Genomics 116 (2024) 110929

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Research Paper

Peering into the gaps: Long-read sequencing illuminates structural variants and genomic evolution in the Australasian snapper

Julie Blommaert [a,*], Jonathan Sandoval-Castillo [b], Luciano B. Beheregaray [b], Maren Wellenreuther [a,c]

[a] The New Zealand Institute for Plant and Food Research, Nelson, New Zealand
[b] Molecular Ecology Laboratory, College of Science and Engineering, Flinders University, Bedford Park, South Australia, Australia
[c] School of Biological Sciences, The University of Auckland, Auckland, New Zealand

ARTICLE INFO

Keywords:
Genetic resources
Sparidae
Transposons
Centromeres
Genetic variation
Fisheries genomics

ABSTRACT

Even before genome sequencing, genetic resources have supported species management and breeding programs. Current technologies, such as long-read sequencing, resolve complex genomic regions, like those rich in repeats or high in GC content. Improved genome contiguity enhances accuracy in identifying structural variants (SVs) and transposable elements (TEs). We present an improved genome assembly and SV catalogue for the Australasian snapper (*Chrysophrys auratus*). The new assembly is more contiguous, allowing for putative identification of 14 centromeres and transfer of 26,115 gene annotations from yellowfin seabream. Compared to the previous assembly, 35,000 additional SVs, including larger and more complex rearrangements, were annotated. SVs and TEs exhibit a distribution pattern skewed towards chromosome ends, likely influenced by recombination. Some SVs overlap with growth-related genes, underscoring their significance. This upgraded genome serves as a foundation for studying natural and artificial selection, offers a reference for related species, and sheds light on genome dynamics shaped by evolution.

1. Introduction

Species management, conservation, and breeding programme outcomes can be greatly enhanced with the consideration of genomic data [52,71]. Prior to the routine and widespread availability of DNA sequencing, genetic insights into wild populations and breeding programmes were gained through allozyme studies, eventually moving on to higher resolution markers (e.g. microsatellites and single-nucleotide polymorphisms), and most recently to whole genomes. Each of these technologies has provided some level of knowledge about genetic diversity in populations, how it segregates and is inherited, and this has been fundamental to many conservation, management, and breeding programmes [12]. Most genome-wide data available to date has been produced via high-throughput, short-read sequencing, which often leads to fragmented assemblies especially around GC-rich and repeat-rich regions (e.g. [20]). The increased accessibility and accuracy of long-read sequencing technologies has lead to an increase in genome quality allowing researchers to explore more complex variants in these difficult regions [52].

Single-nucleotide polymorphisms (SNPs) are commonly the main or only type of variant that have thus far been included in population and functional genomic studies, however, it is becoming increasingly clear that structural variants (SVs) can provide additional insights into genomic function [70,71]. Where SVs and SNPs have both been included in a single study, SVs have typically been found to impact a greater portion of the genome than SNPs [14,76,77], and impact phenotypes [53]. However, given that SVs are more complex than SNPs, technological limitations often preclude their detection, especially when SVs are larger than and cannot be spanned by short sequencing reads [40]. Additionally, SVs are often enriched in conjunction with more repetitive regions of the genome, such as satellite repeats, centromeres and transposable elements (TEs) [8]. Biologically, this can be due to the activity of TEs and the errors that small regions of homology in the genome can cause during the process of DNA repair or recombination [48]. Technologically, the repetitive nature of some of these regions can also increase mis-mapping of (short) sequencing reads, introducing a potential source of error in variant calling [40]. The use of long-read sequencing in producing genome assemblies, and even in population

* Corresponding author.
E-mail address: julie.blommaert@plantandfood.co.nz (J. Blommaert).

https://doi.org/10.1016/j.ygeno.2024.110929
Received 5 May 2024; Received in revised form 25 August 2024; Accepted 26 August 2024
Available online 29 August 2024

resequencing, can reduce these errors, and increase confidence in SV calls. The routine generation of highly contiguous reference genomes (e.g., [20,52,65]) now allows researchers to go beyond genomic analyses based on SNPs and create large catalogues of SVs within and between species, even in less studied taxa.

Within the teleost family Sparidae (seabreams), there are currently genome sequences available for seven of around 130 seabream species (∼5 % of Sparidae species) on NCBI, and only two of these are chromosome level assemblies (*Sparus aurata* (NCBI accession GCA_900880675.2) and *Acanthopagrus latus*, [39]). Moreover, neither of the chromosome assemblies include reported locations of centromeres, despite previous publication of centromeric satellite sequence from *S. aurata* [25]. Of the available genomes, even those for commercially important species, such as *Pagrus major* are still highly fragmented [59] and no genome assemblies are available for the six species in this family listed as endangered or critically endangered [32].

The Australasian snapper (*Chrysophrys auratus,* hereafter referred to as snapper) is an ecologically [47], and culturally [58] important species of the family Sparidae that supports significant recreational and commercial fisheries [21] in New Zealand and Australia. Both historical and present-day climate change and anthropogenic activities, such as extractive fisheries, have put pressures on this species [5,30,51,58,66]. These pressures have led to recent declines in several stocks across the species range and have become the focus of fisheries management plans and recovery efforts [18,23]. There is some evidence of fisheries induced size-selection on wild snapper stocks [30]. In addition to changes in species abundance, changes in species range have been recorded, possibly related to fluctuations in temperature [51]. Management of snapper populations in Australia and New Zealand must consider all these factors, but current methods employed to assess fisheries are not only labour and cost intensive but also have an inherent degree of uncertainty [7,10,11,46]. In addition, snapper has been identified as a promising candidate to diversify the aquaculture sector and has been bred in New Zealand in captivity since 1994, with genomic selection for improved growth being integrated into the breeding since 2016 [2,3,14,53–55]. Knowledge of the population structure and the extent of genetic diversity that is segregating and underpinning economically and ecologically important traits provides fundamental information to support breeding decisions in snapper. This knowledge, however, relies heavily on having a high-quality genome assembly for this species. With recent advances in sequencing technologies and associated downstream analyses, we can now obtain whole genome datasets that include more comprehensive catalogues of SVs and TEs.

In this study, we present an improved genome assembly from a species within the teleost family, Sparidae by using long-read data. We then use this enhanced assembly to 1) improve genome annotation, focusing on genes, TEs and centromeres; 2) analyse whole-genome data from wild snapper and map the intraspecific diversity of structural genomic variation across the genome; and 3) examine the interaction between SVs, TEs and evolutionary forces which may impact their distribution in the genome. This new assembly provides an important general molecular resource for Sparidae and a specific one for snapper in Australia and New Zealand, where it is both an emerging candidate for aquaculture, and a species of fisheries and indigenous importance. Our study exemplifies an approach to discover novel genetic variants that might underpin phenotypic traits and fitness in fisheries resources and is therefore of relevance to fisheries management efforts and to studies of genomic evolution.

2. Methods

2.1. Sample collection and sequencing

To obtain long-read data, we collected high molecular weight DNA from liver tissue from a female snapper kept in captivity at The New Zealand Institute for Plant and Food Research Limited (PFR) finfish facility in Nelson. This tissue was first stored in RNALater and then digested in G2 digestion buffer with proteinase K at 50 °C for 80 min prior to DNA extraction. DNA was extracted using a chloroform:isoamyl alcohol extraction followed by sodium acetate/isopropanol precipitation. Following this, DNA was suspended in Tris-EDTA buffer. DNA was sequenced on a PacBio (Sequel II) at the McDonnell Genome Institute at Washington University, USA, and produced 38.9 Gbp of continuous long reads (3,978,058 reads). This amounts to approximately 48× coverage of the genome (Fig. 1A).

