Genome assembly and isoform analysis of a highly heterozygous New Zealand fisheries species, the tarakihi (Nemadactylus macropterus)

Yvan Papa, Maren Wellenreuther, Mark A. Morrison, Peter A. Ritchie

1 School of Biological Sciences, Victoria University of Wellington, Wellington 6012, New Zealand
2 Seafood Production Group, The New Zealand Institute for Plant and Food Research Limited, Nelson 7010, New Zealand
3 School of Biological Sciences, The University of Auckland, Auckland 1010, New Zealand
4 National Institute of Water and Atmospheric Research, Auckland 1010, New Zealand

*Corresponding author: Te Toki A Rata, Gate 7, Kelburn Parade, Wellington 6012, New Zealand. Email: peter.ritchie@vuw.ac.nz

Abstract

Although being some of the most valuable and heavily exploited wild organisms, few fisheries species have been studied at the whole-genome level. This is especially the case in New Zealand, where genomics resources are urgently needed to assist fisheries management. Here, we generated 55 Gb of short Illumina reads (92x coverage) and 73 Gb of long Nanopore reads (122x) to produce the first genome assembly of the marine teleost tarakihi [Nemadactylus macropterus (Forster, 1801)], a highly valuable fisheries species in New Zealand. An additional 300 Mb of Iso-Seq reads were obtained to assist in gene annotation. The final genome assembly was 568 Mb long with an N50 of 3.37 Mb. The genome completeness was high, with 97.8% of complete Actinopterygii Benchmarking Universal Single-Copy Orthologs. Heterozygosity values estimated through k-mer counting (1.00%) and bi-allelic SNPs (0.64%) were high compared with the same values reported for other fishes. Iso-Seq analysis recovered 91,313 unique transcripts from 15,515 genes (mean ratio of 5.89 transcripts per gene), and the most common alternative splicing event was intron retention. This highly contiguous genome assembly and the isoform-resolved transcriptome will provide a useful resource to assist the study of population genomics and comparative eco-evolutionary studies in teleosts and related organisms.

Keywords: fish, genomics, Iso-Seq, marine, teleost, transcriptome

Introduction

The tarakihi or jackass morwong (Nemadactylus macropterus, Centrarchiformes: Cirrhitidae, NCBI Taxon ID: 76931) is a species of demersal marine teleost fish that is widely distributed around all inshore areas of New Zealand and along the southern coasts of Australia. It is distinguishable from other New Zealand “morwongs” by the black saddle across its nape (Roberts et al., 2015) and displays a single elongated pectoral fin ray that is characteristic of Nemadactylus species (Ludt et al., 2019). The species and its genus have been recently moved from the Cheilodactyliidae to the Latridae following extensive revision of the taxonomy of both families, which until then was poorly understood (Kimura et al., 2018; Ludt et al., 2019). Tarakihi is an important commercial and recreational inshore fishery, especially in New Zealand, where more than 5,000 tonnes are harvested every year (Fisheries New Zealand, 2018). Like many other fisheries species, tarakihi stocks have been heavily fished over the past century. As a result, the spawning biomass is now concerning depleted to numbers below the fisheries management soft limit of 20% on the east coast of New Zealand, where fishing effort is highest (Langley, 2018). Low effective population size and spawning biomass are of concern for the long-term sustainability of this species, particularly with added and increasing environmental pressures due to global warming. Climate change is already having an impact on marine ecosystems and is expected to affect the distribution and productivity of many fisheries species (Burrows et al., 2011; Ramos et al., 2018; Babcock et al., 2019).

The application of genome-wide markers for tarakihi fisheries management has been limited by the lack of a reference genome. Consequently, the first step in developing new genomic resources for this species is to assemble a high-quality reference genome that can be used to develop high-resolution markers for determining the genetic stock structure. This will offer the potential to estimate gene flow levels and detect adaptive genetic variation (Papa, Morrison, et al., 2022). Incorporating adaptive genetic variation, along with neutral variation, will greatly improve how the genetic data can be used for fisheries management (Bernatchez et al., 2017; Benestan, 2019; Papa, Oosting, et al., 2021; Thomson et al., 2021).

Genome assembly quality, contiguity, and completeness depend on the available technology used. While short-read Illumina sequencing produces highly accurate reads, their short length (<200 bp) makes them problematic for the assembly of highly repetitive segments of the genome. Complex genomes often result in highly fragmented assemblies (Koren et al., 2012; Rice & Green, 2019). Combining short reads with less accurate long-read sequencing technologies leads to more contiguous
genome assemblies (Austin et al., 2017; Zimin, Puju, et al., 2017; Zimin, Stevens, et al., 2017; Tan et al., 2018; Dhar et al., 2019; Jiang et al., 2019; Wiley & Miller, 2020). On the other hand, the relatively young circular consensus sequencing (CCS) PacBio technology produces reads that are both thousands of bp long and highly accurate. CCS long-read DNA sequencing can be applied to DNA (HiFi reads) and RNA (i.e. isof orm sequencing or Iso-Seq). Iso-Seq allows for the sequencing of complete, uninterrupted mRNAs, which enables the accurate characterization of isoforms (An et al., 2018; Byrne et al., 2019; Gao et al., 2019; Hoang & Henry, 2021). Iso-seq has been used to annotate de novo genome assemblies of nonmodel organisms like the cave nectar bat (Eonycteris spelaea) (Wen et al., 2018), the pharaoh ant (Monomorium pharaonis) (Gao et al., 2020), the red-eared slider turtle (Trachemys scripta elegans) (Simison et al., 2020), and the sponge gourd (Luffa spp.) (Pootakham et al., 2020), allowing for the characterization of both gene functions and alternative splicing (AS) patterns in these species.

The main goal of this study was to complete the first tarakihi genome assembly. This was achieved by using a combination of short-read Illumina and long-read Nanopore sequencing data. Four assembly pipelines were compared, three of which used algorithms implemented in MaSuRCA for hybrid assembly, and a fourth pipeline was based on a trial run of low-coverage DNA sequence reads (4 Gb) generated using the PacBio HiFi platform. Iso-Seq data were used to assist with gene annotation and the identification of gene isoforms.

