 orthogroups were shared among all the species examined here, of which 12,600 consisted entirely of single-copy genes (Supplementary Table 6).

Phylogenetic reconstruction using these single-copy genes yielded robust phylogenetic relationships, with all branches supported by 100% bootstrap values (Fig. 1a). Furthermore, our tree topology is consistent with previous studies (Litsios et al. 2014; Litsios and Salamin 2014; Marcionetti et al. 2019). Recovered at the base of the tree was *P. biaculeatus*, with the *A. ocellaris/A. percula* complex at the root of all other anemonefish, and four major clades: (1) *A. frenatus* and *A. melanopus* and its sister species *A. clarkii*, (2) the skunk anemonefishes *A. akallopisos* and *A. perideraion*, (3) the closely related species *A. polymnus* and *A. sebae*, and (4) an Indian Ocean clade represented by *A. bicinctus* and *A. nigripes*. Generally situated in the inferior half of the tree, *A. clarkii* is neither the most ancestral nor the most derivative species, but a species with an intermediate level of evolution within the Amphiprioninae subfamily (Litsios et al. 2014; Litsios and Salamin 2014). Interestingly, this tree differs from the one reported in Ryu et al. (2022). Pomacentrids (anemonefishes in particular) have long been a challenge in systematics due to their high diversity and intraspecific variation (Tang et al. 2021), thus future analyses including more species, especially those located close to the base of the tree, might be critically important in establishing a well-resolved phylogeny.

Identification of specific pigmentation genes in *A. clarkii*

To identify gene families that are enriched in the *A. clarkii* genome, we counted the orthogroups with genes in all species, and then selected those in which numbers were $\geq 2\times$ higher in *A. clarkii* than other anemonefishes. Interestingly, most of these were associated with processes related to melanocyte development and melanosome transport (Supplementary Table 7). Keratin type II (orthogroup ID OG0000508) and receptor tyrosine protein kinase *erbB3* (orthogroup ID OG0001853) stood out, as these contained four and three genes in *A. clarkii* compared with two and one orthologs in other anemonefish species, respectively (however, *A. ocellaris* contained two *erbB3* orthologs) (Fig. 4a; Supplementary Table 7). Keratins are major structural proteins in epithelial cells that influence the distribution and arrangement of melanosomes (Gu and Coulombe 2007), which ultimately impact the color patterning of animals. Indeed, mutations in keratin domains can cause hyper-/hypo-pigmented phenotypes (Uttam et al. 1996). In particular, keratin type II has been implicated in the production of color in frogs (in morphs that have black dorsum and legs) (Stuckert et al. 2021). Here, one gene was identified as keratin type II cytoskeletal 8-like isoform, whereas the other three were annotated as keratin type II cytoskeletal cochlear-like (Supplementary Fig. 7), a key component of the large transcellular cytoskeletal network in the cochlea's organ of Corti (that contributes to hearing) (Mogensen et al. 1998).