2.2. Genome assembly and assessment

Long-read PacBio sequencing reads were overlapped using minimap2 v2.22 with -ava-pb flag and then assembled using miniasm v0.2 [37] under default settings. This assembly was then polished using Racon v1.4.7 [67]. Additionally, the reads were assembled using flye v2.8.3 [35] under default settings. This assembly was then processed with purge_dups v1.1.2 [28] with a low coverage cutoff of 3, mid coverage cutoff of 40 and a high coverage cutoff of 85. Following this, previously reported Bionano sequencing, which was NLRS data from a different individual fish, (Fig. 1C, [14]) was used to further improve the new genome assembly. This was done by performing de novo assembly and hybrid scaffolding within the Bionano Access software suite. The resulting genome assembly was then screened for contaminants using blobtoolkit v4.1.4 [15], and any contigs that had no blast hits and met at least one of the following additional criteria were removed- shorter than 5000 bp, GC content inconsistent with the rest of the assembly, or coverage inconsistent with the rest of the assembly. Mitochondrial DNA was searched for by BLAST, but none was identified, likely due to previous steps focusing on coverage.

The genome assembly was assessed initially via QUAST v4.2 [29] and BUSCO v5.2.2 [41] using the BUSCO Actinopterigii gene set (3640 genes). To enable comparison to the genome assembly reported in [14] (Fig. 1B), RagTag v2.1.0 scaffold [1] with default settings was used to place the contigs and scaffolds from this study onto the linkage groups reported there. For further analyses, we proceeded with this scaffolded version of the genome assembly. The genome assembly reported here will be referred to as chrAur2_scaffold, while the previously published version will be referred to as chrAur1.

To assess synteny between this newly produced genome assembly, and that of another species in the same family (Sparidae), we compared the new assembly with the one from the closely related yellowfin seabream (*Acanthopagrus latus*). To achieve this, the two genomes were aligned using MUMMER [42] under default settings and visualised using the circlize [27] package in R [50]. Based on this synteny, existing linkage group names were updated to be consistent with chromosome numbering of other published Sparidae genomes.

2.3. Genome annotation

The programme LiftOff v1.6.3 [60] was used to lift gene annotations from chrAur1 [14] and *A. latus* [39] to chrAur2_scaffold using default settings. Transposable elements were identified and classified using RepeatModeller2 [22] within the dfam te-tools container v1.4 and the ltr option. The resulting library was then used to mask each assembly with RepeatMasker [61] within the dfam te-tools container v1.4. A total of 137 SNPs that have been previously identified as contributing to growth in snapper [3,53,55] were located in the new genome assembly using SNPlift v1.0.4 [43].

2.4. Centromere annotation

Putative centromeres were identified using two approaches. First, we used the RepeatObserver [19] package to predict centromere position using DNA walks to detect repeats in the genome and produce repeat diversity measures. These DNA walks are transformed into measures of

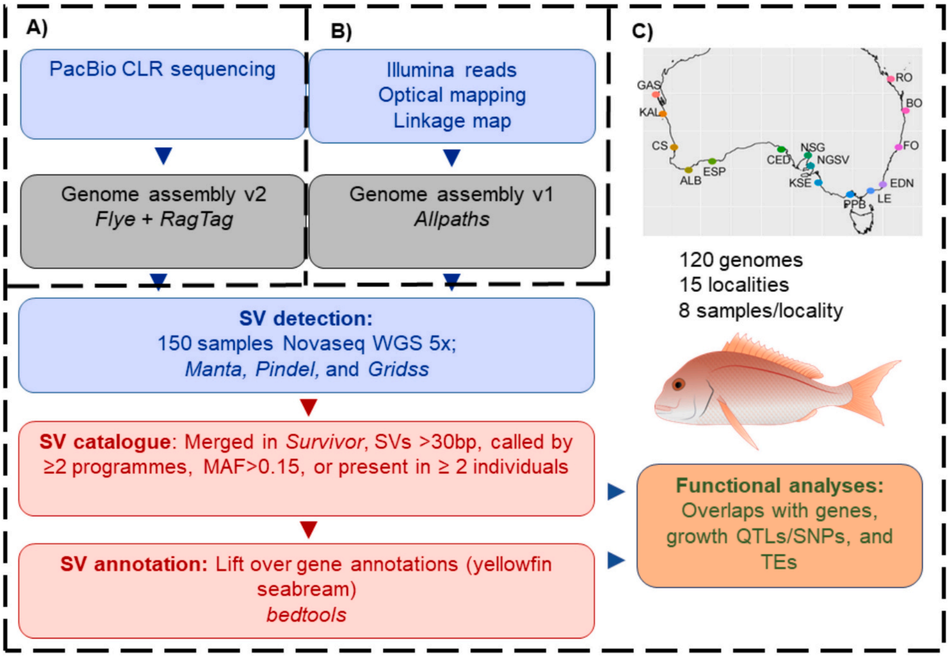
A)
PacBio CLR sequencing
Genome assembly v2
Flye + RagTag
B)
Illumina reads
Optical mapping
Linkage map
Genome assembly v1
Allpaths
C)
GAS
KAL
CS
ALB
ESP
CED
NSG
NGSV
KSE
PPB
LE
EDN
FO
BO
RO
120 genomes
15 localities
8 samples/locality
SV detection:
150 samples Novaseq WGS 5x;
Manta, Pindel, and Gridss
SV catalogue: Merged in Survivor, SVs >30bp, called by ≥2 programmes, MAF>0.15, or present in ≥ 2 individuals
SV annotation: Lift over gene annotations (yellowfin seabream)
bedtools
Functional analyses:
Overlaps with genes, growth QTLs/SNPs, and TEs

Fig. 1. An overview of the DNA sequencing and assembly methods used to detect structural variants (SVs). A) The assembly strategy for the new genome; B) The assembly strategy for the previously published genome [14] used for comparison; C) The sampling strategy to obtain snapper from around Australia, resequence them, and call SVs.

repeat diversity, including Shannon Diversity, across each chromosome of the assembly presented in this study. Long stretches of Ns appeared to produce artefacts in the Shannon Diversity index along the chromosome, therefore we reduced long stretches of Ns to 10 consecutive Ns for this analysis. In the RepeatObserver output, the summarised centromere positions did not always agree with the Shannon diversity plots. Depending on the genomic context, different measures from RepeatObserver give a more reliable centromere prediction (C. Elphinstone, pers. comm.). So here, centromere positions were predicted from the windows output by the multiple Shannon diversity approaches. A putative centromere was identified where all centromere predictions based on Shannon diversity windows, excluding outliers, were within 5 kbp of each other.

Second, we used CentroMiner from the quarTeT toolkit [38] to identify putative centromeres based on identification of repetitive element types. CentroMiner was run on the chromosomes of the assembly using default settings. We visually summarised the results of RepeatObserver and CentroMiner and added information about the gene and TE density across each chromosome in 1 Mbp windows to provide additional support for putative centromeres.

