

## DATA NOTE

# Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly

Mun Hua Tan<sup>1,2,3,†</sup>, Christopher M. Austin<sup>1,2,3,†</sup>, Michael P. Hammer<sup>4</sup>, Yin Peng Lee<sup>2,3</sup>, Laurence J. Croft<sup>1,5</sup> and Han Ming Gan<sup>1,2,3,\*</sup>

<sup>1</sup>Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria 3220, Australia, <sup>2</sup>Genomics Facility, Tropical Medicine and Biology Platform, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500, Petaling Jaya, Selangor, Malaysia, <sup>3</sup>School of Science, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500, Petaling Jaya, Selangor, Malaysia, <sup>4</sup>Museum and Art Gallery of the Northern Territory, Darwin 0801, Australia and <sup>5</sup>Malaysian Genomics Resource Centre Berhad, Mid Valley City 59200, Kuala Lumpur, Malaysia

\*Correspondence address: Han Ming Gan, PhD, Building Ka, Level 4, Room 4.338, Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Waurn Ponds, Victoria 3216, Australia. Tel: +61-490786277; Fax: +61-352272017; E-mail: [han.gan@deakin.edu.au](mailto:han.gan@deakin.edu.au)

†Equal contribution

## Abstract

**Background:** Some of the most widely recognized coral reef fishes are clownfish or anemonefish, members of the family Pomacentridae (subfamily: Amphiprioninae). They are popular aquarium species due to their bright colours, adaptability to captivity, and fascinating behavior. Their breeding biology (sequential hermaphrodites) and symbiotic mutualism with sea anemones have attracted much scientific interest. Moreover, there are some curious geographic-based phenotypes that warrant investigation. Leveraging on the advancement in Nanopore long read technology, we report the first hybrid assembly of the clown anemonefish (*Amphiprion ocellaris*) genome utilizing Illumina and Nanopore reads, further demonstrating the substantial impact of modest long read sequencing data sets on improving genome assembly statistics. **Results:** We generated 43 Gb of short Illumina reads and 9 Gb of long Nanopore reads, representing approximate genome coverage of 54× and 11×, respectively, based on the range of estimated k-mer-predicted genome sizes of between 791 and 967 Mbp. The final assembled genome is contained in 6404 scaffolds with an accumulated length of 880 Mb (96.3% BUSCO-calculated genome completeness). Compared with the Illumina-only assembly, the hybrid approach generated 94% fewer scaffolds with an 18-fold increase in N<sub>50</sub> length (401 kb) and increased the genome completeness by an additional 16%. A total of 27 240 high-quality protein-coding genes were predicted from the clown anemonefish, 26 211 (96%) of which were annotated functionally with information from either sequence homology or protein signature searches. **Conclusions:** We present the first genome of any anemonefish and demonstrate the value of low coverage (~11×) long Nanopore read sequencing in improving both genome assembly contiguity and completeness. The near-complete assembly of the *A. ocellaris* genome will be an invaluable molecular resource for supporting a range of genetic, genomic, and phylogenetic

Received: 14 November 2017; Revised: 11 December 2017; Accepted: 27 December 2017

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

studies specifically for clownfish and more generally for other related fish species of the family Pomacentridae.

**Keywords:** clownfish; long reads; genome; transcriptome; hybrid assembly

## Data Description

The clown anemonefish, *Amphiprion ocellaris* (Fig. 1, NCBI Taxon ID: 80 972, Fish Base ID:6509), is a well-known tropical marine fish species among the nonscientific community especially following the Pixar film *Finding Nemo* and its sequel *Finding Dory* [1]. The visual appeal of *A. ocellaris* due to its bright coloration and behaviour and ease of husbandry have maintained a strong global demand for this species in the marine aquarium trade, driving a fine balance between positive environmental awareness and sustainable ornamental use [1, 2]. Further, given high survival rates and ability to complete their life cycle in captivity, captive-breeding programs to partially sustain their global trade have been successful [3]. For the scientific community, *A. ocellaris* or anemonefishes in general are actively studied due to their intriguing reproductive strategy, i.e., sequential hermaphroditism [4–7] and mutualistic relationships with sea anemones [8–12]. Phenotypic body colour variation based on host-anemone use and geography also pose additional questions regarding adaptive genetic variation [13].

