GigaScience, 6, 2017, 1–6

doi: 10.1093/gigascience/gix063
Advance Access Publication Date: 19 July 2017
Data Note

DATA NOTE

De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read

Christopher M. Austin[1,2,3,†], Mun Hua Tan[1,2,3,†], Katherine A. Harrisson[4], Yin Peng Lee[2,3], Laurence J. Croft[3,5], Paul Sunnucks[4], Alexandra Pavlova[4] and Han Ming Gan[1,2,3,*]

[1]Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria 3220, Australia, [2]Genomics Facility, Tropical Medicine and Biology Platform, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500, Petaling Jaya, Selangor, Malaysia, [3]School of Science, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500, Petaling Jaya, Selangor, Malaysia, [4]School of Biological Sciences, Monash University, Clayton Campus, Clayton, Victoria, Australia and [5]Malaysian Genomics Resource Centre Berhad, Boulevard Signature Office, Kuala Lumpur, Malaysia

*Correspondence address. Han Ming Gan, Building Ka, Level 4, Room 4.338, Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria 3220, Australia. Tel: +61-490786277; E-mail: han.gan@deakin.edu.au
†Equal contribution.

Abstract

One of the most iconic Australian fish is the Murray cod, *Maccullochella peelii* (Mitchell 1838), a freshwater species that can grow to ~1.8 metres in length and live to age ≥48 years. The Murray cod is of a conservation concern as a result of strong population contractions, but it is also popular for recreational fishing and is of growing aquaculture interest. In this study, we report the whole genome sequence of the Murray cod to support ongoing population genetics, conservation, and management research, as well as to better understand the evolutionary ecology and history of the species. A draft Murray cod genome of 633 Mbp (N_{50} = 109 974bp; BUSCO and CEGMA completeness of 94.2% and 91.9%, respectively) with an estimated 148 Mbp of putative repetitive sequences was assembled from the combined sequencing data of 2 fish individuals with an identical maternal lineage; 47.2 Gb of Illumina HiSeq data and 804 Mb of Nanopore data were generated from the first individual while 23.2 Gb of Illumina MiSeq data were generated from the second individual. The inclusion of Nanopore reads for scaffolding followed by subsequent gap-closing using Illumina data led to a 29% reduction in the number of scaffolds and a 55% and 54% increase in the scaffold and contig N_{50}, respectively. We also report the first transcriptome of Murray cod that was subsequently used to annotate the Murray cod genome, leading to the identification of 26 539 protein-coding genes. We present the whole genome of the Murray cod and anticipate this will be a catalyst for a

Received: 4 May 2017; **Revised:** 12 June 2017; **Accepted:** 11 July 2017

range of genetic, genomic, and phylogenetic studies of the Murray cod and more generally other fish species of the *Percichthydae* family.

Keywords: Murray cod; long reads; genome; transcriptome; hybrid assembly

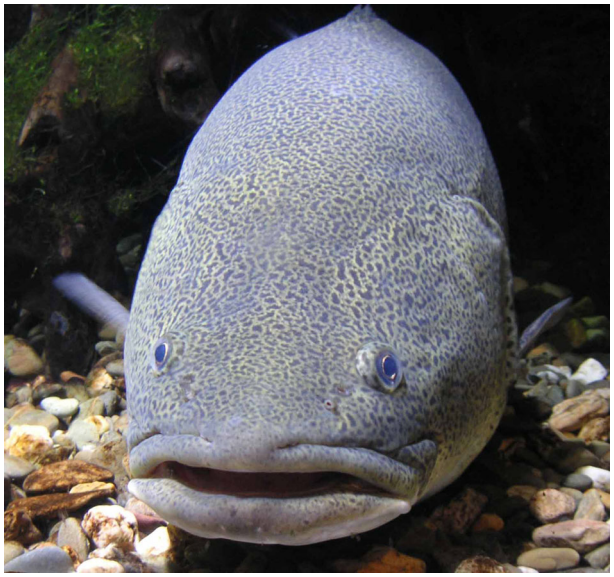

Figure 1: The iconic Murray cod. Photo: Paul Sunnucks.