2.5. Resequencing, read mapping and variant calling

For the SVs detection, short-read resequencing of 120 fish sampled from 15 locations across Australia (Fig. 1C) was performed. The samples were obtained from commercial and recreational fisheries and are a subset of the samples used on a ddRAD reduced genome representation study [9–11]. DNA was extracted using a modified salting-out protocol [62], and libraries were prepared by Novogene (Hong Kong) using a NEBNext DNA Library Prep Kit (350 bp). Whole genome sequencing was done on NovaSeq X with PE 150 bp runs resulting in an average coverage of 5× per individual. The sequences were trimmed, and quality filtered using AdapterRemoval v2.3.1 [57]. The sequencing reads which passed quality control were mapped to both chrAur1 [14] and chrAur2_scaffold using SNAP [75], and individual BAM files were used to detect SVs (Fig. 1A and B).

We employed Manta v1.6 [17], Pindel v2.0.5 [72], and Gridss v2.13.2 [13] algorithms, each utilising different information from PE short reads. These methods were chosen to provide more confidence in SV detection. The results from each algorithm were merged using the Survivor package v1.0.7 [33]. In each individual sample, SVs larger than 30 bp, and called by two or more algorithms were selected. Although 50 bp is a conventional cutoff for SVs detection, we selected 30 bp since, at this size, SVs were still consistently detected by 2 or more algorithms, and short SVs may be as biologically relevant as those >50 bp. These SVs were then filtered based on a minor allele frequency (MAF) greater than 0.015, or presence in at least 2 individuals. This resulted in a catalogue of filtered SVs.

The detected SVs were functionally annotated using assembly annotation and bedtools v2.30.0 [49]. We applied three different gene overlap thresholds: any overlap (1 or more bp), 20 % of the gene overlap, and the entire gene contained within the SV. The annotation was broadly classified into annotated genes, unclassified transcripts, and non-coding RNA. The same overlap thresholds were used to classify SV overlap with TEs, and TE categories taken into consideration were DNA transposons, Long Interspersed Nuclear Elements (LINEs), Long Terminal Repeats (LTRs), Rolling Circles (RCs), Short Interspersed Nuclear Elements (SINEs), and Unknown. To compare these overlaps to those expected by random chance, genomic features were randomly shuffled and compared to SVs using bedtools shuffle and bedtools overlap 1000 times and averages compared to observed values.

We used ANGSD v.0938 [36] to detect and filter genome-wide SNPs, and then to estimate the genotype likelihood of each sample. These genotype likelihoods were used to calculate linkage with ngsLD v1.1.1 [24]. The average D^2 per sliding window of 1 Mbp was calculated using an in-house script available at https://github.com/Yuma248/MELFUwgrs. A model selection was run to determine which variable- LD, position on chromosome, or mapping quality- had the most explanatory power for number of SVs per window.

In order to analyse how many additional SVs could be annotated due to genome assembly improvements from gap filling, we first found N-gaps over 100 bp in chrAur1 [14], extended their coordinates by 1000 bp to each side using bedtools slop. This enabled lifting these windows over to chrAur2_scaffold using LiftOff [60]. Then, SVs were counted using bedtools intersect using default values.

3. Results

3.1. A more contiguous snapper genome assembly

The data and pipeline used here have resulted in a more contiguous genome assembly than chrAur1 (Supplementary Fig. 1 A, Supplementary Table 1). The genome assembly produced by flye was more contiguous than that produced by miniasm and resulted in 1023 contigs totalling 757.78 Mbp which were able to be placed onto 364 scaffolds (N50: 33.1 Mbp), including 24 linkage groups, based on chrAur1. After scaffolding, the orphan linkage group (LG25) reported chrAur1 has largely been eliminated in chrAur2_scaffold. In chrAur1, LG25 represented 3.82 Mbp of the genome, and in chrAur2_scaffold it represents 33 kpb of the genome. chrAur2_scaffold was 97.6 % complete when Benchmarking Universal Single-Copy Orthologs (BUSCO) genes were considered. Of these complete BUSCO genes, 30 (0.8 %) were duplicated, and an additional 32 BUSCO genes (0.9 %) were found in fragmented copies (Supplementary Table 1). This genome assembly also led to a reduction in N-gaps, where these represented ~8 % of chrAur1, but only 2 % of chrAur2_scaffold. In chrAur1, there were 14,609 N-gaps over 100 bp in length, while in chrAur2_scaffold there were 1370.

3.2. Synteny analyses

Overall, chrAur2_scaffold displays high synteny to chrAur1 [14] (Supplementary Fig. 2B), with each chrAur1 linkage group having an average of 97.8 % and a minimum of 92.2 % coverage with chrAur2_scaffold chromosomes. However, some chromosomes showed structural differences between the two genome assembly versions (e.g., chromosome 4). Additionally, there is remarkable broad scale synteny between the snapper and the three other high quality Sparidae genome assemblies (Supplementary Fig. 1B and 2). Because of this high synteny, we have used the *A. latus* genome, as a reference for gene lift over and to name the chromosomes, which are different to the linkage group numbers reported in chrAur1. On average, 93.3 % of the base pairs per chromosome of the chrAur2_scaffold genome aligned to the *A. latus* genome assembly. In almost all chrAur2_scaffold chromosomes, at least 90 % of base pairs aligned to those in the *A. latus* genome, with only chromosome 14 aligning slightly less than 90 % (88.9 %).

3.3. Genome annotation

Of the 30,057 features classed as genes in the *A. latus* genome annotation, 26,115 were able to be lifted over to the ChrAur2_scaffold assembly. Of these, 22,054 were protein coding and 4002 were lncRNAs (Supplementary Table 2, Fig. 2). Other small RNAs (e.g., tRNA, snoRNA, etc.) were lifted over, though not as completely as genes (Supplementary Table 2). Of the 36,397 annotation genes in ChrAur1, 35,282 were lifted over to ChrAur2_scaffold. Originally, 33,863 were placed on LGs [14], and in ChrAur2_scaffold, 34,928 annotations were placed on chromosome scaffolds. However, since none of these include functional annotations or protein evidence, the annotations lifted over from *A. latus* were used in subsequent analyses. In the ChrAur2_scaffold assembly, 54 BUSCO genes (49 complete, 2 duplicate, 3 fragmented) could not be found in the annotation lift over set. Overall, this represents a fairly complete gene annotation set, though there are annotations that could not be transferred from the *A. latus* genome annotation and a small number of BUSCO genes found in the assembly that do not overlap with any lifted over gene annotations.

Improving the genome assembly saw marginal increases in transposon annotation, with approximately 3 % more of the genome being covered by TEs than in the previously published genome assembly [14]. These increases were across all TE types, including those unclassified. The largest relative increase was in the LTRs, which accounted for 0.74 % of the previous genome assembly, but now represent 1.45 % of the genome assembly, nearly double the proportion (Supplementary Table 3, Fig. 2). TEs seemed largely evenly distributed across chromosomes, with a few hotspots (e.g. a small peak on chromosome 2). There also seemed to be a trend for TEs to increase gradually in density towards the chromosome ends.