In recent years, concurrent with the advent of long read sequencing technologies [14], several studies have explored combining short but accurate Illumina reads with long but less accurate Nanopore/PacBio reads to obtain genome assemblies that are usually more contiguous with higher completeness than assemblies based on Illumina-only reads [15–19]. To further contribute to the evaluation of long read technology in fish genomics [15], we sequenced the whole genome of *A. ocellaris* using Oxford Nanopore and Illumina technologies and demonstrate that hybrid assembly of long and short reads greatly improved the quality of genome assembly.

## Whole-genome sequencing

Tissues for genome assembly and as reference material were sourced from the collection of the Museum and Art Gallery of the Northern Territory (NTM). The samples used for DNA extraction and subsequent whole-genome sequencing were from freshly vouchered captive bred *A. ocellaris* specimens, representing a unique black and white colour phenotype found only in the Darwin Harbour region, Australia (NTM A3764, A4496, A4497).

Genomic DNA was extracted from multiple fin clip and muscle samples using the E.Z.N.A. Tissue DNA Kit (Omega Bio-tek, Norcross, GA, USA). For Illumina library prep, approximately 1 µg of gDNA from isolate A3764 was sheared to 300 bp using a Covaris Focused-Ultrasonicator (Covaris, Woburn, MA, USA) and subsequently processed using the TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Paired-end sequencing was performed on a single lane of HiSeq 2000 (Illumina, San Diego, CA, USA) located at the Malaysian Genomics Resource Centre Berhad. Two additional libraries were constructed from specimen NTM A3764, and both libraries were sequenced on the MiSeq (2 × 300 bp setting), located at the Monash University Malaysia Genomics Facility.

To generate Oxford Nanopore long reads, approximately 5 µg of gDNA was extracted from isolates NTM A4496 and A4497,



**Figure 1:** The clown anemonefish (*Amphiprion ocellaris*). Photo by Michael P. Hammer.

size-selected (8–30 kb) with a BluePippin (Sage Science, Beverly, MA, USA), and processed using the Ligation Sequencing 1D Kit (Oxford Nanopore, Oxford, UK) according to the manufacturer's instructions. Three libraries were prepared and sequenced on 3 different R9.4 flowcells using the MiniION portable DNA sequencer (Oxford Nanopore, Oxford, UK) for 48 hours.

## Sequence read processing

Raw Illumina short reads were adapter-trimmed with Trimmomatic v.0.36 (ILLUMINACLIP:2:30:10, MINLEN:100; Trimmomatic, RRID:SCR\_011848) [20], followed by a screening for vectors and contaminants, using Kraken v.0.10.5 (Kraken, RRID:SCR\_005484) [21] based on the MiniKraken DB. Kraken-unclassified reads, i.e., nonmicrobial/viral origin, were aligned to the complete mitogenome of NTM A3764 (see the Mitogenome Assembly section) to exclude sequences of organellar origin. This results in a total of 42.35 Gb of “clean” short reads. Nanopore reads were base-called from their raw FAST5 files using the Oxford Nanopore proprietary base-caller, Albacore, version 2.0.1. Applying a minimum length cutoff of 500 bp, this study produced a total of 8.95 Gbp in 895 672 Nanopore reads ( $N_{50}$ : 12.7 kb). Sequencing statistics are available in Supplementary Table 1.

## Genome size estimation

K-mer counting with the “clean” Illumina reads was performed with Jellyfish v.2.2.6 (Jellyfish, RRID:SCR\_005491) [22], generating k-mer frequency distributions of 17-, 21-, and 25-mers. These histograms were processed by GenomeScope [23], which estimated a genome size of 791 to 794 Mbp with approximately 80% of unique content and a heterozygosity level of 0.6% (Supplementary Fig. 1). Given that we had previously excluded adapters as well as sequences from contaminant or organellar sources, the max kmer coverage filter was not applied (*max kmer coverage*: -1). A separate estimation performed by BBMap [24] estimated a haploid genome size of 967 Mbp. The genome sizes estimated from both approaches are within the range of sizes listed for

other *Amphiprion* species (792 Mb–1.2 Gb) as reported on the Animal Genome Size Database [25].