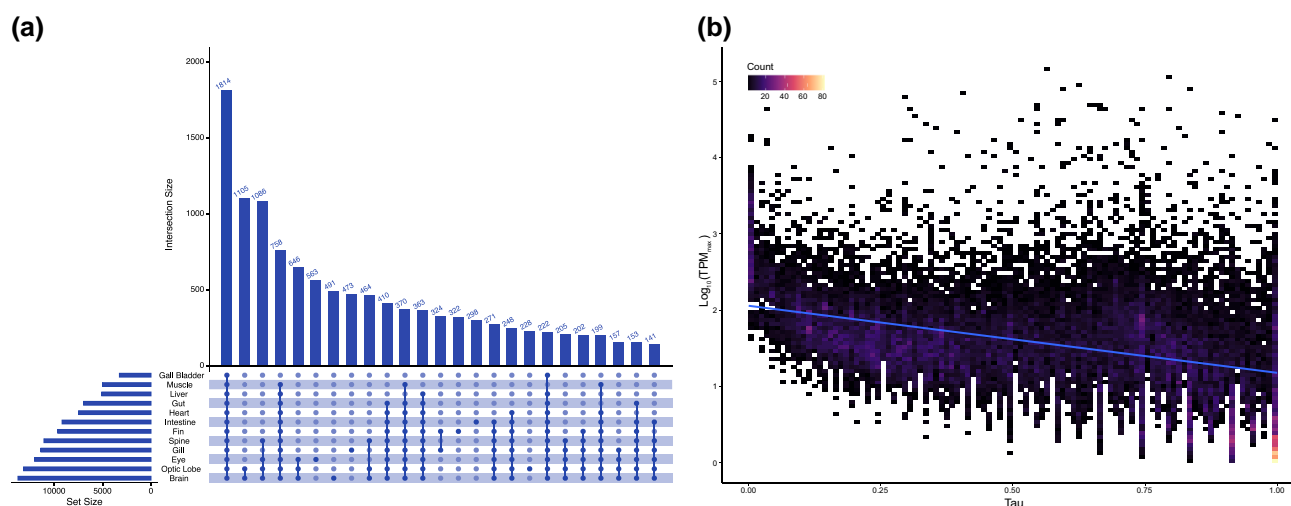


Fig. 3. a) Upset plot displaying the number of genes expressed (intersection size) in individual and combinations of different tissues. Transcripts per million (TPM) values of >10 were used as a threshold for gene expression. Note that “rest of brain” and “cerebrum” tissues are combined to give the final “brain” tissue counts. b) Two-dimensional histogram displaying the relationship between the maximum TPM and tissue specificity index (Tau, τ) of each gene. Trendline displays Pearson’s correlation between τ and \log_{10} .

Interestingly, however, keratin type II cytoskeletal cochlear-like has been found to be highly expressed in trout skin (Djurdjević et al. 2019). On the other hand, receptor tyrosine protein kinase *erbB3* belongs to the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases (ErbB), a group of proteins that have essential roles in regulating cell proliferation and differentiation (Stein and Staros 2006; Wieduwilt and Moasser 2008). In particular, *erbB3* signaling is required for the formation of new melanocytes

during metamorphosis (Hultman et al. 2009). Mutations in this gene result in a phenotype with fewer melanophores (i.e. picasso mutant in zebrafish) (Budi et al. 2008).

ErbB3b genes in *A. clarkii*

Three sequences annotated as *erbB3b* (herein referred to as *erbB3b* gene1, *erbB3b* gene2, and *erbB3b* gene3) were identified in *A. clarkii*, two were identified in *A. ocellaris*, and only one in all the other