Data Description

Population genetic and evolutionary studies on Australian freshwater fish are of special interest in relation to conservation, biogeography, and adaptive responses and have been studied using a range of molecular techniques [1–8]. A limitation to a more complete understanding of the genetics and evolution of Australian inland fish species is the lack of genome-level resources [9]. The Murray cod, *Maccullochella peelii* (NCBI Taxon ID: 135761, Fishbase ID: 10311), is one of Australia's most iconic large (up to ~1.8 metres) and long-lived (≥48 years) predatory fish species that occurs across highly variable and heterogeneous riverine environments of inland Australia (Fig. 1). Despite being widespread, the Murray cod is a threatened species under national legislation (Environment Protection and Biodiversity Conservation Act 1999), and populations are intensively managed through programs such as habitat restoration, provision of environmental flows, and stocking.

Sampling, Library Construction, and Sequencing

Sequencing data from 2 Murray cod individuals were merged for whole genome assembly. The first individual was obtained from an Australian fish market in 2014 [5]. Genomic DNA was extracted from multiple fin clip and muscle samples using DNAeasy Blood and Tissue Kits (Qiagen, Halden, Germany). A 300-bp insert library was prepared from the purified gDNA using the TruSeq DNA sample prep kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions and subsequently sequenced (2 × 100 bp, 1 × 100 bp configurations) on a HiSeq 2000 (Illumina, San Diego, CA, USA), located at the Malaysian Genomics Resource Centre Berhad. For sequencing on the MinION, gDNA was extracted from the remaining fin clip and muscle tissues that were collected in 2014. However, due to DNA degradation associated with long-term storage, an additional size selection (8–30 kb) with a BluePippin was performed to reduce the representation of short reads (Sage Science, Beverly, MA, USA). Seven individual libraries (2 1D preps and 5 2D preps) were prepared and sequenced on 7 R9 flowcells using the MinION portable DNA sequencer (Oxford Nanopore, Oxford, UK) according to the manufacturer's instructions. The second Murray cod, isolate KMC200 (=MCC0324) [2], was sampled from the Lachlan River in New South Wales in 2006, and its library has been previously constructed and sequenced at the Monash University Malaysia Genomics Facility for a mitogenome-based population genetics study [2, 5]. Given that the whole mitogenome of isolate KMC200 (= MCC0324; GenBank accession number: KT337332.1) exhibits a 100% nucleotide identity to that of the first individual (GenBank accession number: NC_023807.1), indicating a recently shared maternal ancestry [2, 5], its remaining library was re-sequenced on 3 separate MiSeq runs (2 × 250 bp configuration) to improve the sequencing coverage of the Murray cod genome. A total of 70.6 Gb (47.4 Gb and 23.2 Gb from HiSeq and MiSeq runs, respectively) and 804 Mb (N_{50}: 4438 bp, longest read: 129 945 bp) of nucleotide sequence were generated on the Illumina platforms and the Oxford Nanopore MinION device, respectively.

Genome Characteristics

Jellyfish v. 2.2.6 [10] was used to obtain a frequency distribution of 17-, 21-, 25-, and 31-mers in a subset (~20 Gb) of the raw HiSeq sequence reads, and the histograms were uploaded to GenomeScope for estimation of genome size, repeat content, and heterozygosity, based on a kmer-based statistical approach [11]. The resulting analysis shows that the haploid genome size was between 640 and 669 Mbp for the Murray cod (Fig. 2), a figure smaller than the 812 Mbp (C-value: 0.83 pg) estimated size reported on the Animal Genome Size Database [12, 13]. This smaller estimate may be due to an additional parameter introduced in GenomeScope, set to exclude extremely high-frequency kmers as these likely represent organelle sequences or other contaminants that can inflate the genome size [11]. Further, the 21-mer analysis (with "max kmer coverage" set at 1000) on GenomeScope also indicates 14.3% repeat content and a low heterozygosity of 0.103%. To test that both Murray cod isolates possess the same genome characteristics, the Jellyfish and GenomeScope analysis was repeated for 23 Gb of MiSeq sequence reads, which resulted in comparable results (643 to 673 Mbp haploid genome size, 15.7% repeat content, a low heterozygosity of 0.113%) (Supplementary Fig. S1; for combined data set, see Supplementary Fig. S2). Further repeat-content analysis and masking is performed in subsequent sections in this study (see "Repeat-content analysis").