Materials and methods

Tissue collection and nucleotide extraction

Tissues for Illumina and Nanopore sequencing were collected from a freshly vouchered N. macropterus specimen (male, standard length: 285 mm, weight: 460 g). The specimen was a captive-bred from Plant & Food Research, Nelson, New Zealand (Fig. 1a) and is thereby referred to as TARdn1 (for “tarakihi de novo”). A caudal fin clip and a heart piece were stored in 96% EtOH, and a kidney piece was stored in DESS (20% DMSO, 0.25 M EDTA, NaCl saturated solution). Total genomic DNA was extracted from these tissues using a high-salt extraction protocol adapted from that of Aljanabi & Martinez (1997), which included an RNase treatment, and then suspended in Tris-EDTA buffer (10 mM Tris–HCl pH 8.0 and 1 mM EDTA). The integrity of DNA fragments was assessed by gel electrophoresis in 1% agarose. The purity and quantity of DNA (concentration >200 ng/µl, A260/280=1.8, A60/230=0.1) were estimated with CLARIOstar spectrometer (BMG Labtech). Purified DNA samples were sent to Annoroad Gene Technology Co. Ltd (Beijing, China) and NextOmics Biosciences Co. Ltd (Wuhan, China) for Illumina and Nanopore library preparation and sequencing.

Tissues for HiFi sequencing and Iso-Seq were obtained from a wild specimen (male, standard length: 255 mm) captured by a recreational fisherman at Kau Bay, in Wellington Harbour (New Zealand), thereby referred to as TARdn2 (Fig. 1b). Tissues were collected c. 6 h after capture and flash-frozen in liquid nitrogen. Five pieces of tissues were sent to BGI Tech Solutions Co. Ltd (Hong Kong, China): one tissue (heart, weight: 2 g) for DNA extraction and HiFi sequencing and four tissues (liver, white muscle, brain, and spleen, weight: c. 150 mg) for RNA extraction and Iso-Seq. DNA and RNA were extracted by BGI using a phenol-chloroform method.

Library preparations and sequencing

Before sequencing, the genome size of N. macropterus was estimated to be about 700 Mb (see Methods in Supplementary File 1). The quantity of Illumina and Nanopore bases to be sequenced was tuned for a deep 85× Illumina coverage (c. 60 Gb) and 140× Nanopore coverage (c. 100 Gb), following sequencing provider recommendations. For Illumina reads, DNA samples were sheared for a fragment insert size of 350 ± 50 bp. Approximately 200 million of 150 bases pair-end reads were generated using the HiSeq X System (Illumina). Raw Illumina reads were filtered for adapter contamination, base uncertainty, and Quality Value. For Nanopore sequencing, a DNA library of 20–40 Kb fragments was loaded into two flow cells on PromethION (Oxford Nanopore Technologies). The HiFi library was prepped with SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) and CCS was performed on one-third of an SMRT Cell 8 M with a PacBio Sequel II sequencer. Four Iso-Seq libraries of 0–5 Kb insert sizes (one per tissue) were generated using the SMRTbell Express Template Prep Kit 2.0. The multiplexed libraries were sequenced on one SMRT Cell 8 M with a PacBio Sequel II sequencer, resulting in 3.6 million polymerase reads from which subreads were extracted (see Methods in Supplementary File 1 for more details on library preparation, sequencing, and preliminary filtering).

Quality, contamination, and mitochondrial filtering

Primary quality filtering resulted in 405.2 million Illumina pair-end reads (60.78 Gb). Quality metrics of these filtered reads were visualized with FastQC v0.11.7 (Andrews, 2018) before proceeding to the next steps. Kraken v2.0.7-beta (Wood et al., 2019) was used to detect and filter contamination from archaeal, bacterial, viral, and human sequences based on the MiniKraken2 v2 8GB database (Wood, 2019). The 9.25% of reads that were classified as contaminants were discarded, leading to 367.8 million noncontaminated
reads (55.16 Gb) (Table 1). The mitogenome sequence (16,650 bp) was then retrieved and discarded from the genome assembly (see details in Methods in Supplementary File 1). A total of 99.18 Gb was obtained from the raw unfiltered Nanopore reads, with an average read length above 6 Kb and a maximum length above 1 Mb (Table 1). Nanopore reads were filtered and trimmed for minimum length (500 bp) and a minimum average read quality score of 7 (c. 80% base call accuracy) (see details in Methods in Supplementary File 1).

### Genome size, coverage, and heterozygosity estimation postsequencing

Genome size and sequencing coverage based on the Illumina sequence reads were performed with a k-mer frequency analysis of 17, 21, and 27-mers using jellyfish v2.2.10 (Marçais & Kingsford, 2011) and GenomeScope (Vurture et al., 2017) (see details in Supplementary File 1: Methods). A haploid genome size of c. 516–520 Mb, with a high heterozygosity level of 1.01–1.07% and a duplication level of 0.98–1.10%, was estimated (Fig. 2, Supplementary Fig. 1 in Supplementary File 2). This estimated genome size was consistent, albeit c. 150 Mb lower than the size estimated presequencing. However, it is common for k-mer estimated size and genome assembly size to be smaller than the size estimated with C-value (Austin et al., 2017; Jansen et al., 2017; Feron et al., 2020). The heterozygous coverage of 40× was considered sufficient for performing genome assembly. The heterozygosity of TARdn1 was estimated a second time by calling SNPs from the Illumina reads aligned to the final assembly (see details of parameters in Supplementary File 1: Methods and Supplementary Fig. 2 in Supplementary File 2).