## Hybrid genome assembly

Short reads used for assemblies described in this study were only trimmed for adapters, but not for quality. Both short-read-only and hybrid *de novo* assemblies were performed with the Maryland Super-Read Celera Assembler v.3.2.2 (MaSuRCA, [RRID:SCR.010691](#)) [26]. During hybrid assembly, errors were encountered in the fragment correction step of the Celera Assembler (CA; Celera assembler, [RRID:SCR.010750](#)). To overcome this, given that the CA assembler is no longer maintained, we disabled the *frgcorr* step based on one of the developer's recommendations, and the hybrid assembly was subsequently improved with 10 iterations of Pilon v.1.22 (Pilon, [RRID:SCR.014731](#)) [27], using short reads to correct bases, fix misassemblies, and fill assembly gaps. To assess the completeness of the genome, Benchmarking Universal Single-Copy Orthologs v.3.0.2 (BUSCO, [RRID:SCR.015008](#)) [28] was used to locate the presence or absence of the Actinopterygii-specific set of 4584 single-copy orthologs (OrthoDB v9).

The short-read-only and hybrid assemblies yielded total assembly sizes of 851 Mb and 880 Mb, respectively. Statistics for assemblies for each Pilon iteration are available in Supplementary Table 2. Inclusion of Nanopore long reads for a hybrid assembly representing approximately  $\times 11$  genome coverage led to a 94% decrease in the number of scaffolds (>500 bp) from 106 526 to 6404 scaffolds and an 18-fold increase in the scaffold  $N_{50}$  length from 21 802 bp to 401 715 bp (Table 1). In addition, the genome completeness was also substantially improved in the hybrid assembly, with BUSCO detecting complete sequences of 96.3% (4417/4584) of single-copy orthologs in the Actinopterygii-specific dataset.

## Transcriptome sequencing and assembly

Total RNA extraction from RNashield-preserved whole-body and muscle tissues of isolate A4496 used Quick-RNA MicroPrep (Zymo Research Corpt, Irvine, CA, USA) according to the manufacturer's protocols. After assessing total RNA intactness on the TapeStation2100 (Agilent), mRNA was enriched using NEB-Next Poly(A) mRNA Magnetic Isolation Kit (NEB, Ipswich, MA, USA) and processed with NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA). Libraries from both whole-body and muscle tissues were sequenced on a fraction of MiSeq V3 flowcell ( $1 \times 150$  bp). Single-end reads from both libraries in addition to 2 publicly available *A. ocellaris* transcriptome sequencing data (SRR5253145 and SRR5253146, Bioproject ID: PRJNA374650) were individually assembled using Scallop v0.10.2 [29] based on HiSat2 [30] alignment of RNA-sequencing reads to the newly generated *A. ocellaris* genome. The transcriptome assemblies were subsequently merged using the tr2aacds pipeline from the EvidentialGene [31] package and similarly assessed for completeness using BUSCO, version 3 [28]. The final nonredundant transcriptome assembly, which was subsequently used to annotate the *A. ocellaris* genome, contains 25 264 contigs/isotigs (putative transcripts) with an accumulated length of 68.4 Mb and BUSCO-calculated completeness of 92.8% (Table 1).

## Genome annotation

Protein-coding genes were predicted with the MAKER v.2.31.9 genome annotation pipeline (MAKER, [RRID:SCR.005309](#)) [32]. A