Genome Assembly

Illumina reads were trimmed with *platanus_trim* v. 1.0.7 (*-q* 20, *-l* 35) and assembled with the Platanus v. 1.2.4 assembler to account for a potential increase in genome heterozygosity due to the use of sequencing data from 2 individuals with shared

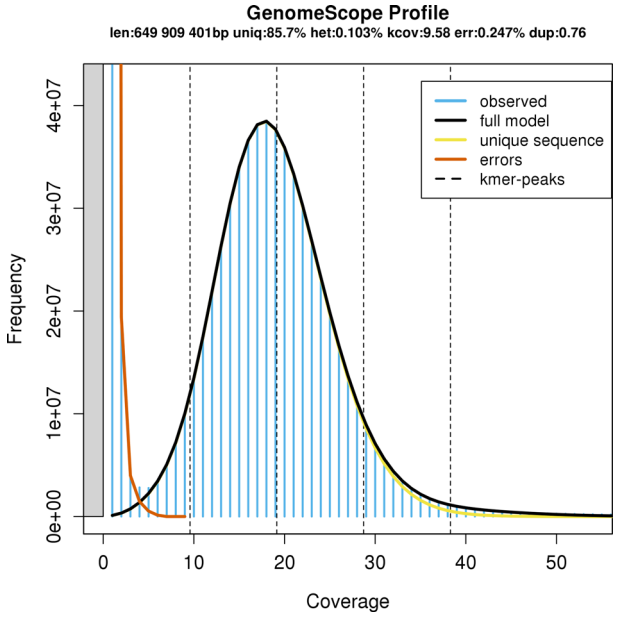

Figure 2: Estimation of genome size, repeat content, and heterozygosity by GenomeScope, based on 21-mers in HiSeq sequence reads (max kmer coverage at 1000).

maternal ancestry [14]. The initial assembly is 622 Mb in length, comprising 80 098 scaffolds with an N_{50} of 68 937 bp (Table 1). The assembly was subsequently scaffolded with SSPACE-LongRead v. 1–1 (BLASR aligner, default settings, minimum 3 links [long reads] required for scaffolding) [15] using long-read MinION sequences (0.93 × coverage), which was base-called offline with Albacore/ONT Sequencing Pipeline Software v. 0.7.4 followed by further gap-filling with Illumina reads using GapFiller v. 1–10 [16].

By adding only 804 Mb of Nanopore reads, we observed improvements in the final 633 Mb assembly, reducing the number of scaffolds (≥500 bp) by 29%, from 25 642 to 18 198, and increasing the scaffold N_{50} by 55%, from 70 993 bp to 109 975 bp. In addition, based on results from read alignment performed with Bowtie2 v. 2.3.2 [17], only a small percentage of the scaffolds representing less than <0.005% of the total assembly size were unique to 1 donor (Supplementary Data 1).