### Illumina + nanopore hybrid assembly

De novo genome assembly of short- and long-reads was performed with the Maryland Super-Read Celera Assembler pipeline, MaSuRCA (Zimin et al., 2013; Zimin, Puiu, et al., 2017). This is one of the most common assemblers for performing short- and long-reads hybrid genome assemblies of eukaryotes, with consistently good results across studies (Tan et al., 2018; Jiang et al., 2019; Thai et al., 2019) (see details on the MaSuRCA pipeline in Supplementary File 1: Methods). The hybrid Illumina + Nanopore assembly was run on MaSuRCA v3.2.9 (see parameters in Supplementary File 1). MaSuRCA v3.2.9 uses a modified version of the CABOG assembler (Miller et al., 2008) for the final assembly of corrected mega-reads. However, later releases of MaSuRCA included the Flye assembler (Kolmogorov et al., 2019) as a supposedly faster and more accurate alternative tool for the same step. To compare both methods, a second assembly was run on MaSuRCA v3.4.1 with the same parameters as above, but this time using FLYE_ASSEMBLY = 1. The Flye assembly was subsequently polished with POLCA (Zimin & Salzberg, 2020) as implemented in MaSuRCA v3.4.1 on default settings, using the clean Illumina reads to fix substitutions and indel errors.

### HiFi sequencing and assembly

Assembly of HiFi reads was performed with hiiasm v0.13 (Cheng et al., 2021) using default parameters. Another assembly was also tentatively performed with HiCanu as implemented in Canu v2.1.1 (Nurk et al., 2020), with an estimated genome size of 600 Mb. However, the read coverage estimated (6.68x) was lower than the minimum coverage allowed by HiCanu (10x), so the assembly could not be completed.
Quality assessment and comparison of assemblies

After each assembly, basic contiguity statistics were computed with bmap v38.31 (Bushnell, 2018) script stats.sh. To assess the completeness of the assemblies, the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool v3.0.2 (Simão et al., 2015) was used with parameter -p sp to zebrafish on the Actinopterygii odb9 orthologs set, which contains 4,584 single-copy orthologs that are present in at least 90% of ray-finned fish species.

The quality of the CABOG and Flye assemblies was further compared by mapping clean Illumina reads back to the assemblies themselves with bwa-kit v0.7.15 using bwa mem -a -M. The resulting alignment files were also used to plot Feature Response Curves (FRC) (Vezzi et al., 2012b) with FRCbam v5b3f53e-0 (Vezzi et al., 2012a). This allowed comparison of the quality of the assemblies without relying on contiguity, by plotting the accumulation of error “features” along the genome (e.g. areas with low or high coverage, numbers of unpaired reads, and misoriented reads). The presence of unmerged haplotigs in the CABOG and the Flye polished assembly was investigated by using mimap v2.16 (Li, 2018) with parameters -ax map-ont -secondary = no to map the clean Nanopore reads back to the assembly and then analyzing the resulting alignment with Purge Haplotigs v1.1.1 (Roach et al., 2018) command hists.

The last quality check of the CABOG and Flye polished assemblies was done by plotting assemblies against each other and against two chromosome-level fish assemblies using MashMap 2.0 (Jain et al., 2018) with a minimum mapping segment length of 500 bp and a minimum identity of 85% (for comparison between tarakahi assemblies) and 90% (for comparison between different species). To visualize the presence of potential misassemblies on the longest scaffolds, the results from MashMap were used to plot the mappings of these scaffolds between different assemblies with a custom R script (plot_mashmap_scaffolds.R). The first fish chromosome-level assembly used for comparison was the mandarin fish Siniperca chuatsi (SinChu7, GCA_001952085.1) because it was the phylogenetically closest chromosome-level assembly (Centrarchiformes, Centrarchoidei) available on NCBI at the time this analysis was performed. The second was the Austrasian snapper Chrysophrys auratus (SNA1, https://www.genomics-aotearoa.org.nz/data), to compare with a well-curated specimen from a more evolutionarily distant species.

Final visualization of contiguity and completeness of the genome assemblies was generated with assembly-stats v17.02 (Challis, 2017) as implemented in the grpiccoli container (Piccoli, 2021).

Genome repetitive elements detection

Repetitive elements (REs) in the N. macropterus genome were identified both by de novo modeling and based on repeats homology using RepeatModeler v2.0.1 (Flynn et al., 2020) and RepeatMasker v4.1.1 (Smit et al., 2013) as implemented in Dfam TE Tools container v1.2 (https://github.com/Dfam-consortium/TETOols); see Supplementary File 1: Methods for more details).

Iso-Seq analysis

Iso-Seq subreads were processed with the SMRTLink v9.0 Iso-Seq pipeline. Circular consensus sequences were generated and converted into high-quality (predicted accuracy > 0.99), polished, nonredundant isoforms (see details in Supplementary File 1: Methods). These isoforms were screened for REs against the N. macropterus custom repeat library with RepeatMasker v4.1.1. Transcripts with ≥70% bases masked were considered REs. Identified REs were discarded from further analyses using a custom bash script for filtering (Count_filter_N_isoseqrepeats.bash) and categorized using a custom R script (R_characterize_transcripts.R).

AS events in the repeat-cleaned Iso-Seq reads were counted and classified with SUPPA v2.3 (Trindaco et al., 2018) with default parameters. These results were compared with reported AS values for other animal species from studies that also used SUPPA on Iso-Seq reads. Results reported were compiled for the zebrafish (Danio rerio) (Nudelman et al., 2018), the goldfish (Carassius auratus auratus) (Gan et al., 2021), the Wuchang bream (Megalobrama amblycephala) (Chen et al., 2021), the whiteleg shrimp (Litopenaeus vannamei) (Zhang et al., 2019), and the cave nectar bat (Eonycteris spelaea) (Wen et al., 2018).

Genome annotation

The N. macropterus genome was annotated using the MAKER v2.31.10 (Holt & Yandell, 2011) pipeline, using only the complex repeats for hard masking. SNAP v2013.11.29 (Korf, 2004) and Augustus v3.3.1 (Stanke et al., 2004) were used for ab initio gene prediction. Gene predictions were also inferred from the TARdn2 Iso-Seq transcripts and from protein homology by using protein sequences of zebrafish (D. rerio), three-spined stickleback (Gasterosteus aculeatus), spotted gar (Lepisosteus oculatus), Nile tilapia (Oreochromis niloticus), medaka (Oryzias latipes), Japanese puffer (Takifugu rubripes), green spotted puffer (Tetraodon nigroviridis), and southern platyfish (Xiphophorus maculatus) that were downloaded from Ensembl release version 103 (Kersey et al., 2016). Functional annotation of predicted proteins was performed with blast +v2.6.0 against the NCBI nonredundant protein sequences database (NR) and with InterProScan v5.50-84.0 (Jones et al., 2014). Finally, low-quality genes shorter than 50 amino acids were identified with AGAT v0.6.0 (Dainat, 2021) and filtered out. Genes with an incomplete open reading frame (ORF) were flagged. Genome annotation was also inspected visually with JBrowse v1.1.10 (Skinner et al., 2009). See all details of the genome annotation pipeline in Supplementary File 1: Methods.