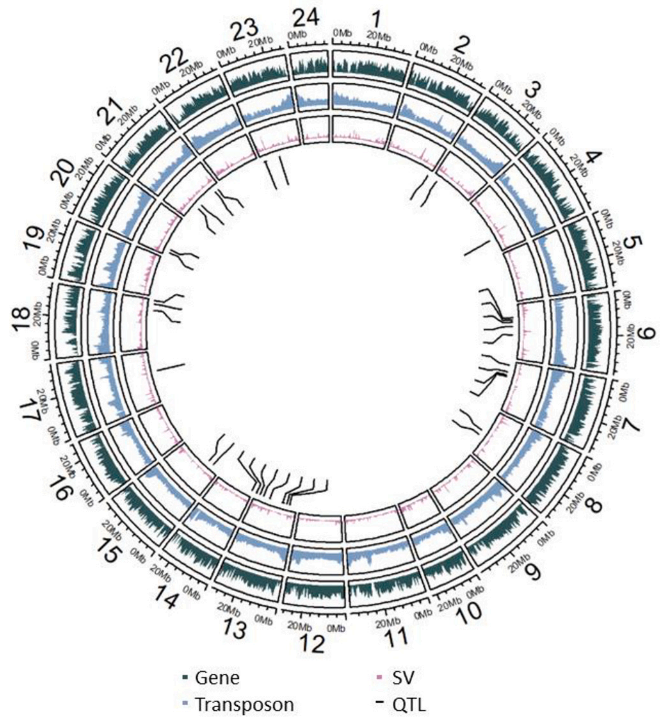

Fig. 2. A circos plot of the gene density, transposable element (TE) density, and structural variant (SV) density across the *Chrysophrys auratus* genome assembly as well as markers for quantitative trait loci (QTLs) and single-nucleotide polymorphisms (SNPs) previously found to be implicated in growth.

Of the 137 SNPs implicated in growth in previous work (Ashton et al., 2019; [53,55]) 125 SNPs were lifted over from old to new assemblies, 54 of which were on chromosome 13 (Fig. 2). Chromosome 13 corresponds to the linkage group 16 in chrAur1, and this linkage group has been strongly implicated in growth by the previous work (Ashton et al., 2019; [53,55]).

3.4. Identification of putative centromeres

By combining the results from RepeatObserver and CentroMiner we were able to putatively identify centromeres in a number of chromosomes with varying levels of confidence. Overall, the estimated centromere positions were smaller with CentroMiner (mean: 0.14 Mbp) than RepeatObserver (mean: 1.98 Mbp). On four chromosomes (chromosomes 2, 14, 16, and 21), the midpoints of the putative centromeres from both methods were within 5 Mbp of each other (Supplementary Table 4, red asterisk on Supplementary Fig. 3). However, the expected pattern of decreased gene density and increased TE density at putative centromere sites does not occur at any of these four sites. For nine chromosomes (chromosomes 3, 5, 11, 12, 17, 18, 19, 22, 24- blue asterisk on Supplementary Fig. 3), one centromere detection method predicted a centromere at a chromosomal location where the gene density decreased, and TE density increased. However, on four of these chromosomes (chromosomes 5, 11, 12, 24), there was a centromere prediction from either RepeatObserver or CentroMiner which was further than 5 Mbp from the other prediction and also did not align with a pattern of increased gene density and decreased TE density. On chromosome 9, the predicted centromere locations from CentroMiner and RepeatObserver are 25.2 Mbps apart, but both correspond with the gene and TE density patterns expected around centromeres. Two chromosomes (chromosomes 1 and 4) did not produce any predicted centromeres from either two software, but each have a location on the chromosome displaying the TE/gene density pattern that may occur around centromeres. A further 8 chromosomes (chromosomes 6, 7, 8, 10, 13, 15, 20, and 23) had only one predicted centromere location and this location did not display the expected pattern of TE and gene density. However, in some cases (e.g. chromosome 7), the density of both features dropped at the putative centromere. For the 13 chromosomes where putative centromeres could be identified with some confidence, six could be said to be acrocentric, five metacentric, and a further two submetacentric based on these putative centromere locations.

The output of CentroMiner also includes satellite sequences identified at putative centromeres. When these were compared to previously reported *S. aurata* centromeric satellite reported [25] via pairwise alignments, the maximum sequence identity was 37 % and the mean 26 %.

3.5. Improved variant discovery

The number of detected SVs ranged from over 160,000 by Manta to just over 2 million in Gridss. After merging and filtering for high-quality SVs, the catalogue contained 96,890 SVs. Among these, there were 31,946 deletions (DEL), 1591 duplications (DUP), 22,260 inversions (INV), 7041 insertions (INS), and 34,052 translocations (TRA). While most of these SVs were smaller than 500 bp, a substantial number of them exceeded 1 Mb. In total, they covered approximately 23.5 Mb, which represents around 3.1 % of the genome.

In comparison with running the same variant detection pipeline on the previous assembly, we detected around 35,000 more SVs (Fig. 3A, Table 1). Of these ~35,000 SVs, 6568 were in or near an N-gap that was lifted over from chrAur1. This represents around 18 % of the additional SVs. While the number of indels remained similar, we observed more than double the number of duplications, inversions, and translocations (Fig. 3A, Table 1). Additionally, the new assembly allowed for an improved detection of large structural variants, with more than double the number of SVs larger than 1 Mb (Fig. 3B, Supplementary Table 5). This is reflected by the fact that the increase in the number of structural variants was only 57.5 %, while there was a 2.3-fold increment in genome coverage.

Table 1
Number of structural variants (SVs) detected in each version of the snapper genome assembly.

	chrAur1	chrAur2_scaffold
Deletions	30,908	31,946
Duplications	664	1591
Inversions	9311	7041
Insertions	7534	22,260
Translocations	13,116	34,052
Total SV	61,533	96,890
# bp covered by SVs	7,138,750	23,558,660
% of genome covered by SVs	0.9502	3.1359