**Table 1:** Genome and transcriptome statistics of the clownfish (*Amphiprion ocellaris*) genome

	Illumina ( $\geq 500$ bp)	Illumina + Nanopore ( $\geq 500$ bp)
<b>Genome assembly</b>		
Contig statistics		
Number of contigs	133 997	7810
Total contig size, bp	851 389 851	880 159 068
Contig $N_{50}$ size, bp	15 458	323 678
Longest contig, bp	204 209	2051 878
Scaffold statistics		
Number of scaffolds	106 526	6404
Total scaffold size, bp	852 602 726	880 704 246
Scaffold $N_{50}$ size, bp	21 802	401 715
Longest scaffold, bp	227 111	3111 502
GC/AT/N, %	39.6/60.2/0.14	39.4/60.5/0.06
BUSCO genome completeness		
Complete	3691 (80.5%)	4417 (96.3%)
Complete and single copy	3600 (78.5%)	4269 (93.1%)
Complete and duplicated	91 (2.0%)	148 (3.2%)
Fragmented	534 (11.6%)	63 (1.4%)
Missing	359 (7.9%)	104 (2.3%)
<b>Transcriptome assembly</b>		
Number of contigs	25 364	
Total length, bp	68 405 796	
Contig $N_{50}$ size, bp	3670	
BUSCO completeness		
Complete	4253 (92.8%)	
Complete and single-copy	4128 (90.1%)	
Complete and duplicated	125 (2.7%)	
Fragmented	127 (2.8%)	
Missing	204 (4.4%)	
<b>Genome annotation</b>		
Number of protein-coding genes	27 420	
Number of functionally annotated proteins	26 211	
Mean protein length	514 aa	
Longest protein	29 084 aa (titin protein)	
Average number (length) of exon per gene	9 (355 bp)	
Average number (length) of intron per gene	8 (1532 bp)	

total of 3 passes were run with MAKER2; the first pass was based on hints from the assembled transcripts as RNA-seq evidence (*est2genome*) and protein sequences from 11 fish species downloaded from Ensembl (Ensembl, [RRID:SCR.002344](#)) [33] (*protein2genome*), whereas the second and third passes included gene models trained from the first (and then second) passes with *ab initio* gene predictors SNAP (SNAP, [RRID:SCR.002127](#)) [34] and Augustus (Augustus: Gene Prediction, [RRID:SCR.008417](#)) [35]. In the final set of genes predicted, sequences with annotation edit distance (AED) values of less than 0.5 were retained. A small AED value suggests a lesser degree of difference between the predicted protein and the evidence used in the prediction (i.e., fish proteins, transcripts). This resulted in a final set of 27 240 protein-coding genes with an average AED of 0.14 (Table 1). A BUSCO analysis on the completeness of the predicted protein dataset detected the presence of 4259 (92.9%) single-copy orthologs from the Actinopterygii-specific dataset.