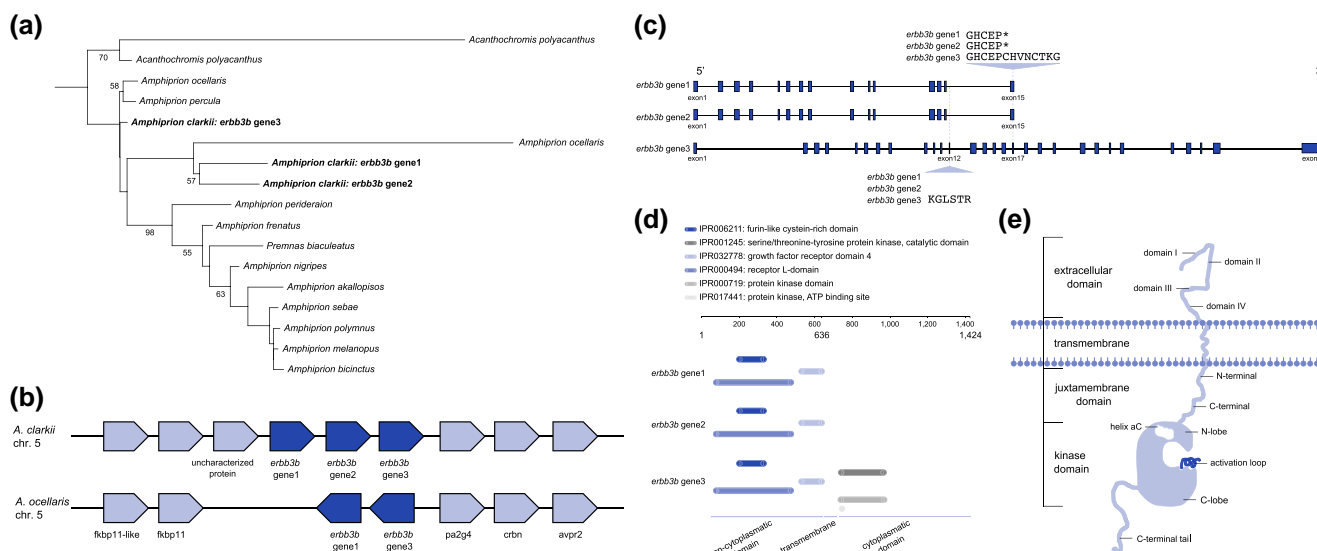


Fig. 4. a) Maximum-likelihood phylogeny of protein sequences from the *erbB3b* gene in anemonefish. Bootstrap support values (%) above 50 are shown in each branching node. b) Syntenic *erbB3b* genes between *Amphiprion clarkii* and the false clownfish *Amphiprion ocellaris*. The two *erbB3b* genes identified in *A. ocellaris* are orthologous to the *erbB3b* gene1 and *erbB3b* gene3 in *A. clarkii*. Dark blue-colored boxes indicate the *erbB3b* gene and light blue-colored boxes represent the flanking orthologous genes. Each gene is arranged considering the transcriptional direction, and the physical distance is ignored. c) Structure of *erbB3b* genes in *A. clarkii*. Both short sequences (*erbB3b* gene1 and *erbB3b* gene2) are comprised of 15 exons whereas the long sequence (*erbB3b* gene3) consists of 29 exons. The short sequences have a gap that corresponds to exon12 of the long sequence and end in the same position as exon17 of the long gene. Exons are represented by blue-colored boxes and introns by a black solid line. Each exon and intron is represented considering the size and physical distance. Asterisk (*) represents the stop codon. d) Functional analysis of the *erbB3b* genes as implemented by InterPro. Protein domains are color-coded in the legend. e) Structure diagram of the *erbB3b* protein [adapted from Li et al. (2013)] including an extracellular ligand binding domain, a transmembrane helix domain, and an intracellular tyrosine kinase domain. The extracellular domain is made of a tandem repeat of leucine-rich (domains I and III) and cysteine-rich segments (domains II and IV). The intracellular domain is a continuation of the transmembrane region and is divided into a juxtamembrane domain, a kinase domain, and a C-terminal tail. The kinase domain includes the N-lobe, helix aC, activation loop, and C-lobe.

anemonefish species (Fig. 4, a and b). Notably, *erbb3b* gene1 and *erbb3b* gene2 are much shorter (636 amino acids) compared with *erbb3b* gene3 (1,424 amino acids) (Fig. 4c). In the case of *A. ocellaris*, one short and one long sequence was retrieved (Fig. 4b). Synteny analysis between *A. clarkii* and *A. ocellaris* revealed all genes are located in tandem on chromosome 5, and are flanked by peptidyl-prolyl cis-trans isomerase FKBP11 and proliferation-associated 2G4 genes (Fig. 4b). The protein sequences of *erbb3b* gene1 and *erbb3b* gene2 