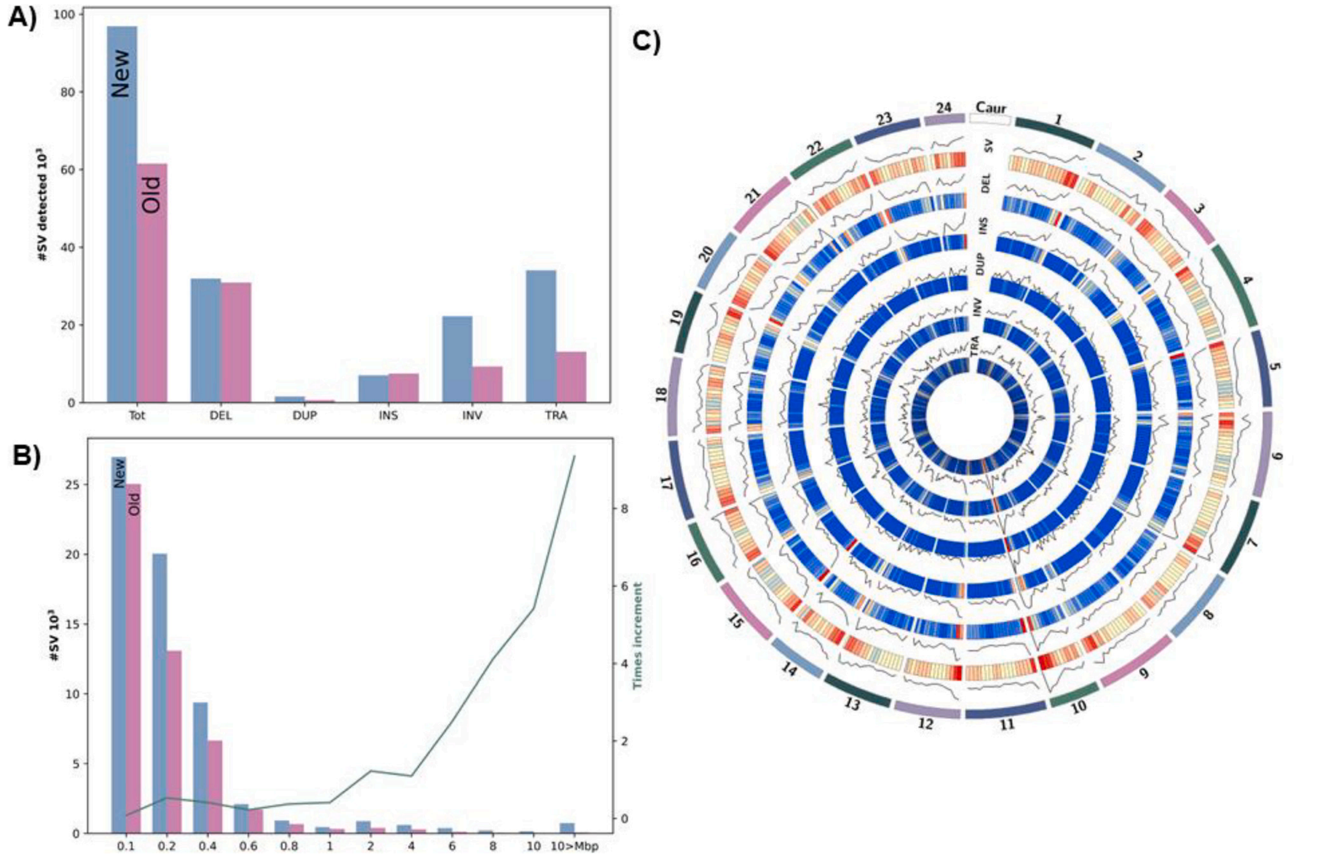

Fig. 3. A) The number of structural variants (SVs) called in the new versus old *Chrysophrys auratus* genome assemblies, shown as a total and by SV type. B) The number of SVs called in the new versus old *C. auratus* genome assemblies by SV size. C) The distribution of SVs across the *C. auratus* genome, both overall and split by SV type.

3.6. Structural variant distribution and overlap with other genomic features

The SVs were evenly distributed across the 24 chromosomes (Fig. 3C, 4, Supplementary Table 6). However, within chromosomes, SV numbers (but not bp occupied; Supplementary Fig. 4) increased towards chromosome ends (Fig. 4). This pattern was consistent across all types of SVs, but not all sizes. The increase in SV number towards chromosome ends was only observed in smaller (under 100 bp) SVs (Supplementary Fig. 5).

A total of 78,649 SVs overlapped with genes. Considering any type of overlap, 19,077 unique genes were affected by one or more SVs. This number reduced to 904 genes that were entirely covered by an SV. These figures respectively represent approximately 57.8 % and 2.7 % of the total genes identified in the snapper genome (Fig. 5A; Supplementary Table 7). When considering any overlap, the random permutations showed more overlaps than observed, but when considering 20 % overlap or 100 % overlap, the random permutations showed significantly fewer overlaps than what was observed. The observed overlaps fell outside of the ranges produced in the random permutations. While some of these genes are transcripts with unknown functions, the majority are well-characterised genes.

Furthermore, of the filtered SVs, 32 were found within 1000 bp of SNPs identified as important growth quantitative trait loci (QTLs) in previous studies (Fig. 5B). Excluding translocations, for which sizes were not reported, these SVs had an average size of 804 bp, a median of 104 bp, minimum of 51 bp and a maximum of 5431 bp. Twelve were deletions, 3 duplications, 2 insertions, 6 inversions, and 9 translocations.

Considering TEs and SVs, both showed a distribution pattern where their density increased towards the chromosome ends. Surprisingly, the overlaps between SVs and TEs were lower or comparable to SV overlaps with genes. Overall, 6.21 % of TEs had any overlap with SVs (Fig. 5A; Supplementary Table 8), but when considering 20 % or 100 % overlap, the proportion of TEs was comparable to the proportion of genes with this same overlap (4.07 % and 2.80 % respectively; Fig. 5A; Supplementary Table 8). In all cases (any overlap, 20 % overlap, 100 % overlap), the random permutations showed significantly more overlaps than what was observed, with the observed overlaps falling outside of the ranges produced in the random permutations. SV number and TE number were significantly correlated with each other ($p < 0.05$; Fig. 5C). To determine if recombination was a common causal factor for the SV and TE distribution patterns, we considered the explanatory power of linkage disequilibrium (LD), distance from the chromosome end, and mapping quality on SV number. LD had the strongest explanatory power with chromosome position and mapping quality making small contributions to explanatory power (Supplementary Table 9).

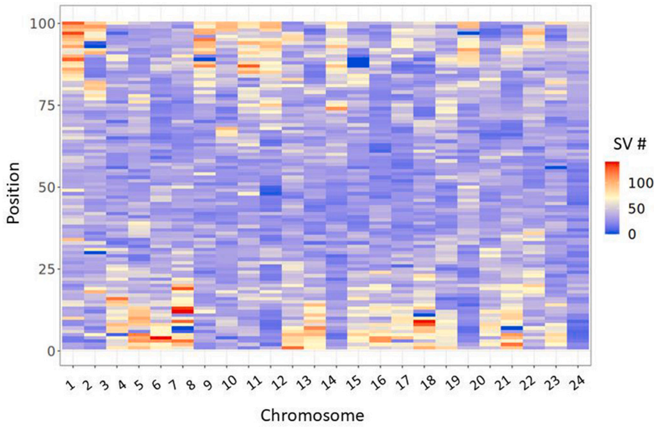

Fig. 4. The distribution of structural variant (SV) numbers across each chromosome, shown with all chromosomes scaled to the same size and window sizes set to 1 % of chromosome.

4. Discussion

In this study, we focused on improving the genome assembly and annotation of a teleost from the family Sparidae, the Australasian snapper (*C. auratus*), using long-read sequencing data and analysing whole-genome data from wild snapper to map the diversity of structural genomic variation. This work resulted in significant improvements in the genome assembly, which provided valuable insights into non-genic genome annotations and the structural variants (SVs) present in the snapper genome, with wider relevance to the family Sparidae. Not only is the genome assembly presented here more contiguous, but the scaffolds have fewer N-filled gaps, and we saw an improvement in the assembly of repetitive regions, while quality and completeness of genic regions stayed high. Additionally, we putatively identified the location of 14 centromeres and noted that SVs and TEs were distributed across chromosomes in a similar manner and that this pattern is likely distributed by the evolutionary effects of recombination.