Results

Genome sequencing

Illumina sequencing reads filtering (i.e. quality, contamination, and mitochondria) resulted in a final data set of 54.91 Gb short reads (Table 1) with a c. 92x depth of coverage. The GC content was 43%, and the overall sequence read quality was high. Both forward and reverse reads passed all the FastQC criteria; that is, they were never flagged for poor quality (Supplementary Fig. 3 in Supplementary File 2). Although there was a small bias in in base sequence contents of the first c. 10 bases, this was expected because of the nonrandom nature of the hexamer priming step during sequencing (Hansen et al., 2010). This slight deviation from uniformity in sequence content was not considered an issue because there is no quantitative step involved in the analyses based on the short reads. Nanopore sequencing, filtering, and trimming resulted in 9.18 million reads (73.39 Gb), or 122x coverage, with an average read length of 8 Kb (Table 1), a mean read quality of 7.9, and an N50 length of 9.5 Kb. A total of 285,997 CCS HiFi reads (4.01 Gb) and 91,602 repeat-free, nonredundant, high-quality Iso-Seq transcripts (312.31 Mb) were also obtained.

Assemblies comparison and quality assessment

The Flye assembly reduced the number of scaffolds by more than half compared with the CABOG assembly (Table 2). The scaffold
N50 length of the Flye assembly was almost twice as long, and the number of complete BUSCOs was higher. The Flye assembly size was also more consistent with the haploid genome size pre-estimated by k-mer counting (c. 520 Mb) than the CABOG assembly size. Interestingly, the Flye assembly also corrected a misassembly of the first scaffold of the CABOG assembly (see below). Polishing the Flye assembly resulted in the correction of 43,080 substitution errors and 42,783 deletion errors. The polished assembly had the same number of scaffolds and contigs, but a few hundred fewer bases, and one missing BUSCO was recovered into an additional single-copy BUSCO. The hifiasm assembly performed on the HiFi reads did not produce satisfactory results compared with the Illumina + Nanopore hybrid assemblies, with 6–10 times more scaffolds, an N50 length 50 times smaller, and a BUSCO completeness lower than 90%. This was most probably due to the low coverage of HiFi reads (c. 6.5x) used for this sequencing trial.

Approximately 99.7% of Illumina reads could be mapped back to the CABOG assembly, and 99.8% to both Flye assemblies, making the Flye assemblies slightly more accurate according to that metric. The Flye polished assembly had a slightly higher proportion of “proper-pairs” reads mapped (86.23%) than the unpolished assembly (85.7%). FRC curves showed that both Flye assemblies were more accurate than the CABOG assembly (Fig. 3). Moreover, while both the unpolished and polished Flye assemblies produced a very similar curve, for the same genome coverage, the polished Flye assembly always had a slightly lower number of cumulative errors than the unpolished assembly (Supplementary Fig. 4 in Supplementary File 2). While there was evidence of the presence of unmerged haplotigs in the CABOG assembly (Fig. 4a), no evidence was detected in the Flye polished assembly (Fig. 4b); thus, a filtering step was not required. Trailing Ns were not present in the Flye polished assembly either.

Interestingly, the longest scaffold of the CABOG assembly, scaffold 1, was 5 Mb longer than the longest scaffold of the Flye assembly (Table 2). Between-scaffolds alignment scores obtained from MashMap (Supplementary Fig. 5 in Supplementary File 2) were used to visualize a potential misassembly at that scaffold. The longest scaffold of the CABOG assembly corresponded indeed to the two longest scaffolds of the polished Flye assembly, scaffolds 1 and 2 (Fig. 5a). The CABOG scaffold 1 is highly likely to have been misassembled since it also corresponds to two long regions in two different linkage groups (i.e. chromosomes) in chromosome-level assemblies of both S. chuatsi and C. auratus. This is not the case for scaffold 1 in the polished Flye assembly (Fig. 5b). This supported the interpretation that the “correct” longest scaffold is the one from the polished Flye assembly.

The Flye polished assembly provided the best results and thus was used in all subsequent analyses. This final genome assembly consisted of 567,902,715 bases in 1,214 scaffolds, with a scaffold N50 length of 3.37 Mb and a proportion of gaps of 0.001% (Table 2, Fig. 6). The base composition was A: 28.10%, T: 28.15%, G: 21.87%, C: 21.88%, and overall standard deviation of GC content was 3.87%. The BUSCO completeness was very good overall, with >95% of the single-copy Actinopterygii orthologs retrieved in the final assembly (Table 2, Fig. 6). The contiguity and completeness were high compared with those of other Illumina + Nanopore hybrid assemblies (Table 3).

**Estimation of heterozygosity**

Variant calling of Illumina reads against the polished assembly resulted in a total of 3,654,819 SNPs. By dividing this number by the size of the genome, this corresponded roughly to a heterozygosity level of 0.64%. This is lower than the level estimated by k-mer frequency (c. 1.00%). However, it is common for heterozygosity estimated by k-mer frequency to be lower than estimated by called SNPs, because the SNP calling approach is more conservative (Thai et al., 2019). Nevertheless, the heterozygosity estimated for

![Fig. 3. FRC curves for the CABOG, Flye, and Flye polished assemblies of tarakihi. The Y-axis represents the cumulative size of the assembly and the X-axis is the cumulative number of potential errors (i.e. “features”). Assemblies for which the curves are steeper are considered more accurate.](image)