are almost identical, with both comprising of 15 exons, a gap that corresponds to exon 12 in *erbb3b* gene3, and a stop codon at the same position as exon 17 in the *erbb3b* gene3 (Fig. 4c). Furthermore, functional analysis using the InterPro database (Blum et al. 2021) revealed six protein domains: (1) the extracellular growth factor receptor domain IV, (2) the furin-like cysteine-rich domain, (3) the receptor L-domain, and the cytoplasmic domains (4) serine/threonine-tyrosinase protein kinase, (5) protein kinase (catalytic subunit), and (6) protein kinase ATP-binding site (Hanks et al. 1988; Cho et al. 2002). Interestingly, neither *erbb3b* gene1 or *erbb3b* gene2 have the cytosolic protein kinase domains but only the extracellular ligand domain (Fig. 4d).

The structure of *erbb3b* (Fig. 4e) is typical in the ErbB receptor tyrosine kinase family. It includes an extracellular ligand binding domain of 600–630 amino acids, a transmembrane helix domain, and an intracellular domain of ~600 amino acids that includes the tyrosine kinase and regulatory sequences (Li et al. 2013). The extracellular domain itself is made of a tandem repeat of leucine-rich segments that make up the ligand binding (domains I and III), and cysteine-rich domains (II and IV, with the former containing the dimerization arm). The intracellular domain is a continuation of the transmembrane region and is divided into a juxtamembrane region (which is in turn divided into N-terminal and C-terminal), kinase domain, and C-terminal tail. Located in the N-terminal extremity of the catalytic domain, there is a lysine residue that has been shown to be involved in ATP-binding. The kinase domain includes an N-lobe, helix α C, activation loop, and C-lobe (Li et al. 2013). However, unlike other ErbB family members, *erbb3* lacks endogenous kinase activity (Jura et al. 2009; Li et al. 2013). Thus, phosphorylation of target proteins only occurs if ligand binding leads to dimerization with other tyrosinase kinase receptors, such as *erbb2*, that do have kinase activity (Jura et al. 2009; Li et al. 2013). Following ligand binding, intracellular pathways are then triggered, resulting in the formation of new melanocytes (Hultman et al. 2009). Given that *A. clarkii* is a polymorphic species in terms of pigmentation, particularly melanization, finding higher copy numbers of a gene implicated in melanophore development calls for further analysis.