4.1. The improved genome assembly increases the known catalogues of genetic variation fourfold

The genome assembly presented here represents a substantial improvement over the snapper genome version 1 [14], resulting in increased contiguity and completeness. The quality of this genome assembly is comparable to the current best available assemblies for the Sparidae family [39]. The more contiguous genome assembly provides a better reference for future studies and facilitates the identification of structural variants and genes relevant to snapper biology. The low percentage of duplicated and fragmented BUSCO genes suggests a high level of completeness in the gene annotation, providing a basis for enhancing the understanding of the genetic basis of snapper traits and functions.

The high synteny observed between the chrAur2_scaffold genome assembly and the previous version (chrAur1) supports the reliability and accuracy of the new assembly. The reduction in the percentage of Ns in chrAur2_scaffold, and the near complete resolution of a formerly unplaced orphan linkage group (LG25) represent significant improvements in genome assembly completeness and have lead to an increase in TE discovery and SV annotation, as well as identification of putative centromere locations. However, further data is required to understand if the structural differences identified between chrAur2_scaffold and chrAur1 suggest either assembly errors or biological differences between the individuals used to produce these two genome assemblies. The high nucleotide-level synteny between the Australasian snapper and *A. latus* provides confidence in the overall genome assembly and structural arrangements. Given the evolutionary distance between these species [56], the high synteny is notable and has enabled the lift over of a high proportion of genes (both protein-coding and lncRNAs). The total gene number is consistent with previous snapper gene annotation work [69]. Neither the BUSCO gene annotation or the gene lift over from *A. latus* suggest that there are any major gene duplications or losses. Future work should focus on functional annotation of the *C. auratus* genome to a similar standard as *A. latus* incorporating transcriptome and proteome data [39].

The overall increase in TE annotation in this assembly suggests that the repetitive and difficult to assemble regions of the genome showed the greatest improvements with this application of long read sequencing, though quantifying this improvement with the LTR index [45] was still not possible due to the low number of LTRs. Transposon annotation found a marked increase in LTRs, but it is still notable that less than 2 % of the snapper genome appears to consist of LTRs. That most of the classified transposons were DNA transposons is consistent with other fish species and recent work [46]. Future work should focus on TE evolution across Sparidae, the low number of LTRs in these

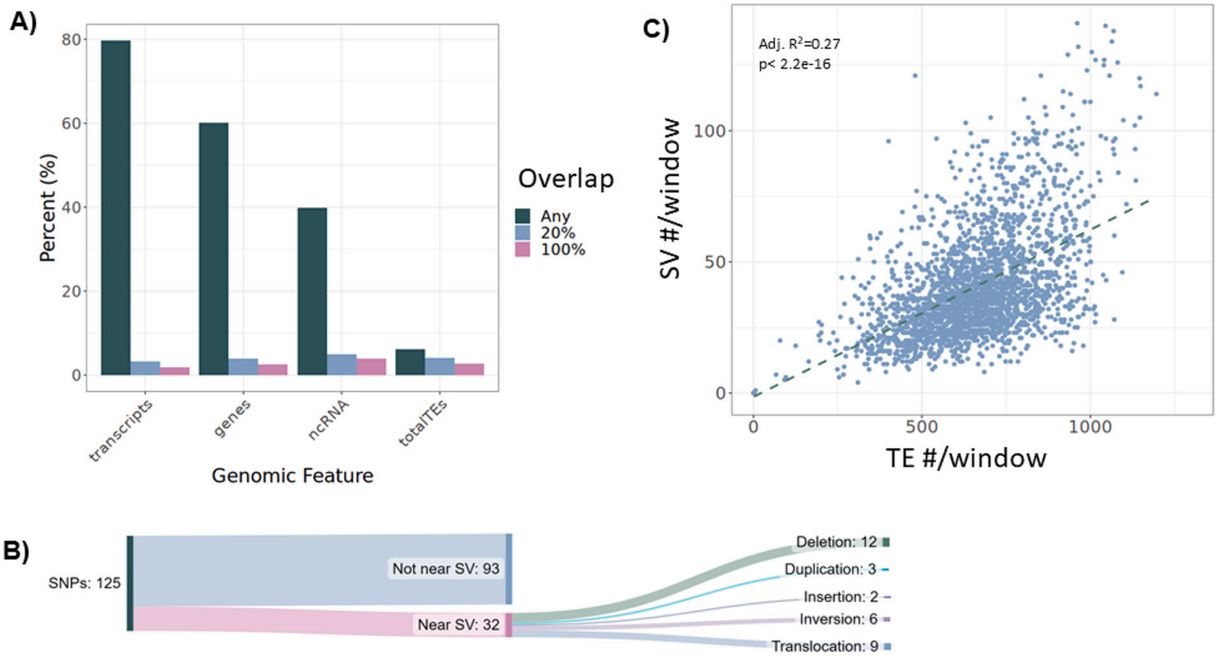

Fig. 5. A) The numbers of genes and transposable elements (TEs) that overlap with structural variants (SVs) at all, by more than 20 % or by 100 %. B) The number of growth single-nucleotide polymorphisms (SNPs) which are within 1000 bp of an SV, further broken down by SV type. C) A linear regression between the number of SVs and TEs per window shown in Fig. 4.

genomes, the interactions between TEs and other genomic features and variants, and the contribution of TEs to important traits like sex and growth.

While we could not confidently identify putative centromeres on all chromosomes, to our knowledge, this is the first report of putative centromere annotation in a sea bream genome assembly. While centromere satellite sequence has been reported in *S. aurata* [25], this has not been linked to locations in genome assemblies. We were able to predict centromere location with some degree of confidence for 14 chromosomes, and identify some centromere associated satellite sequences. There were ten chromosomes which either had no centromere predictions or had minimal support for the predictions. Considering the challenges of assembling and predicting centromeres only from DNA sequence, this is not unexpected. Of note, one chromosome (chromosome 9) had centromere predictions at two locations with approximately equal support. This potentially suggests that this chromosome may be an example of formation of a neocentromere or evolutionarily new centromere (ENC) on this chromosome. There are robust examples of neocentromere and ENC formation in other species [26,34]. While cytogenetic and epigenetic data are essential for confirming centromere location, many of the predictions presented here are supported by the identification of satellite DNA at the putative centromere locations. Since centromeres are epigenetically determined, further data that would support centromere identification would rely on karyotypic and ChIP-seq approaches targeted towards a species-specific centromere-specific histone H3 variant centromere protein A [63].

Centromeres are structurally complex regions of chromosomes, often containing repetitive sequences that pose challenges for genome assembly, making them particularly difficult to sequence and analyse accurately [63]. Identifying centromeres helps to pinpoint regions in the genome that play a crucial role in cell division, as they are responsible for ensuring proper chromosome segregation during mitosis and meiosis. Understanding the structure and organization of centromeres is hence essential for unravelling the mechanisms underlying chromosome inheritance and stability [6,31]. Considering the synteny across seabream genomes but the relatively low sequence identity between satellite centromeres reported in *S. aurata* and those found here, this would be an interesting avenue for future research. As in other taxa with conserved genomes, centromere changes may play a role in speciation of sea breams [6,34,68].