![Fig. 4. Read depth histograms of the tarakihi genome assemblies contigs, obtained by mapping the clean Nanopore reads back to the assembly. A unimodal distribution with a peak equal to the sequencing reads depth is expected for a haplotig-free assembly. Another peak at half the sequencing reads depth (arrow) is indicative of the presence of unmerged haplotigs. a) CABOG assembly and b) Flye polished assembly.](image)
TARdn1 is one of the highest reported for fish species. To our knowledge, this is the highest heterozygosity estimated for a fish through k-mer analysis, with other reported values ranging from 0.1% (Tibetan loach Triplophysa tibetana and Murray cod Maccullochella peelii) to 0.9% (Java medaka Oryzias javanicus) (Vij et al., 2016; Austin et al., 2017; Gong et al., 2018; Ge et al., 2019; Nguinkal et al., 2019; Yang et al., 2019; Lu et al., 2020; Takehana et al., 2020; Zhang et al., 2020; Zheng et al., 2021). Even the heterozygosity estimated through SNPs (0.64%) is high compared with estimations from other fishes using the same method [e.g. large yellow croaker: 0.36% (Wu et al., 2014), grass carp: 0.25% (Wang et al., 2015)]. This result is even more striking in that the variant analysis was very stringent in our case by retaining only high-quality bi-allelic SNPs. This reinforces the recent findings that N. macropterus is a species with a historically large population that displays a particularly high genetic diversity (Papa, Halliwell, et al., 2021).

### Repetitive elements and genes annotation

REs represented 30.45% of the genome or a total of 172,911,032 bp. Although the proportion of REs in fish genomes can vary greatly at scales from 10 to 60% (Yuan et al., 2018), the proportion of repeat
elements in *N. macropterus* is at par with the proportion observed in other Centrarchiformes [Largemouth bass (*Micropterus salmoides*): 33.79%; Big-eyed mandarin fish (*Siniperca knerii*): 26.55% (Lu et al., 2020; Sun et al., 2021)] and for Perciformes in general (Yuan et al., 2018). Of the REs known in the databases, interspersed repeats accounted for 27.62% of the genome, including 10.87% of DNA transposons and 6.17% of retro-elements (long interspersed nuclear elements (LINEs), long terminal repeat retrotransposons (LTR), short interspersed nuclear elements (SINEs), and Penelope (PLE), in that order. The rest of the repeat elements consisted of simple sequence repeats (Supplementary Table 2 in Supplementary File 2). After filtering for length, the final predicted gene set included 20,169 protein-coding genes with a mean length of 13,832 bp, among which 95.5% had a maximum Annotation Edit Distance (AED) < 0.5. An AED value of 0 indicates an exact match between the intron/exon coordinates of an annotation and the aligned transcriptome and proteins evidence, while an AED of 1 indicates no support from evidence. The mean exon length was 229 bp, and the mean intron length in CDS was 1,184 bp. More than 98% of the genes were functionally annotated by at least one of the two methods used (blastp, 98.2%; InterProScan, 82.8%).

### Iso-Seq analysis

Of the 93,949 full-length polished, nonredundant Iso-Seq transcripts, 2,347 were classified as REs and were filtered out from downstream analyses. For each of these RE transcripts, the main RE elements included DNA elements (801), LINEs (639), LTRs (464), SINEs (94), rRNAs (47), low complexity/simple repeats (33), rolling circles (26), satellites (16), and retroposons (2), as well as one LINE/LTR hybrid, and 224 unknown RE. The final non-RE Iso-Seq data set included 91,313 unique transcripts from 15,515 genes. The mean transcript per gene ratio was 5.89, with a median of 3 and a maximum of 211 (Fig. 7a). This is higher than the values recently reported for humans (3.62) and two species of bats (1.92).
and 1.49), but lower than for pharaoh ants (9) (Wen et al., 2018; Gao et al., 2020). Less than 5% of genes had more than 20 different transcripts. The predicted proteins of both genes that produced the most transcripts (respectively 211 and 164 transcripts) were collagen alpha chains isoforms (XP_006787735.1: collagen alpha-2(I) chain-like isoform X2, XP_020490299.1: collagen alpha-1(V) chain-like isoform X1), implicated in the structural integrity of the cellular matrix (GO:0005201).

A total of 26,644 AS events were detected in the tarakihi transcriptome (Fig. 7b). The most frequent AS event was the retention of intron (46%), while “alternative last exons” and “mutually exclusive exons” were the rarest (<1% each). Some examples of these AS events were visualized in the tarakihi genome (Fig. 8). Comparison of the frequency of AS events in the tarakihi with other species showed that the trends are globally similar across organisms (Fig. 9). Most organisms show relatively high occurrences of retained introns (RIs), alternative 3’ and 5’ splice sites (A3 and A5), alternative first exons (AF), and to a lesser degree skipping exons (SE), compared with alternative last exons (AL) and mutually exclusive exons (MX). Figure 9 also shows that tarakihi, goldfish, and cave nectar bat may have a better representation of the AS events proportions owing to a much deeper coverage than the Wuchang bream, zebrafish, and whiteleg shrimp (although values for MX and SE were not reported for the goldfish study). While it is the most common AS event in both tarakihi and goldfish, the proportion of RI events is much higher in tarakihi than in the proportion of other events. While intron retention was thought until recently to be the least prevalent AS form in animals, it is now clear that this is not the case (as shown in the studies in Fig. 9 but also e.g. Wang et al. 2019; Gao et al. 2020). RI events are widely used across organisms to tune down the levels of transcription of some genes in cells and tissues depending on their function (Braunschweig et al., 2014).
Genome size
The size of the tarakihi genome was consistent with values for fish genomes that have been reported so far. A recent review of publicly available fish genome assemblies (comprising 244 species) showed that the average genome length of fish is 872.64 Mb but varies between c. 300 Mb and c. 4.5 Gb (Fan et al., 2020). The genome size of *N. macroperus* (568 Mb) is several hundred Mb shorter than the two other published Centrarchichorme genomes, for the largemouth bass *M. salmoides* (964 Mb) and the big-eye mandarin fish *S. knerii* (732.1 Mb) (Lu et al., 2020; Sun et al., 2021). However, *N. macroperus* is still evolutionarily far apart from these two species. The largemouth bass and the big-eye mandarin fish both belong to the Centrarchoidei suborder, which is thought to have split from Cirrhitioidei at least 70 million years ago (Sancliangco et al., 2016).