Validation of *erbb3b* genes in *A. clarkii* and other anemonefish species

Through performing PCR, all three *erbb3b* genes described above were confirmed to be present in the *A. clarkii* genome (Supplementary Fig. 2). However, while our bioinformatic analysis identified only one ortholog for the other species (except *A. ocellaris*, for which we identified two), PCR and Sanger sequencing highlighted the presence of two *erbb3b* genes in all other anemonefish species tested here. Alignment of these sequences to the *erbb3b* genes from *A. clarkii* indicates these two copies correspond to one short and one long copy of *erbb3b* (orthologous to *erbb3b* gene1 and *erbb3b* gene3, respectively). To further validate these results, we performed a BLASTn search (using the parameters “-task blastn -evalue 10^{-10} -max_target_seqs 5”) of the three *A. clarkii* *erbb3b* genes against the genomes of all the other species (Marcionetti et al. 2018, 2019; Lehmann et al. 2019; Ryu et al. 2022),

with the results matching genes annotated as receptor tyrosine protein kinase *erbb3* (Supplementary Table 8). However, these matches corresponded to only 25–35% of the *A. clarkii* *erbb3b* genes total length (10,283 bp for *erbb3b* gene1 and 20,883 bp for *erbb3b* gene3 from start to stop codons). With the exception of the *A. ocellaris* and *A. percula* genomes (for which we obtained two and one match, respectively), the genome assemblies of all other anemonefish species studied here are primarily based on Illumina short-read technology and are therefore highly fragmented (resulting in gaps and reduced contiguity). Therefore, genome quality likely underpins why our bioinformatic analysis detected one *erbb3b* gene in these species, while PCR detected two. Nevertheless, with three copies of the *erbb3b* gene detected in *A. clarkii*, it is clear that *A. clarkii* possess a higher number than all other anemonefish species.

Implications of additional *erbb3b* genes in *A. clarkii*

Interestingly, alternative splice variants encoding different isoforms have been characterized for *erbb3*. One isoform (p45-sErbB3) encoded by a 2.1 kb transcript lacks the transmembrane and cytoplasmic domains, and is secreted outside the cell to modulate the activity of the membrane-bound isoform (Chen et al. 2007; Lin et al. 2008). However, *erbb3b* gene1 and *erbb3b* gene2 do not seem to be the result of alternative splicing. We bioinformatically checked that reads mapped to each gene were not shared by other genes, and also performed PCR with the forward primers of each short form (*erbb3b* gene1 and *erbb3b* gene2) and the reverse primer of the long *erbb3b* gene3. Results from this analysis did not detect any bands (Supplementary Fig. 8), thus suggesting that *erbb3b* genes in *A. clarkii* might be tandemly duplicated genes.

Most new proteins appear to evolve from pre-existing proteins via varying degrees of modification (Andersson et al. 2015), with gene duplication certainly being a prerequisite to acquire novel functions (Zhang 2003). The widespread distribution of tandemly duplicated gene complexes supports a strong link between gene duplication and genetic novelty, but also gene expression (Zhang 2003; Rogers et al. 2017). More specifically, a study (Menghi et al. 2016) showed that ErbB receptors are recurrent tandemly duplicated genes with increased expression levels. Increased receptor expression has indeed been shown to be a main mechanism of deregulation in the ErbB signaling pathways (Roskoski 2014). There is also evidence indicating that *erbb3* increases the transcription of other genes even if the cytoplasmic pathways are not activated (i.e. truncated protein) (Bian et al. 2021). Thus, taking into account all the above, we hypothesize that the two short copies (*erbb3b* gene1 and *erbb3b* gene2) identified here could have roles in modulating the activity of the full-length protein (*erbb3b* gene3) by either (1) providing additional dimerization arms for interaction with other ErbB receptors (such as *erbb2*) and/or (2) functioning as a transcriptional co-activator of other genes necessary for the development of new melanocytes.