The detection and characterisation of more structural variants provides insights into the genetic diversity and potential adaptive functions within snapper. In the present work, SVs calling was compared across the improved and original assemblies, and the current assembly allowed for the identification of a greater number of SVs, especially large structural variants larger than 1 Mb. Nearly 20 % of the additional SVs called in chrAur2_scaffold were found in or near a location which had been a gap in chrAur1. Since locations of gaps in chrAur1 were determined by aligning the flanking regions of N-gaps, SVs were also counted if they fell within the windows that included these flanks. This likely overestimates the number of SVs in filled gaps. However, because our SV calling relies on mapping read pairs, some of the counted SVs had a read in the flanking region of the filled gap. These SVs would also have been missed in chrAur1, so the over-estimate may not be so drastic. This indicates that the enhanced genome quality and contiguity have enabled a more accurate and comprehensive assessment of genomic variation in the snapper genome.

Structural variants are increasingly found to be associated with complex traits, and form part of the polygenic architecture of traits such as those related to growth. Increasingly, this is being recognised and incorporated into breeding programmes [53,77], and conservation planning [71]. Previous work has identified SVs as a major source of genetic diversity in snapper compared with SNPs [14], and that SVs can play important roles in complex traits [53]. SV calls in this study were based on consensus calls across different algorithms to ensure that our relatively low-coverage resequencing per sample did not lead to false calls. This approach will also be applicable to other studies where resources must be carefully allocated, such as those in conservation biology or animal breeding programmes [71]. Here, we identified SV associations with genetic variants that are associated with growth in snapper, namely, three growth SNPs which overlapped directly with SVs, and a further 32 which were within 1000 bp of SVs, suggesting some combined contribution of these variants to growth. These new discoveries were enabled by access to previous QTL studies on snapper and showcase how studies can be combined to reveal new putative candidate associations for complex growth traits. Considering the relatively small number of growth SNPs included in this study, it is difficult to form conclusions regarding enrichment of SVs around growth SNPs and how this might interact with gene function, but future studies that combine SNP and SV associations could consider if the interaction of different types of variants is enriched in complex traits.

4.2. Chromosomal ends show accumulation of variants

The number of SVs increases in concert with TE density, increasing towards the ends of the chromosomes. Since LD seems to be a major explanatory variable in this SV distribution pattern, recombination is likely involved in SV formation in snapper, as reported in other species [8,48]. Correlations between SVs and TEs have been reported in other species (e.g. [16,39,78]), and can be explained by the tendency of TEs to cause SVs either directly by their movements around the genome, or indirectly through errors in recombination or DNA repair based on homology [48]. The interaction between TEs and SVs is complex and the role of recombination in this interaction may explain the pattern of TE and SV distribution observed here, where both increase towards the chromosome ends. Since mapping quality did not seem to offer much explanatory power for the distribution of SVs in this genome, it seems unlikely that the shared pattern of SV and TE distribution is a technical artefact caused by mis-mapping in repetitive regions. Other small SV and TE hotspots may be related to genomic features such as centromeres, the location of which remains to be confirmed with additional data. Various approaches can help untangle the impacts of recombination across the genome, while considering TEs (which can both be removed by recombination and also increase recombination rate locally), SVs (which can be caused by recombination or locally decrease recombination by reducing homology between chromosome pairs), and other factors such as centromeres, telomeres, and chromatin state. Various studies have leveraged high-quality, long read sequencing to investigate some of these interactions in the sub-telomeric regions especially, via population long-read sequencing (e.g. [4,73,74]) and pangenome maps (e.g. [44]). In addition to the population genomic based inferences here, linkage maps from breeding programmes (e.g. [14]), pangenome mapping, and gamete-based approaches such as those proposed in [48], can disentangle these interactions. These approaches linked to the resources available in snapper make this a promising area for future research.

4.3. Implications for fisheries management and breeding

Snapper has long been an important food source for the indigenous peoples of New Zealand and has been more recently targeted by commercial fisheries [21,47,58]. Additionally, this species is a promising future candidate for aquaculture, and has thus been captively bred in New Zealand under a selective breeding programme [2,3,14,54]. The wild populations in New Zealand and Australia have been under pressure from extractive fisheries, especially in modern times, and together with global climate change, populations are in decline and likely to change their geographic ranges [5,18,23,30,66].

Understanding the genomic landscape of wild populations provides insights into the genetic substrate available for adaptive evolution in this species [71]. Future work should seek to identify adaptive SVs to assess population health, estimate genetic diversity, and monitor the impact of environmental stressors on fisheries. We have added to the number of high-quality genome assemblies available for seabreams and provided a catalogue of SVs that can help guide genomics research across the *Sparidae* family more generally. These resources may be especially relevant for improving the fragmented *P. major* [59] genome assembly and assisting in population genomics of endangered related species [32].

Moreover, the information obtained from the improved genome assembly and variant analyses will be instrumental in enhancing the ongoing breeding programme for snapper aquaculture. Identifying genomic regions associated with important traits, such as sex, growth, disease resistance, and stress tolerance, will facilitate targeted breeding efforts to produce more resilient and economically valuable snapper stocks.

5. Conclusion

Uncovering the full extent of genomic variation that segregates in species provides fundamental insight into the genetic architecture underpinning phenotypic traits related to fitness, and crucial information to inform fisheries management. This discovery has long been hampered by technological limitations, but recent advances in sequencing technologies now allow this information to be garnered even in non-model species. This study represents a significant improvement in the genome assembly of the ecologically and culturally important species, snapper, which provides a greater understanding of the genomic variation and structural variants for this fishery and aquaculture species. This improved genome assembly, annotation, centromere identification on over half of the chromosomes, and comprehensive variant analyses provide valuable insights into the genetic diversity within this ecologically, economically, and culturally important species. The findings have important implications for fisheries management, conservation efforts, and aquaculture breeding programmes, enabling more informed decisions to sustainably manage this valuable resource in the face of ongoing environmental and anthropogenic pressures. Additionally, we have made note of and investigated the interaction between TEs, SVs, and recombination, which is often speculated about or overlooked in genomic studies. While the experiments required to further tease apart these relations are technically challenging, the resources (i.e. a breeding programme with pedigree and access to gametes for direct sequencing) are available for snapper and this represents a great opportunity to further dissect the basis of this relationship to offer greater insight into the evolution of recombination. As sequencing technologies continue to advance, further research will undoubtedly refine our understanding of the genomic intricacies of the Australasian snapper and other marine species.

CRediT authorship contribution statement

Julie Blommaert: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jonathan Sandoval-Castillo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luciano B. Beheregaray:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Maren Wellenreuther:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

None.

Data availability

Genomic Data: PacBio Sequence data and the genome assembly are located in the managed Aotearoa Genomic Data Repository (AGDR; [64]) and can be accessed via application at https://data.agdr.org.nz/. Benefits Generated: In recognition of the taonga status of snapper (tāmure), the associated sequencing data generated in this study have been deposited into AGDR, with access managed by application which relies on feedback from Māori as partners in scientific research.

Acknowledgements

We thank Elena Hilario and Igor Ruza for their help with high molecular DNA extraction, Chen Wu for bioinformatic advice, and Sara Montanari and Dafni Anastasiadi for proofreading. We also thank the rest of the Aquatic Phenomics and Genomics research group at Plant and Food Research for troubleshooting and discussions. We thank the New Zealand Ministry of Business, Innovation, and Employment for funding (SSIF Fund for Data Science under contract VUW RTVU1914 and "Accelerated breeding for advanced seafood production" C11X1603). We thank the Australian Research Council (LP180100756 grant to LBB) for financial support. We thank PacBio for funding long read sequencing through the PacBio SMRTGrant. We would like to thank Genomics Aotearoa for data management and hosting. We would also like to thank the three reviewers for their constructive and detailed comments which improved this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2024.110929.