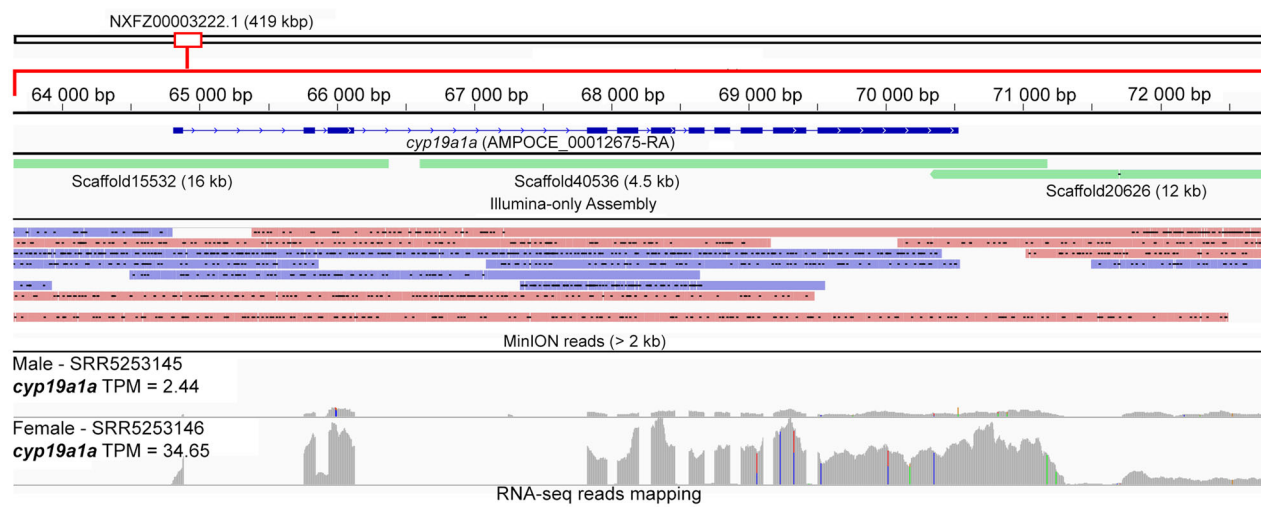


Figure 2: Mapping of MinION long reads, Illumina-assembled scaffolds, and RNA-sequencing reads of male and female *A. ocellaris* to the genomic region containing the *cyp19a1a* gene. Transcripts per million (TPM) values were calculated using Kallisto, version 0.43.1 [46].

Further, to infer the putative function of these predicted proteins, NCBI's *blastp* v.2.6.0 (-*evalue* 1e-10, -*seg* yes, -*soft\_masking* true, -*lcase\_masking*; BLASTP, [RRID:SCR.001010](#)) [36] was used to find homology to existing vertebrate sequences in the nonredundant (NR) database. Applying a hit fraction filter to include only hits with  $\geq 70\%$  target length fraction, the remaining unannotated sequences were subsequently aligned to all sequences in the NR database. With this method, 20 107 proteins (74%) were annotated with a putative function based on homology. Additionally, InterProScan v.5.26.65 (InterProScan, [RRID:SCR.005829](#)) [37] was used to examine protein domains, signatures, and motifs present in the predicted protein sequences. This analysis detected domains, signatures, or motifs for 26 211 proteins (96%). Overall, 96% of the predicted clownfish protein-coding genes were functionally annotated with information from at least 1 of the 2 approaches.

### Mitogenome recovery via genome skimming

Genome skimming [38, 39] was performed on 3 additional *A. ocellaris* individuals from known localities (Supplementary Table 3). Mitogenome assembly was performed with MITObim, version 1.9 (MITObim, [RRID:SCR.015056](#)) [40], using the complete mitogenome of *A. ocellaris* (GenBank: NC009065.1) as the bait for read mapping. The assembled mitogenomes were subsequently annotated with MitoAnnotator [41]. Consistent with the original broodstock collection from northern Australia, the captive-bred black and white *A. ocellaris* NTM A3764 exhibits strikingly high whole-mitogenome nucleotide identity (99.98%) to sample NTM A3708 as a wild collection from Darwin Harbour, Australia. In addition, the overall high pair-wise nucleotide identity ( $> 98\%$ ) of NTM A3764 to newly generated and publicly available *A. ocellaris* whole mitogenomes further supports its morphological identification as *A. ocellaris* (Supplementary Table 3).

### Identification of the *cyp19a1a* gene associated with sexual differentiation

The validated *cyp19a1a* enzyme of *Danio rerio* (Uniprot: O42145) was used as the query (*E*-value = 1e-10) for *blastp* search against the predicted *A. ocellaris* proteins. The top *blast* hit,

AMPOCE\_00 012675-RA (71.5% protein identity to O42145), was searched (tblastn) against the NCBI TSA database (Taxon: *Amphiprion*) and showed strikingly high protein identity (99%) to a translated RNA transcript from *Amphiprion bicinctus* (c183337\_g1.i2: GDCV01327693) [5]. The *cyp19a1a* gene codes for a steroidogenic enzyme that converts androgens into estrogens [42] were recently shown to be instrumental during sex change in *Amphiprion bicinctus*, as evidenced by significant correlation and differential expression of this gene between males and mature females [5]. We also observed a similar profile based on mapping of RNA reads from the publicly available male and female transcriptomes of *A. ocellaris* to the *cyp19a1a* gene region as visualized using the Integrative Genomics Viewer (Fig. 2) [43]. The *A. ocellaris cyp19a1a* gene is located on a 419-kb scaffold and is spanned by multiple Minimap2-aligned Nanopore reads [44]. It is noteworthy that in the Illumina-only assembly, this gene is fragmented and located on 3 relatively short scaffolds (Fig. 2).