Finally, as *erbb3b* has been shown to be essential for the development of melanocytes and promoting adult pigmentation pattern metamorphosis in zebrafish (Budi et al. 2008), we also measured the expression of each *erbb3b* gene in melanistic and orange *A. clarkii* juveniles. Surprisingly, none of the three *erbb3b* genes were differentially expressed (Supplementary Fig. 9). Ideally, *erbb3b* expression levels would have been measured from skin tissue and not entire juveniles as the latter could dilute the expression signal. Unfortunately, skin tissue from the original adults used for genome sequencing was unavailable, and due to the size of the body, it was not possible to separate the black

and orange skin of juveniles. Although *erbb3* is robustly expressed in skin, it is also expressed in most epithelial tissues including the intestine and liver epithelium (Wieduwilt and Moasser 2008). Indeed, high expression levels of all three *erbb3b* genes in *A. clarkii* were observed in the liver, gall bladder, and intestines (Supplementary Fig. 10), which might explain the similar expression levels between black and orange juveniles. This result does not necessarily preclude a link between the additional *erbb3b* gene *A. clarkii* possesses and the melanism polymorphism of this species. Future research should endeavor to better characterize these *erbb3b* genes and investigate their implications for melanism in *A. clarkii* and other fish.

Conclusion

Here, we assembled a highly contiguous and complete chromosome-scale genome of the yellowtail clownfish *A. clarkii* using PacBio long reads and Hi-C chromatin conformation capture technologies. We annotated 25,050 protein-coding genes with 97% completeness of conserved actinopterygian genes, the highest level among anemonefish genomes available so far. Furthermore, we identified a higher number of *erbb3b* genes in *A. clarkii* compared with other anemonefish species thus suggesting a link between this gene and the natural melanism polymorphism in *A. clarkii*. 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 *A. clarkii* as an emerging model organism for molecular, ecological, developmental, and environmental studies of reef fishes.

Data availability

The genomic and transcriptomic sequencing reads have been deposited in the NCBI GenBank database under the BioProject ID: PRJNA813357. The chromosome-scale genome assembly has been deposited in the GenBank database under the accession number: JALBFV000000000. Genome assembly, annotation, proteome, and mitogenome for *A. clarkii* are also available in the Dryad Repository: <https://datadryad.org/stash/share/odvtvEuWTbDTQ43BWODojR4gFyKGlmcB199DbikJQSc>.

Supplemental material available at G3 online.

Acknowledgments

We thank Lilian Carlu for the anemonefish drawings displayed in Figs. 1 and 2.

Funding

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

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

- Andersson DI, Jerlström-Hultqvist J, Näsval J. Evolution of new functions de novo and from preexisting genes. *Cold Spring Harbor Perspect Biol.* 2015;7(6):a017996. doi:10.1101/cshperspect.a017996.
- Bell L, Moyer J, Numachi K. Morphological and genetic variation in Japanese populations of the anemonefish *Amphiprion clarkii*. *Mar Biol.* 1982;72(2):99–108. doi:10.1007/BF00396909.
- Bian C, Li R, Wen Z, Ge W, Shi Q. Phylogenetic analysis of core melanin synthesis genes provides novel insights into the molecular basis of albinism in fish. *Frontiers in Genetics.* 2021;12:707228. doi:10.3389/fgene.2021.707228.
- 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. doi:10.1038/ng.3802.
- Blum M, Chang HY, Chuguransky S, Grego T, Kandasamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 2021;49(D1):344–354. doi:10.1093/nar/gkaa977.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.
- Brüna 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):lqaa108. doi:10.1093/nargab/lqaa108.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59–60. doi:10.1038/nmeth.3176.
- Budi EH, Patterson LB, Parichy DM. Embryonic requirements for ErbB signaling in neural crest development and adult pigment pattern formation. *Development.* 2008;135(15):2603–2614. doi:10.1242/dev.019299.
- Camp EF, Hobbs JPA, De Brauwier M, Dumbrell AJ, Smith DJ. Cohabitation promotes high diversity of clownfishes in the coral triangle. *Proc R Soc B: Biol Sci.* 2016;283(1827):20160277. doi:10.1098/rspb.2016.0277.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972–1973. doi:10.1093/bioinformatics/btp348.
- Catalano KA, Dedrick AG, Stuart MR, Puritz JB, Montes HR, Jr, Pinsky ML. Quantifying dispersal variability among nearshore marine populations. *Mol Ecol.* 2021;30(10):2366–2377. doi:10.1111/mec.15732.
- Chen N, Ye X-C, Chu K, Navone NM, Sage EH, Yu-Lee L-Y, Logothetis CJ, Lin S-H. A secreted isoform of ErbB3 promotes osteonectin expression in bone and enhances the invasiveness of prostate cancer cells. *Cancer Res.* 2007;67(14):6544–6548. doi:10.1158/0008-5472.CAN-07-1330.
- Condon K. 2020. tispec: Calculates tissue specificity from RNA-seq data. <https://github.com/roonysgalbi/tispec>.
- Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics.* 2017;33(18):2938–2940. doi:10.1093/bioinformatics/btx364.
- De Brauwier M, Camp E, Jompa J, Smith DJ. High levels of heterospecific cohabitation among anemonefishes in Hoga Island, Indonesia. *Mar Biodivers.* 2016;46(1):19–20. doi:10.1007/s12526-015-0343-7.
- Djurdjević I, Furmanek T, Miyazawa S, Sušnik Bajec S. Comparative transcriptome analysis of trout skin pigment cells. *BMC Genomics.* 2019;20(1):1–15. doi:10.1186/s12864-019-5714-1.