Discussion
The advances in DNA sequencing technologies have made it clear how valuable reference genome assemblies are for the study of biology and conservation, resulting in a global effort to assemble the genomes of as many organisms as possible (Koopfli et al., 2015; Worley et al., 2017; Fan et al., 2020). Here, we present the first genome assembly of the tarakihi, a valuable commercial fisheries species, and the first representative out of the c. 60 species of the Cirrhitioidei suborder to have a whole genome sequenced. While performing a hybrid assembly of Illumina and Nanopore reads and allowing phased genomes. Nonetheless, the present genome assembly has already been successfully used as a resource for population structure analyses and the study of adaptive selection (Papa, Morrison, et al., 2022) and in the context of comparative genomics (Papa, Wellenreuther, et al., 2022). The highly accurate transcriptome will also surely be a valuable resource for future studies.

Data availability
All genomic sequences and associated metadata are deposited on the Genomics Aotearoa repository (https://repo.data.nesi.org.nz/) under projects “Tarakihi genome” (https://doi.org/10.57748/aabm-5y92) and “Tarakihi transcriptome” (https://doi.org/10.57748/xkhy-3t51). All scripts used in the analyses are openly available on GitHub at https://github.com/yvanpapa/tarakihi_genome_assembly.

Supplemental material available at G3 online.

Acknowledgments
The authors are thankful to Igor Ruza, David Ashton, and Matthew Wylie (Plant & Food Research, Nelson) for assisting in the sampling of the captive-bred specimen and to Nick Johnston for capturing and providing the wild specimen. They are also grateful for advice and assistance from the Removing Fisheries Juvenile Habitat Bottleneck Technical Advisory group, and in particular, they wish to acknowledge the support and consultation from Laws Lawson (Te Ohu Kaimoana), Jeremy Helson (Fisheries Inshore New Zealand), and Carol Scott (Southern Inshore Fisheries Management Company Limited). They thank Tom Oosting and Holly Jackson (Victoria University of Wellington) for proofreading the manuscript.

Author contributions
YP conceptualized the study, prepared the methodology, did software and formal analysis, validated and investigated the study, collected resources, did data curation, wrote the original draft, wrote, reviewed, and edited the article, visualized the study. MW collected resources, wrote, reviewed, and edited the article, supervised the study, acquired funds. MAM wrote, reviewed, and edited the article, supervised the study, acquired funds. PAR conceptualized the study, collected resources, wrote, reviewed, and edited the article, supervised the study, involved in project administration, acquired funds.

Literature cited
Fisheries New Zealand. Fisheries Assessment Plenary: Stock Assessment and Stock Status Volume 3: Pipi to Yellow-Eyed Mullet. New Zealand: Ministry for Primary Industries; 2018.

Aljanabi SM, Martinez I. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res. 1997;25(22):4692–4693. https://doi.org/10.1093/nar/25.22.4692.

An D, Cao HX, Li C, Humbeck K, Wang W. Isoform sequencing and State-Of-Art applications for unravelling complexity of plant transcriptomes. Genes (Basel). 2018;9(1):43. https://doi.org/10.3390/genes9010043.

Andrews S. FastQC: A quality control tool for high-throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Austin CM, Tan MH, Harrison KA, Lee YP, Croft LJ, Sunnucks P, Pavlova A, Gan HM. 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. https://doi.org/10.1093/gigascience/gix063.

Babcock RC, Bustamante RH, Fulton EA, Fulton DJ, Haywood MDE, Hobday AJ, Kenyon R, Matear RJ, Plagányi EE, Richardson AJ, et al. Severe continental-scale impacts of climate change are happening now: extreme climate events impact marine habitat forming communities along 45% of Australia’s Coast. Front Mar Sci. 2019;6:411. https://doi.org/10.3388/fmars.2019.00411.

Benestan L. Population genomics applied to fishery management and conservation. In: Oleksiak M and Rajora O, editors. Population Genomics: Marine Organisms. Berlin: Springer; 2019. p. 399–421. https://doi.org/10.1007/13836_2019_66.

Funding
This work was supported by the Ministry of Business, Innovation and Employment (CO1X1618) and a Victoria University of Wellington Doctoral Scholarship.

Conflicts of interest
None declared.
Bernatchez L, Wellenreuther M, Araneda C, Ashton DT, Barth JMI, Beacham TD, Maes GE, Martinsohn JT, Miller KM, Naish KA, et al. Harnessing the power of genomics to secure the future of seaweed. Trends Ecol Evol. 2017;32(9):665–680. https://doi.org/10.1016/j.tree.2017.06.010.

Braunschweig U, Barbosa-Moraes NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 2014;24(11):1774–1786. https://doi.org/10.1101/gr.177790.114.

Burrows MT, Schoeman DS, Buckley LB, Moore P, Poloczkanska ES, Brander KM, Brown C, Bruno JF, Duarte CM, Halpern BS, et al. The pace of shifting climate in marine and terrestrial ecosystems. Science. 2011;334(6056):652–655. https://doi.org/10.1126.science.1210288.

Bushnell B. BBMap Short Read Aligner. Berkeley: University of California; 2018. http://sourceforge.net/projects/bbmap.

Byrne A, Cole C, Volden R, Vollmers C. Realizing the potential of full-length transcriptome sequencing. Philos Trans R Soc B Biol Sci. 2019;374(1786):20190097. https://doi.org/10.1098/rstb.2019.0097.

Chaliss R. rjchallis/assembly-stats 17.02. Zenodo; 2017. https://doi.org/10.5281/zenodo.322347.

Chen Y et al. Genom-wide integrated analysis revealed functions of lincRNA–miRNA–mRNA interaction in growth of intermascular bones in Megalobrama amblycephala. Front Cell Dev Biol. 2021;8:603815. https://doi.org/10.3389/fcell.2020.603815.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–175. https://doi.org/10.1038/s41392-020-01056-5.

Dinan T. AGAT: Another Gff Analysis Toolkit to handle annotations of the Indian blue peacock (Pavo cristatus) genome using Oxford Nanopore technology and Illumina sequencing. GigaScience. 2019;8(5):1–13. https://doi.org/10.1093/gigascience/giz038.