Genome completeness was estimated using 2 separate programs, CEGMA (CEGMA, RRID:SCR_015055) and BUSCO v. 3.0 (BUSCO, RRID:SCR_015008). For BUSCO analysis (-m geno –sp zebrafish settings), the genome was searched against the actinopterygii database (actinopterygii_odb9), which was constructed from 20 fish species consisting of 4584 orthologs. Final genome completeness percentages of 94.2% and 91.94% were estimated by BUSCO and CEGMA, respectively. Further, both analyses also indicate a slight improvement in the genome completeness with the inclusion of Nanopore reads for scaffolding and subsequent gap-closing using Illumina short reads. The small amount of available Nanopore long reads in this study resulted in a limitation in the program of choice in assembly as well as scaffolding. At the time of this study, we chose to use SSPACE-LongRead [15] as a scaffolder as it has been used in several genome assembly publications utilizing Nanopore reads, albeit mostly bacterial genome assemblies [18, 19], reviews [20–22], and benchmarking studies [23–26], as well as some eukaryotic genome assemblies that utilized BAC or fosmid libraries or PacBio long read data [27–29]. While no formal testing on eukaryotic genomes and Nanopore long reads was done by Boetzer and Pirovano [15] in their publication, there is mention of the potential of the method applied on Nanopore reads and eukaryotic assemblies. We have found SSPACE-LongRead to be effective in the scaffolding of the Murray cod contigs as elaborated earlier and also in Table 1. Though the gene content appears to support the validity of the assembly, this study does not include further assessment or verification of the accuracy of the scaffold extensions by SSPACE-LongRead [15]. It is noteworthy, however, that a greater range of assembly and scaffolder programs has become available for large eukaryotic genomes; these programs are worth exploring in future studies [24, 26, 30–32].

Repeat-Content Analysis

To identify repeats in the assembly, a *de novo* repeat library was first built with RepeatModeler v. 1.0.4 (RepeatModeler, RRID:SCR_015027) [33] using default parameters based on the larger scaffolds (≥5 kb) in the assembly. RepeatMasker v. open-4.0.7 (RepeatMasker, RRID:SCR_012954) [34] was then used to align sequences from the whole assembly to the RepeatMasker Combined Library (Dfam_Consensus 20 170 127 [35] and RepBase 20 170 127 [36]) as well as the *de novo* repeat library to screen for repeats and low-complexity sequences in the assembly. Repeat sequences were estimated to account for 23.38% (148 Mb) of the Murray cod assembly presented in this study.

Transcriptome Assembly

Total RNA was extracted using the RiboPure RNA purification Kit (Thermo Fisher Scientific, Waltham, MA, USA) from the liver, brain, and muscle tissues of a juvenile Murray cod that was collected from a natural population in Broken Creek under a DELWP collecting permit and euthanized using approved procedures under a Monash ethics permit (BSCI/2012/19). Thirty μL of 300 ng/μL of each RNA extract was pooled and processed as a single sample using the TruSeq RNA library kit (Illumina, San Diego, CA, USA) to generate a 160-bp insert size library. The library was subsequently sequenced on 1 lane of HiSeq2000 (2 × 100 bp configuration) at the Ramaciotti Centre for Gene Function Analysis. A total of 376 million reads were generated and preprocessed with Trimmomatic v. 0.32 (*leading: 3, trailing: 3, slidingwindow: 4:20, minlen: 75*; Trimmomatic, RRID:SCR_011848) [37]. These reads were then assembled *de novo* using Trinity v. r20140717 (Trinity, RRID:SCR_013048) [38], producing a 305-Mb transcriptome consisting of 321 855 transcripts.

Genome Annotation

The MAKER2 genome annotation pipeline [39] predicted protein-coding genes using 3 approaches: (i) homology to fish proteins, (ii) assembled transcripts as RNA-seq evidence, and (iii) *de novo* gene predictors. Protein sequences from 11 other fish species on Ensemble and the set of Murray cod transcripts assembled in this study were aligned to the genome in a preliminary MAKER run as evidence to retrain *ab initio* gene predictors such as Augustus (Augustus: Gene Prediction, RRID:SCR_008417) [40] and SNAP [41]. These higher-quality gene models are then used in subsequent runs to predict the final set of Murray cod protein-coding genes. The pipeline identified 26 539 genes with an average annotation edit distance (AED) of 0.187 [42].