- Emms DM, Kelly S. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):1–14. doi:10.1186/s13059-019-1832-y.
- Fautin DG. The anemonefish symbiosis: what is known and what is not. *Symbiosis.* 1991;10:23–46.
- Fautin DG, Allen GR. *Field Guide to Anemonefishes and Their Host Sea Anemones.* Perth (WA): Western Australian Museum; 1992.
- Fautin DG, Allen GR. *Anemone Fishes and Their Host Sea Anemones: A Guide for Aquarists and Divers.* Perth (WA): Western Australian Museum; 1997.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. Repeatmodeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117(17):9451–9457. doi:10.1073/pnas.1921046117.
- Gremme G, Steinbiss S, Kurtz S. Genometools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform.* 2013;10(3):645–656. doi:10.1109/TCBB.2013.68.
- Gu LH, Coulombe PA. Keratin function in skin epithelia: a broadening palette with surprising shades. *Cell Struct Dyn.* 2007;19(1):13–23. doi:10.1016/j.ceb.2006.12.007.
- Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36(9):2896–2898. doi:10.1093/bioinformatics/btaa025.
- Hanks S, Quinn A, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science.* 1988;241(4861):42–52. doi:10.1126/science.3291115.
- Hattori A. Small and large anemonefishes can coexist using the same patchy resources on a coral reef, before habitat destruction. *J Animal Ecol.* 2002;71(5):824–831. doi:10.1046/j.1365-2656.2002.00649.x.
- Hultman KA, Budi EH, Teasley DC, Gottlieb AY, Parichy DM, Johnson SL. Defects in ErbB-dependent establishment of adult melanocyte stem cells reveal independent origins for embryonic and regeneration melanocytes. *PLoS Genet.* 2009;5(7):e1000544. doi:10.1371/journal.pgen.1000544.
- Hyun-Soo C, Leahy Daniel J. Structure of the extracellular region of HER3 reveals an interdomain tether. *Science.* 2022;297:1330–1333. doi:10.1126/science.1074611.
- Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z. Getorganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 2020;21(1):241. doi:10.1186/s13059-020-02154-5.
- Jura N, Shan Y, Cao X, Shaw DE, Kuriyan J. Structural analysis of the catalytically inactive kinase domain of the human EGF receptor 3. *Proc Natl Acad Sci U S A.* 2009;106(51):21608–21613. doi:10.1073/pnas.0912101106.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772–780. doi:10.1093/molbev/mst010.
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28:1647–1649. doi:10.1093/bioinformatics/bts199.
- 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. doi:10.1038/s41587-019-0201-4.
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019;37(5):540–546. doi:10.1038/s41587-019-0072-8.
- Kronenberg ZN, Hall RJ, Hiendleder S, Smith TP, Sullivan ST, Williams JL, Kingan SB. FALCON-Phase: integrating PacBio and Hi-C data for phased diploid genomes. *BioRxiv* 327064. <https://doi.org/10.1101/327064>, 2018, preprint: not peer reviewed.
- Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform.* 2017;18(2):205–214. doi:10.1093/bib/bbw008.
- Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front Zool.* 2014;11(1):1–8. doi:10.1186/s12983-014-0081-x.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–359. doi:10.1038/nmeth.1923.
- Laudet V, Ravasi T. *Evolution, Development and Ecology of Anemonefishes: Model Organisms for Marine Science.* Boca Raton (FL): CRC Press; 2022.
- Lehmann R, Lightfoot DJ, Schunter C, Michell CT, Ohyanagi H, Mineta K, Foret S, 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. doi:10.1111/1755-0998.12939.
- 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):293–296. doi:10.1093/nar/gkab301.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
- Li Q, Yuan Z, Cao B. The function of human epidermal growth factor receptor-3 and its role in tumors. *Oncol Rep.* 2013;30(6):2563–2570. doi:10.3892/or.2013.2754.
- Lin S-H, Cheng C-J, Lee Y-C, Ye X, Tsai W-W, Kim J, Pasqualini R, Arap W, Navone NM, Tu S-M, et al. A 45-kDa ErbB3 secreted by prostate cancer cells promotes bone formation. *Oncogene.* 2008;27(39):5195–5203. doi:10.1038/onc.2008.156.
- Litsios G, Pearman PB, Lanterbecq D, Tolou N, Salamin N. The radiation of the clownfishes has two geographical replicates. *J Biogeogr.* 2014;41(11):2140–2149. doi:10.1111/jbi.12370.
- Litsios G, Salamin N. Hybridisation and diversification in the adaptive radiation of clownfishes. *BMC Evol Biol.* 2014;14(1):1–9. doi:10.1186/s12862-014-0245-5.
- Lorin T, Brunet FG, Laudet V, Volff JN. Teleost fish-specific preferential retention of pigmentation gene-containing families after whole genome duplications in vertebrates. *G3 (Bethesda).* 2018;8(5):1795–1806. doi:10.1534/g3.118.200201.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27(6):764–770. doi:10.1093/bioinformatics/btr011.
- Marcionetti A, Rossier 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. doi:10.1111/1755-0998.12772.
- 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 Biol Evol.* 2019;11(3):869–882. doi:10.1093/gbe/evz042.
- Menghi F, Inaki K, Woo X, Kumar PA, Grzeda KR, Malhotra A, Yadav V, Kim H, Marquez EJ, Ucar D, et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci U S A.* 2016;113(17):2373–2382. doi:10.1073/pnas.1520010113.
- Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics.* 2018;34(13):42–50. doi:10.1093/bioinformatics/bty266.
- Militz TA, McCormick MI, Schoeman DS, Kinch J, Southgate PC. Frequency and distribution of melanistic morphs in coexisting