### Conclusion

We present the first clownfish genome co-assembled with high-coverage Illumina short reads and low-coverage ( $\sim 11\times$ ) Nanopore long reads. Hybrid assembly of Illumina and Nanopore reads is one of the new features of the MaSuRCA assembler, version 3.2.2, which works by constructing long and accurate mega-reads from the combination of long and short read data. Although this is a relatively computationally intensive strategy with long run times, we observed substantial improvement in the genome statistics when compared with Illumina-only assembly. As Nanopore technology becomes more mature, it is likely that future *de novo* genome assembly will shift toward high-coverage long read-only assembly, followed by multiple iterations of genome polishing using Illumina reads.

### Availability of supporting data

Data supporting the results of this article are available in the GigaDB repository [45]. Raw Illumina and Nanopore reads generated in this study are available in the Sequence Read Archive (SRP123679), whereas the Whole Genome Shotgun project has

been deposited at DDBJ/EMBL/GenBank under the accession NXFZ00000000, both under BioProject PRJNA407816.

## Abbreviations

bp: base pair; CDS: coding sequence; Gb: giga base; kb: kilo base; Mb: mega base; SRA: Sequence read archive; TE: transposable elements; TSA: transcriptome shotgun assembly.

## Additional files

Additional file 1: Figure S1: Genome profiling of *A. ocellaris* based on Illumina short reads.

Additional file 1: Table S1: Summary of raw reads generated from genome and transcriptome sequencing.

Additional file 1: Table S2: Assembly details after each pilon iteration.

Additional file 1: Table S3: Mitogenome similarity of *Amphiprion ocellaris* between the target sample (NTM A3764) and other isolates with known locality; body-colour phenotype is marked where known.

## Competing interests

The authors declare that they have no competing interests.

## Funding

This study was funded by the Monash University Malaysia Tropical and Biology Multidisciplinary Platform.

## References

- Militz TA, Foale S. The “Nemo Effect”: perception and reality of Finding Nemo’s impact on marine aquarium fisheries. *Nat Biotechnol* 2017;18(3):525–7.
- Madduppa HH, von Juterzenka K, Syakir M et al. Socio-economy of marine ornamental fishery and its impact on the population structure of the clown anemonefish *Amphiprion ocellaris* and its host anemones in Spermonde Archipelago, Indonesia. *Ocean Coast Manag* 2014;100(Supplement C):41–50.
- Hall H, Warmolts D. The role of public aquariums in the conservation and sustainability of the marine ornamentals trade. In: Cato JC, Brown CL, eds. *Marine Ornamental Species*. Ames, Iowa, USA: Blackwell Publishing Company; 2008:305–24.
- Madhu R, Madhu K, Rethesh T. Life history pathways in false clown *Amphiprion ocellaris* Cuvier, 1830: a journey from egg to adult under captive condition. *J Marine Biol Assoc India* 2012;54(1):77–90.
- Casas L, Saborido-Rey F, Ryu T et al. Sex change in clownfish: molecular insights from transcriptome analysis. *Sci Rep* 2016;6:35461.
- Buston P. Social hierarchies: size and growth modification in clownfish. *Nature* 2003;424(6945):145–6.
- Kobayashi Y, Horiguchi R, Miura S et al. Sex- and tissue-specific expression of P450 aromatase (*cyp19a1a*) in the yellowtail clownfish, *Amphiprion clarkii*. *Comp Biochem Physiol A Mol Integr Physiol* 2010;155(2):237–44.
- Davenport D, Norris KS. Observations on the symbiosis of the sea anemone *Stoichactis* and the pomacentrid fish, *Amphiprion percula*. *Biol Bull* 1958;115(3):397–410.
- Arvedlund M, Nielsen LE. Do the anemonefish *Amphiprion ocellaris* (Pisces: Pomacentridae) imprint themselves to their host sea anemone *Heteractis magnifica* (Anthozoa: Actinidae)? *Ethology* 1996;102(2):197–211.
- Mariscal RN. An experimental analysis of the protection of *Amphiprion xanthurus* Cuvier & Valenciennes and some other anemone fishes from sea anemones. *J Exp Marine Biol Ecol* 1970;4(2):134–49.
- Hattori A. Coexistence of two anemonefishes, *Amphiprion clarkii* and *A. perideraion*, which utilize the same host sea anemone. *Environ Biol Fish* 1995;42(4):345–53.
- Schmiege PF, D’Aloia CC, Buston PM. Anemonefish personalities influence the strength of mutualistic interactions with host sea anemones. *Marine Biol* 2017;164(1):24.
- Allen GR. *Damselfishes of the World*. Melle, Germany: Mergus Publishers; 1991.
- Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics* 2016;107(1):1–8.
- Austin CM, Tan MH, Harrison KA et al. De novo genome assembly and annotation of Australia’s largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. *Gigascience* 2017;6(8):1–6.
- Gan HM, Lee YP, Austin CM. Nanopore long-read guided complete genome assembly of *Hydrogenophaga intermedia*, and genomic insights into 4-aminobenzenesulfonate, p-aminobenzoic acid and hydrogen metabolism in the genus *Hydrogenophaga*. *Front Microbiol* 2017;8:1880.
- Zimin AV, Puiu D, Hall R et al. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* 2017;6(11):1–7.
- Zimin AV, Stevens KA, Crepeau MW et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience* 2017;6(1):1–4.
- Zimin AV, Puiu D, Luo M-C et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. *Genome Res* 2017;27(5):787–92.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114–20.
- Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15(3):R46.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27(6):764–70.
- Vurture GW, Sedlazeck FJ, Nattestad M et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 2017;33(14):2202–4.
- Bushnell B. *BBMap Short Read Aligner*. Berkeley, CA: University of California; 2016. <http://sourcefor-genet/projects/bbmap> (15 June 2017, date last accessed).
- <http://www.genomesize.com> (15 June 2017, date last accessed).
- Zimin AV, Marçais G, Puiu D et al. The MaSuRCA genome assembler. *Bioinformatics* 2013;29(21):2669–77.
- Walker BJ, Abeel T, Shea T et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;9(11):e112963.
- Simão FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31(19):3210–2.

29. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat Biotechnol* 2017; doi:10.1038/nbt.4020. <https://www.nature.com/articles/nbt.4020#supplementary-information> (15 June 2017, date last accessed).
30. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;12:357. <https://www.nature.com/articles/nmeth.3317#supplementary-information> (15 June 2017, date last accessed).
31. Gilber D. Gene-omes built from mRNA seq not genome DNA. *F1000Res* 2016;5(1695):1.
32. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 2011;12(1):491.
33. Hubbard T, Barker D, Birney E et al. The Ensembl genome database project. *Nucleic Acids Res* 2002;30(1):38–41.
34. Korf I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;5(1):59.
35. Stanke M, Steinkamp R, Waack S et al. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 2004;32(suppl-2):W309–12.
36. Boratyn GM, Camacho C, Cooper PS et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 2013;41(W1):W29–33.
37. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 2001;17(9):847–8.
38. Gan HM, Schultz MB, Austin CM. Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. *BMC Evol Biol* 2014;14(1):19.
39. Grandjean F, Tan MH, Gan HM et al. Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish. *Zool Scripta* 2017; doi:10.1111/zsc.12247.
40. Hahn C, Bachmann L, Chevreur B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res* 2013;41(13):e129.
41. Iwasaki W, Fukunaga T, Isagozawa R et al. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol Biol Evol* 2013;30(11):2531–40.
42. Kallivretaki E, Eggen R, Neuhauss S et al. Aromatase in zebrafish: a potential target for endocrine disrupting chemicals. *Mar Environ Res* 2006;62(90):7.
43. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics* 2013;14(2):178–92.
44. Li H. Minimap2: fast pairwise alignment for long nucleotide sequences. arXiv 2017. <https://arxiv.org/abs/1708.01492> (15 June 2017, date last accessed).
45. Tan MH, Austin CM, Hammer MP et al. Supporting data for “Finding Nemo: hybrid assembly with oxford nanopore and illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly.” *GigaScience Database* 2017. <http://dx.doi.org/10.5524/100397> (15 June 2017, date last accessed).
46. Bray NL, Pimentel H, Melsted P et al. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* 2016;34(5):525–7. <http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html#supplementary-information> (15 June 2017, date last accessed).