NCBI's *blastp* (*-evalue* $1e^{-10}$, *-seg yes*, *-soft_masking true*, *-lcase_masking*, and hit fraction of ≥70% target length; BLASTP,

Table 1: Murray cod assembly and annotation statistics

Genome assembly	Illumina only	Illumina (≥500 bp)	Illumina + Nanopore (≥500 bp)
Number of contigs	95 612	41 152	45 882
Contig N_{50} size	33 442 bp	34 269 bp	52 687 bp
Longest contig	328 477 bp	328 477 bp	501 239 bp
Number of scaffolds	80 098	25 642	18 198
Total scaffold size	622 421 194 bp	609 090 121 bp	633 241 041 bp
Scaffold N_{50} size	68 937 bp	70 993 bp	109 974 bp
Longest scaffold	548 726 bp	548 726 bp	1 119 190 bp
% GC/AT/N	40.7/59.1/0.2	40.7/59.2/0.1	40.4/58.7/0.9
CEGMA completeness	89.52%	84.68%	91.94%
Complete BUSCOs	4228 (92.3%)	4229 (92.3%)	4317 (94.2%)
Complete and single-copy BUSCOs	4115 (89.8%)	4115 (89.8%)	4202 (91.7%)
Complete and duplicated BUSCOs	113 (2.5%)	114 (2.5%)	115 (2.5%)
Fragmented BUSCOs	224 (4.9%)	222 (4.8%)	156 (3.4%)
Missing BUSCOs	132 (2.8%)	133 (2.9%)	111 (2.4%)
Transcriptome assembly			
Number of transcripts	321 855		
Transcriptome size	305 149 376 bp		
Mean transcript length	948.10 bp		
Longest transcript	23 655 bp		
CEGMA completeness	99.19%		
Annotation			
Number of protein-coding genes	26 539		
Mean gene length	10 115.3 bp		
Longest gene	134 909 bp		
With functional annotation	25 607		

RRID:SCR_001010) [43] was used to functionally annotate the gene sequences against vertebrate sequences in the NCBI non-redundant database, after which un-annotated sequences were searched against all sequences in the NCBI non-redundant database. Additional functional annotation was performed with InterProScan (InterProScan, RRID:SCR_005829) [44] to examine motifs, domains, and signatures in the Murray cod protein sequences based on information from public databases, including PANTHER (PANTHER, RRID:SCR_004869) [45], Pfam (Pfam, RRID:SCR_004726) [46], PRINTS (PRINTS, RRID:SCR_003412) [47], PROSITE (PROSITE, RRID:SCR_003457) [48], SMART (SMART, RRID:SCR_005026) [49], SUPERFAMILY (SUPERFAMILY, RRID:SCR_007952) [50], and TIGRFAMs (JCVI TIGRFAMS, RRID:SCR_005493) [51]. As a result, 96.5% of the predicted protein-coding genes were successfully annotated by at least 1 of the 2 methods (*blastp* 69%, InterProScan 96.1%).

Conclusion

Having assembled and annotated the genome of an Australian teleost fish, we anticipate that this will be a catalyst for a range of genetic, genomic, and evolution-related studies of the Murray cod and related fish species (Harrisson et al., submitted for publication). In this study, we demonstrate that, despite its reported high error rate, low-coverage Nanopore long reads are still useful for scaffolding fish genome assembly. However, low-coverage long reads still pose limitations in (i) the full utilization of these reads, e.g., sequence self-correction and the use of long reads itself for assembly and gap-filling, and (ii) the choice of the most suitable assembly and scaffolder programs. Given the relative ease of generating Nanopore MinION reads and continuous improvement in data yield and read accuracy of this sequencing platform, we look forward to overcoming these limitations and to further incorporating Nanopore long read information into eurkaryote genome assemblies, either in hybrid approaches or, ideally and ultimately, in non-hybrid *de novo* assemblies. We anticipate that Nanopore long reads will increasingly complement or even supersede short read data for the *de novo* genome assembly of fish species.