- population of nine clownfish species in Papua New Guinea. *Mar Biol.* 2016;163(10):200. doi:10.1007/s00227-016-2972-1.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: the protein families database in 2021. *Nucleic Acids Research.* 2021;49:D412–D419. doi: 10.1093/nar/gkaa913.
- Mogensen MM, Henderson CG, Mackie JB, Lane EB, Garrod DR, Tucker JB. Keratin filament deployment and cytoskeletal networking in a sensory epithelium that vibrates during hearing. *Cell Motil.* 1998;41(2):138–153. doi:10.1002/(SICI)1097-0169(1998)41:2<138::AID-CM5>3.0.CO;2-A.
- Moyer JT. Geographical variation and social dominance in Japanese populations of the anemonefish *Amphiprion clarkii*. *Jpn J Ichthyol.* 1976;23(1):12–22. doi:10.11369/jji1950.23.12.
- Moyer JT. Influence of temperate waters on the behavior of the tropical anemonefish *Amphiprion clarkii* at Miyake-jima, Japan. *Bull Mar Sci.* 1980;30(1):261–272. doi:10.2307/1444899.
- Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;11(9):1650–1667. doi:10.1038/nprot.2016.095.
- Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinform.* 2014;47(1):11–12. doi:10.1002/0471250953.bi1112s47.
- Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol.* 2009;5(12):e1000598. doi:10.1371/journal.pcbi.1000598.
- Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 2020;21(1):1–27. doi:10.1186/s13059-020-02134-9.
- Rogers RL, Shao L, Thornton KR. Tandem duplications lead to novel expression patterns through exon shuffling in *Drosophila yakuba*. *PLoS Genet.* 2017;13(5):e1006795–e1006795. doi:10.1371/journal.pgen.1006795.
- Roskoski R, Jr. ErbB/HER protein-tyrosine kinases: structures and small molecule inhibitors. *Pharmacol Res.* 2014;87:42–59. doi: 10.1016/j.phrs.2014.06.001.
- Roux N, Miura S, Dussene M, Tara Y, Lee F, de Bernard S, Reynaud M, Salis P, Barua A, Boulahtouf A, et al. The multi-level regulation of clownfish metamorphosis by thyroid hormones. *BioRxiv.* 2022.03.04.482938. doi: 10.1101/2022.03.04.482938, 2022, preprint: not peer reviewed.
- Roux N, Salis P, Lee SH, Besseau L, Laudet V. Anemonefish, a model for Eco-Evo-Devo. *EvoDevo.* 2020;11(1):20. doi:10.1186/s13227-020-00166-7.
- Ryu T, Herrera M, Moore B, Izumiyama M, Kawai E, Laudet V, Ravasi T. A chromosome-scale genome assembly of the false clownfish, *Amphiprion ocellaris*. *G3 (Bethesda).* 2022;12(5):jkac074. doi:10.1093/g3journal/jkac074.
- Salem M, Paneru B, Al-tobasei R, Abdouni F, Thorgaard GH, Rexroad CE, Yao J. Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS ONE.* 2015; 10(3):e0121778. doi:10.1371/journal.pone.0121778.
- Salis P, Roux N, Huang D, Marcionetti A, Mougnot P, Reynaud M, Salles O, Salamin N, Pujol B, Parichy DM, et al. Thyroid hormones regulate the formation and environmental plasticity of white bars in clownfishes. *Proc Natl Acad Sci U S A.* 2021;118(23): e2101634118. doi:10.1073/pnas.2101634118.
- Salis P, Roux N, Soulat O, Lecchini D, Laudet V, Frédérick B. Ontogenetic and phylogenetic simplification during white stripe evolution in clownfishes. *BMC Biol.* 2018;16(1):90. doi:10.1186/s12915-018-0559-7.
- 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. *Mol Biol Evol.* 2018;35(6):1553–1555. doi:10.1093/molbev/msy074.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31(19): 3210–3212. doi:10.1093/bioinformatics/btv351.
- Ståhlberg A, Kubista M, Pfaffl M. Comparison of reverse transcriptases in gene expression analysis. *Clin Chem.* 2004;50(9): 1678–1680. doi:10.1373/clinchem.2004.035469.
- Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30(9): 1312–1313. doi:10.1093/bioinformatics/btu033.
- Stein RA, Staros JV. Insights into the evolution of the ErbB receptor family and their ligands from sequence analysis. *BMC Evol Biol.* 2006;6(1):79. doi:10.1186/1471-2148-6-79.
- Storer J, Hubble R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* 2021;12(1):1–14. doi:10.1186/s13100-020-00230-y.
- Stuckert AM, Chouteau M, McClure M, LaPolice TM, Linderoth T, Nielsen R, Summers K, MacManes MD. The genomics of mimicry: gene expression throughout development provides insights into convergent and divergent phenotypes in a Müllerian mimicry system. *Mol Ecol.* 2021;30(16):4039–4061. doi:10.1111/mec.16024.
- 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.* 2018;7(3):gix137. doi:10.1093/gigascience/gix137.
- Tang KL, Stiassny MLJ, Mayden RL, DeSalle R. Systematics of damselfishes. *Ichthyol Herpetol.* 2021;109(1):258–318. doi:10.1643/i2020105.
- Tao Y, Li J-L, Liu M, Hu X-Y. Complete mitochondrial genome of the yellowtail clownfish *Amphiprion clarkii* (Pisces: Perciformes, Pomacentridae). *Mitochondrial DNA Part A.* 2016;27(1):326–327. doi:10.3109/19401736.2014.892100.
- Tempel S. Using and understanding RepeatMasker. *Methods Mol Biol.* 2012;859:29–51. doi:10.1007/978-1-61779-603-6_2.
- Thongtam Na Ayudhaya P, Areesirisuk P, Singchat W, Sillapaprayoon S, Muangmai N, Peyachoknagul S, Srikulnath K. Complete mitochondrial genome of 10 anemonefishes belonging to *Amphiprion* and *Premnas*. *Mitochondrial DNA Part B.* 2019; 4(1):222–224. doi:10.1080/23802359.2018.1546145.
- UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49(D1):480–489. doi:10.1093/nar/gkaa1100.
- Uttam J, Hutton E, Coulombe PA, Anton-Lamprecht I, Yu QC, Gedde-Dahl T, Jr, Fine JD, Fuchs E. The genetic basis of epidermolysis bullosa simplex with mottled pigmentation. *Proc Natl Acad Sci U S A.* 1996;93(17):9079–9084. doi:10.1073/pnas.93.17.9079.
- 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. doi:10.1093/bioinformatics/btx153.
- 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. doi:10.1371/journal.pone.0112963.

Wieduwilt MJ, Moasser MM. The epidermal growth factor receptor family: biology driving targeted therapeutics. *Cell Mol Life Sci.* 2008;65(10):1566–1584. doi:10.1007/s00018-008-7440-8.

Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17(9):847–848. doi:10.1093/bioinformatics/17.9.847.

Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol.* 2003;18(6):292–298. doi:10.1016/S0169-5347(03)00033-8.

Communicating editor: J. J. Emerson