Fan G, Song Y, Yang L, Huang X, Zhang S, Zhang M, Yang X, Chang Y, Zhang H, Li Y, et al. Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). GigaScience. 2020;9(8):1–7. https://doi.org/10.1093/gigascience/giaa080.

Feron R, Zahm M, Cabau C, Klopp C, Roques C, Bouchez O, Eché C, Vailière S, Donnadieu C, Haffray F, et al. Characterization of a Y-specific duplication/insertion of the anti-Mullerian hormone type II receptor gene based on a chromosome-scale genome assembly of yellow perch, Perca fluviatilis. Mol Ecol Resour. 2020;20(2):531–543. https://doi.org/10.1111/mecr.15313.

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 2020;117(17):9451–9457. https://doi.org/10.1073/pnas.1902461117.

Gar W, Chung-Davidson YW, Chen Z, Song S, Cui W, He W, Zhang Q, Li W, Li M, Ren J. Global tissue transcriptomic analysis to improve genomic annotation and unravel skin pigmentation in goldfish. Sci Rep. 2021;11(1):1–14. https://doi.org/10.1038/s41598-020-80168-6.

Gao Y, Xi F, Liu X, Wang H, Reddy AS, Gu L. Single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) in plants: the status of the bioinformatics tools to unravel the transcriptome complexity. Curr Bioinform. 2019;14(7):566–573. https://doi.org/10.2174/157489361466191024151746.

Gao Q, Xiong Z, Larsen RS, Zhou I, Zhao J, Ding G, Zhao R, Liu C, Ran H, Zhang G. High-quality chromosome-level genome assembly and full-length transcriptome analysis of the pharaoh ant Monomorium pharaonis. GigaScience. 2020;9(12):1–14. https://doi.org/10.1093/gigascience/giaa143.

Ge H, Lin K, Shen M, Wu S, Wang Y, Zhang Z, Wang Z, Zhang Y, Huang Y, Zhou C, et al. De novo assembly of a chromosome-level reference genome of red-spotted grouper (Epinephelus aaraka) using nanopore sequencing and Hi-C. Mol Ecol Resour. 2019;19(6):1461–1469. https://doi.org/10.1111/1755-0998.13064.

Gong G, Dan C, Xiao S, Guo W, Huang P, Xiong Y, Wu J, He Y, Zhang J, Li X, et al. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. GigaScience. 2018;7(11):1–9. https://doi.org/10.1093/gigascience/giy120.

Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12):e131. https://doi.org/10.1093/nar/gkq224.

Hoang NV, Henry JR. Iso-Seq long read transcriptome sequencing. In: Cifuentes A, editors. Comprehensive Foodomics. Amsterdam: Elsevier; 2021. p. 486–500. https://doi.org/10.1016/b978-0-08-100596-5.22729-7.

Holt C, Yandell M. MAKER2: an annotation pipeline and genome database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12(1):491. https://doi.org/10.1186/1471-2105-12-491.

Jain C, Koren S, Dhillon A, Phlippy AM, Aluru S. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics. 2018;34(17):i748–i756. https://doi.org/10.1093/bioinformatics/bty597.

Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien F-A, Swinkel W, Koelewijn A, Palstra AP, Pelster B, Spahn HP, et al. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. Sci Rep. 2017;7(1):7213. https://doi.org/10.1038/s41598-017-07650-6.

Jiang JB, Quatrini AM, Francis WR, Ryan JF, Rodriguez E, McFadden CS. A hybrid de novo assembly of the sea pansy (Renilla muelleri) genome. GigaScience. 2019;8(4):1–7. https://doi.org/10.1093/gigascience/giz026.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka, G, et al. InterProScan 5: genome-scale protein function classification bioinformatics. Bioinformatics. 2014;30(9):1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kadobianskyi M, Schulze L, Schuelke M, Juddewicz B. Hybrid genome assembly and annotation of Danionella translucida. Sci Data. 2019;6(1):156. https://doi.org/10.1038/s41597-019-0161-z.

Kersey PJ, Allen JE, Armean I, Boddou S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Lij, Grabmueller C, et al. Ensemble genomes 2016: more genomes, more complexity. Nucleic Acids Res. 2016;44(D1):D574–D580. https://doi.org/10.1093/nar/gkv1209.

Kimura K, Imamuro H, Kawai T. Comparative morphology and phylogenetic systematics of the families Cheilodactylidae and Latridae (Perciformes: Cirrhitoidae), and proposal of a new classification. Zootaxa. 2018;4536(1):1–72. https://doi.org/10.11646/ zootaxa.4536.1.1.

Koepfli KP, Paten B, O’Brien SJ, Antunes A, Belov K, Bustamante C, Castoe TA, Clawson H, Crawford AJ, Diekians M, et al. The genome 10K project: a work forward. Annu Rev Anim Biosci. 2015;3(1):57–111. https://doi.org/10.1146/annurev-animal-090414-014900.

Kolmogorov M, Yuan J, Lin Y, Pesvner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–546. https://doi.org/10.1038/s41587-019-0072-8.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):695–700. doi:10.1038/nbt.2280.
Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5(1):59. https://doi.org/10.1186/1471-2105-5-59.

Langley AD. Stock assessment of tarakahi off the east coast of mainland New Zealand [New Zealand Fisheries Assessment Report 2018/05]. Ministry for Primary Industries; 2018.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lu L, Zhao J, Li C. High-quality genome assembly and annotation of the big-eye mandarin fish (Siniperca kneri). G3 (Bethesda). 2020;10(3):877–880. https://doi.org/10.1534/g3.119.400930.

Ludt WB, Burridge CP, Chakraborty P. A taxonomic revision of Cheilodactylidae and Latridae (Centrarchiformes: Cirrhitoidei) using morphological and genomic characters. Zootaxa. 2019;4585(1):121–141. https://doi.org/10.11646/zootaxa.4585.1.7.

Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–770. https://doi.org/10.1093/bioinformatics/btr011.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24(24):2818–2824. https://doi.org/10.1093/bioinformatics/btn548.