Availability of supporting data

The data sets supporting the results of this article are available in the *GigaDB* repository [52]. Raw reads (Illumina and Nanopore) are available in the Sequence Read Archive (SRA), and the Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession number LKNJ00000000 (first version), both under BioProject PRJNA290988. Similarly, transcriptome (Illumina) reads are also available in the SRA, and the Transcriptome Shotgun Assembly project has been deposited under accession number GFMM00000000 (first version) as part of BioProject PRJNA383091.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was funded by the Monash University Malaysia Tropical and Biology Multidisciplinary Platform and ARC grant LP110200017 to Monash University, Flinders University, and the University of Canberra, with the following Partner Organizations: University of Montana, ACTEW Corporation, Department

of Sustainability and Environment (Victoria; now the Department of Environment, Land, Water and Planning [DELWP]), Fisheries Victoria (now within Department of Economic Development, Jobs, Transport and Resources), and Melbourne Water. We thank Joanne Kearns and Jarod Lyon from Arthur Rylah Institute (DELWP) and Dean Gilligan and Meaghan Rourke from the New South Wales Department of Primary Industries (NSW DPI) for assistance in sample collection, Catriona Millen for assistance with RNA extraction, and Steven Amish for assistance with preliminary transcriptome assembly. We also acknowledge the Monash University Malaysia High Performance Computing infrastructure for computational resources.

References

1. Pavlova A, Gan HM, Lee YP et al. Purifying selection and genetic drift shaped Pleistocene evolution of the mitochondrial genome in an endangered Australian freshwater fish. Heredity 2017;**118**:466–76.
2. Harrisson K, Pavlova A, Gan HM et al. Pleistocene divergence across a mountain range and the influence of selection on mitogenome evolution in threatened Australian freshwater cod species. Heredity 2016;**116**(6):506–15.
3. Cole TL, Hammer MP, Unmack PJ et al. Range-wide fragmentation in a threatened fish associated with post-European settlement modification in the Murray–Darling Basin, Australia. Conserv Genet 2016;**17**(6):1377–91.
4. Unmack PJ, Sandoval-Castillo J, Hammer MP et al. Genome-wide SNPs resolve a key conflict between sequence and allozyme data to confirm another threatened candidate species of river blackfishes (Teleostei: Percichthyidae: Gadopsis). Mol Phylogenet Evol 2017;**109**:415–20.
5. Austin CM, Tan MH, Lee YP et al. The complete mitogenome of the Murray cod, *Maccullochella peelii* (Mitchell, 1838) (Teleostei: Percichthyidae). Mitochondrial DNA 2016;**27**(1):729–30.
6. Harrisson KA, Yen JDL, Pavlova A et al. Identifying environmental correlates of intraspecific genetic variation. Heredity 2016;**117**(3):155–64.
7. Pavlova A, Beheregaray LB, Coleman R et al. Severe consequences of habitat fragmentation on genetic diversity of an endangered Australian freshwater fish: a call for assisted gene flow. Evol Appl 2017;**10**(6):531–50.
8. Hermoso V, Kennard MJ, Schmidt DJ et al. Species distributions represent intraspecific genetic diversity of freshwater fish in conservation assessments. Freshw Biol 2016;**61**(10):1707–19.
9. Robledo D, Palaiokostas C, Bargelloni L et al. Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev Aquacult 2017; doi:10.1111/raq.12193.
10. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;**27**(6):764–70.
11. Vurture GW. GenomeScope: fast reference-free genome profiling from short reads. bioRxiv 2016. https://doi.org/10.1101/075978.
12. Gregory TR. Animal Genome Size Database. 2017. http://www.genomesize.com (22 August 2017, date last accessed).
13. Hardie DC, Hebert PD. Genome-size evolution in fishes. Can J Fish Aquat Sci 2004;**61**(9):1636–46.
14. Kajitani R, Toshimoto K, Noguchi H et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 2014;**24**(8):1384–95.
15. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 2014;**15**(1):211.
16. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol 2012;**13**(6):R56.
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**(4):357–9.
18. Risse J, Thomson M, Patrick S et al. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience 2015;**4**(1):60.
19. Karlsson E, Lärkeryd A, Sjödin A et al. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep 2015;**5** 1:11996.
20. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 2016;**14**(5):265–79.
21. Laver T, Harrison J, O'Neill PA et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quant 2015;**3**:1–8.
22. Yuan Y, Bayer PE, Batley J et al. Improvements in genomic technologies: application to crop genomics. Trends Biotechnol 2017;**35**(6):547–58.
23. Cao MD, Nguyen SH, Ganesamoorthy D et al. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. Nat Commun 2017;**8**:14515.
24. Warren RL, Yang C, Vandervalk BP et al. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience 2015;**4**(1):35.
25. Madoui M, Engelen S, Cruaud C et al. Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics 2015;**16**(1):327.
26. Gao S, Bertrand D, Chia BKH et al. OPERA-LG: efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome Biol 2016;**17**(1):102.
27. Cruz F, Julca I, Gómez-Garrido J et al. Genome sequence of the olive tree, *Olea europaea*. Gigascience 2016;**5**(1):29.
28. Luo Y, Takeuchi T, Koyanagi R et al. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nat Commun 2015;**6**:8301.
29. Plomion C, Aury J, Amselem J et al. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Resour 2016;**16**(1):254–65.
30. Jansen HJ. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. bioRxiv 2017. https://doi.org/10.1101/101907.
31. Koren S, Walenz BP, Berlin K et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;**27**(5):722–36.
32. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016;**32**(14):2103–10.
33. Smit A, Hubley R. RepeatModeler Open-1.0. 2008–2015. http://www.repeatmasker.org (13 March 2017, date last accessed).
34. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org (6 January 2017, date last accessed).
35. Hubley R, Finn RD, Clements J et al. The Dfam database of repetitive DNA families. Nucleic Acids Res 2016;**44**(D1):D81–9.

36. Jurka J, Kapitonov VV, Pavlicek A et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;**110**(1–4):462–7.
37. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;**30**(15):2114–20.
38. Grabherr MG, Haas BJ, Yassour M et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;**29**(7):644–52.
39. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011;**12**(1):491.
40. Stanke M, Schöffmann O, Morgenstern B et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 2006;**7**(1):62.
41. Korf I. SNAP: Semi-HMM-based Nucleic Acid Parser. Ian Korf homepage. 2013. http://homepage.mac.com/iankorf (6 January 2017, date last accessed).
42. Eilbeck K, Moore B, Holt C et al. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 2009;**10**(1):67.
43. Altschul SF, Gish W, Miller W et al. Basic local alignment search tool. J Mol Biol 1990;**215**(3):403–10.
44. Jones P, Binns D, Chang H-Y et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;**30**(9):1236–40.
45. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2013;**41**(D1):D377–86.
46. Punta M, Coggill PC, Eberhardt RY et al. The Pfam protein families database. Nucleic Acids Res 2012;**40**(D1): D290–301.
47. Attwood TK, Coletta A, Muirhead G et al. The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. Database (Oxford) 2012; doi:10.1093/database/bas019.
48. Sigrist CJA, De Castro E, Cerutti L et al. New and continuing developments at PROSITE. Nucleic Acids Res 2013;**41**(D1):D344–7.
49. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 2012;**40**(D1):D302–5.
50. de Lima Morais DA. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 2011;**39**(Database issue):D427–34.
51. Haft DH, Selengut JD, Richter RA et al. TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res 2013;**41**(D1):D387–95.
52. Austin CM, Tan MH, Harrisson KA. De novo genome assembly and annotation data for the Murray cod (*Maccullochella peelii*), Australia's largest freshwater fish. GigaScience Database 2017; http://dx.doi.org/10.5524/100329.