Ngiunkal JA, Brunner RM, Verleih M, Hansen MP, Brunner RM, Ritchie PA. Comparative phylogeographic structure and historical demography of tarakihi (Nemadactylus macropterus) and king tarakihi (Nemadactylus n. sp.) in New Zealand. N Z J Mar Freshw Res. 2022;56(2):247–271. https://doi.org/10.1080/00288330.2021.1912119.

Papa Y, Morrison MA, Wellenreuther M, Ritchie PA. Chromosome-level genome assembly for the largemouth bass (Micropterus salmoides) using long-read sequencing. Genome Res. 2018;28(9):1415–1425. https://doi.org/10.1101/gr.223586.117.

Sun C, Li J, Dong J, Niu Y, Hu J, Lian J, Li W, Li J, Tian Y, Shi Q, et al. Genome sequence of the euryhaline javafish medaka, Oryzias latipes. Genome Res. 2021;31(9):1291–1305. https://doi.org/10.1101/gr.623566.120.

Papa Y, Halliwell AG, Morrison MA, Wellenreuther M, Ritchie PA. Phylogeographic structure and historical demography of tarakahi (Nemadactylus macropterus) and king tarakihi (Nemadactylus n. sp.) in New Zealand. N Z J Mar Freshw Res. 2022;56(2):247–271. https://doi.org/10.1080/00288330.2021.1912119.

Papa Y, Morrison MA, Wellenreuther M, Ritchie PA. Genomic stock structure of the marine teleost tarakahi (Nemadactylus macropterus) provides evidence of potential fine-scale adaptation and a temperature-associatedcline amid panmixia. Front Ecol Evol. 2022;10.862930. https://doi.org/10.3398/fev.2022.862930.

Papa Y, Oosting T, Valenza-Troubat N, Wellenreuther M, Ritchie PA. Genetic stock structure of New Zealand fish and the use of genomic markers in fisheries management: an overview and outlook. N Z J Zool. 2021;21(1):212–225. https://doi.org/10.1101/1755-0998.13240.

Ramos JE, Pecl GT, Moltschaniwskyj NA, Semmens JM, Souza CA, Strugnell JM. Population genetic signatures of a climate change driven marine range extension. Sci Rep. 2018;8(1):1–12. https://doi.org/10.1038/s41598-018-27351-y.

Race ES, Green RE. New approaches for genome assembly and scaffolding. Annu Rev Anim. 2019;7(1):17–40. https://doi.org/10.1146/annurev-animal-020518-115344.

Roch MJ, Schmidt SA, Borneman AR. Purge Haplotags: allelic contig reassembly for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460. https://doi.org/10.1186/s12859-018-2485-7.

Roberts CD, Stewart AL, Struthers CD. The Fishes of New Zealand. Roberts CD, Stewart AL, Struthers CD, editors. Wellington: Te Papa Press, 2015.

Sanciangco MD, Carpenter KE, Betancur R. Phylogenetic placement of enigmatic percomorph families (Teleostei: Percormorphaceae). Mol Phylogenet Evol. 2016;94:565–576. https://doi.org/10.1016/j.molphy.2015.10.006.

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. https://doi.org/10.1093/bioinformatics/btv351.

Simson WB, Parham FF, Papenfuss TJ, Lam AW, Henderson JB, Brian Simson W, Parham FF, Papenfuss TJ, Lam AW, Henderson JB. An annotated chromosome-level reference genome of the red-eared slider turtle (Trachemys scripta elegans). Genome Biol. Evol. 2020;12(4):456–462. https://doi.org/10.1093/gbe/evaa063.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19(9):1630–1638. https://doi.org/10.1101/gr.094607.109.

Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org.

Stanke M, Steinkamp R, Waack S. Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32(Web Server):W309–W312. https://doi.org/10.1093/nar/gkh379.

Sun C, Li J, Dong J, Niu Y, Hu J, Lian J, Li W, Li J, Tian Y, Shi Q, et al. Genome sequence of the euryhaline javafish medaka, Oryzias latipes: a small aquarium fish model for studies on adaptation to salinity. G3 (Bethesda). 2020;10(3):907–915. https://doi.org/10.1534/g3.119.400725.

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):1–6. https://doi.org/10.1093/gigascience/gix137.

Thai BT, Lee YP, Gan HM, Austin CM, Croft LJ, Trieu TA, Tan MH. Whole genome assembly of the snout otter clam, Lutraria rhynchaea, using Nanopore and Illumina data, benchmarked against bivalve genome assemblies. Front Genet. 2019;10:1158. https://doi.org/10.3389/fgene.2019.01158.

Thomson AI, Archer FI, Coleman MA, Gajardo G, Goodall-Copestake WP, Hoban S, Laikre L, Miller AD, O’Brien D, Pérez-Espana S, et al. Charting a course for genetic diversity in the UN Decade of Ocean Science. Evol Appl. 2021;14(6):1497–1518. https://doi.org/10.1111/eva.13224.

Trincado JL, Entizec BF, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyras E. SUPPA2: fast, accurate, and uncertainty-aware differential
Wood DE, Narzisi G, Langmead B. Improved metagenomic analysis with Feature Response Curves: GAGE and assemblaths. PLoS One. 2012a,7(12):e52210. https://doi.org/10.1371/journal.pone.005220.

Vezzi F, Narzisi G, Mishra B. Reevaluating assembly evaluations with MiniKraken2 v2 8GB database. Johns Hopkins University; 2019.

Wiley G, Miller MJ. A highly contiguous genome for the golden-fronted woodpecker (Melanerpes aurifrons) via hybrid Oxford Nanopore and short read assembly. GigaScience. 2020;10(6):1829–1836. https://doi.org/10.1093/gigascience/giy116.

Zimin AV, Crepeau G, Purushothaman K, et al. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017; 27(5):787–792. https://doi.org/10.1101/gr.219405.116.

Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020;16(6):1–8. https://doi.org/10.1371/journal.pcbi.1007981.

Zimin AV, Stevens KA, Crepeau MW, Purushothaman K, et al. Improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience. 2017,6(1):1–4. doi:10.1093/gigascience/giw016.

Communicating editor: A. McCallion