References

[1] M. Alonge, et al., Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing, Genome Biol. 23 (1) (2022) 258. Available at: https://doi.org/10.1186/s13059-022-02823-7.

[2] D.T. Ashton, et al., Genetic diversity and heritability of economically important traits in captive Australasian snapper (Chrysophrys auratus), Aquaculture 505 (2019) 190–198. Available at: https://doi.org/10.1016/j.aquaculture.2019.02.034.

[3] D.T. Ashton, P.A. Ritchie, M. Wellenreuther, High-density linkage map and QTLs for growth in snapper (*Chrysophrys auratus*), G3 Genes|Genomes|Genetics 9 (4) (2019) 1027–1035. Available at: https://doi.org/10.1534/g3.118.200905.

[4] J. Audoux, et al., DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition, Genome Biol. 18 (2017), https://doi.org/10.1186/s13059-017-1372-2.

[5] R.C. Babcock, 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. 6 (2019), https://doi.org/10.3389/fmars.2019.00411 (Accessed: 31 July 2023).

[6] E. Balzano, S. Giunta, Centromeres under pressure: evolutionary innovation in conflict with conserved function, Genes 11 (8) (2020) 912, https://doi.org/10.3390/genes11080912.

[7] L. Benestan, Population genomics applied to fishery management and conservation, in: M.F. Oleksiak, O.P. Rajora (Eds.), Population Genomics: Marine Organisms, Springer International Publishing (Population Genomics), Cham, 2020, pp. 399–421, https://doi.org/10.1007/13836_2019_66.

[8] E.L. Berdan, et al., Unboxing mutations: connecting mutation types with evolutionary consequences, Mol. Ecol. 30 (12) (2021) 2710–2723, https://doi.org/10.1111/mec.15936.

[9] A. Bertram, Fisheries Genomics of Snapper (Chrysophrys auratus) in Australia, Flinders University, 2024.

[10] A. Bertram, et al., Fisheries genomics of snapper (Chrysophrys auratus) along the west Australian coast, Evol. Appl. 15 (7) (2022) 1099–1114, https://doi.org/10.1111/eva.13439.

[11] A. Bertram, et al., Bioregional boundaries and genomically-delineated stocks in snapper (*Chrysophrys auratus*) from southeastern Australia, bioRxiv (2023), https://doi.org/10.1101/2023.01.16.524335, p. 2023.01.16.524335.

[12] Y.X.C. Bourgeois, B.H. Warren, An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes, Mol. Ecol. 30 (23) (2021) 6036–6071. Available at: https://doi.org/10.1111/mec.15989.

[13] D.L. Cameron, et al., GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing, Genome Biol. 22 (1) (2021) 202, https://doi.org/10.1186/s13059-021-02423-x.

[14] A. Catanach, et al., The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost Chrysophrys auratus, Mol. Ecol. 28 (6) (2019) 1210–1223, https://doi.org/10.1111/mec.15051.

[15] R. Challis, et al., BlobToolKit – interactive quality assessment of genome assemblies, G3 Genes|Genomes|Genetics 10 (4) (2020) 1361–1374, https://doi.org/10.1534/g3.119.400908.

[16] N.-C. Chang, et al., Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression, Genome Res. 32 (7) (2022) 1408–1423, https://doi.org/10.1101/gr.275655.121.

[17] X. Chen, et al., Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications, Bioinformatics 32 (8) (2016) 1220–1222, https://doi.org/10.1093/bioinformatics/btv710.

[18] M. Drew, et al., Snapper Stock Assessment Report 2022, South Australian Research and Development Institute (Aquatic Sciences), Adelaide, 2022, p. 178.

[19] C. Elphinstone, et al., RepeatOBserver: Tandem Repeat Visualization and Centromere Detectio, 2023, https://doi.org/10.1101/2023.12.30.573697.

[20] G. Fan, et al., Initial data release and announcement of the 10,000 fish genomes project (Fish10K), GigaScience 9 (8) (2020), https://doi.org/10.1093/gigascience/giaa080 p. giaa080.

[21] Fisheries Assessment Plenary May 2023 Volume 3, Fisheries New Zealand, Available at: https://fs.fish.govt.nz/Doc/25496/85A%20SNAintro%202023.pdf.ashx, 2023 (Accessed: 31 July 2023).

[22] J.M. Flynn, et al., RepeatModeler2 for automated genomic discovery of transposable element families, Proc. Natl. Acad. Sci. 117 (17) (2020) 9451–9457, https://doi.org/10.1073/pnas.1921046117.

[23] A.J. Fowler, et al., The Status of Australian Fish Stocks Reports, Fisheries Research and Development Corporation, Snapper, 2021.

[24] E.A. Fox, et al., ngsLD: evaluating linkage disequilibrium using genotype likelihoods, Bioinformatics 35 (19) (2019) 3855–3856. Available at: https://doi.org/10.1093/bioinformatics/btz200.

[25] M.A. Garrido-Ramos, et al., Cloning and characterization of a fish centromeric satellite DNA, Cytogenet. Cell Genet. 65 (4) (1994) 233–237, https://doi.org/10.1159/000133637.

[26] Z. Gong, et al., Repeatless and repeat-based centromeres in potato: implications for centromere evolution[C][W], Plant Cell 24 (9) (2012) 3559–3574. Available at: https://doi.org/10.1105/tpc.112.100511.

[27] Z. Gu, et al., Circlize implements and enhances circular visualization in R, Bioinformatics 30 (19) (2014) 2811–2812, https://doi.org/10.1093/bioinformatics/btu393.

[28] D. Guan, et al., Identifying and removing haplotypic duplication in primary genome assemblies, Bioinformatics 36 (9) (2020) 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025.

[29] A. Gurevich, et al., QUAST: quality assessment tool for genome assemblies, Bioinform. (Oxford, Engl.) 29 (8) (2013) 1072–1075, https://doi.org/10.1093/bioinformatics/btt086.

[30] M. Heino, B. Díaz Pauli, U. Dieckmann, Fisheries-induced evolution, Annu. Rev. Ecol. Evol. Syst. 46 (1) (2015) 461–480, https://doi.org/10.1146/annurev-ecolsys-112414-054339.

[31] S. Henikoff, K. Ahmad, H.S. Malik, The centromere paradox: stable inheritance with rapidly evolving DNA, Science 293 (5532) (2001) 1098–1102, https://doi.org/10.1126/science.1062939.

[32] IUCN, The IUCN Red List of Threatened Species, Available at: https://www.iucnredlist.org, 2022 (Accessed: 28 March 2024).

[33] D.C. Jeffares, et al., 'Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast', *nature*, Communications 8 (1) (2017) 14061, https://doi.org/10.1038/ncomms14061.

[34] L.G. Kiazim, et al., Comparative mapping of the macrochromosomes of eight avian species provides further insight into their phylogenetic relationships and avian karyotype evolution, Cells 10 (2) (2021) 362, https://doi.org/10.3390/cells10020362.

[35] M. Kolmogorov, et al., Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (5) (2019) 540–546. Available at: https://doi.org/10.1038/s41587-019-0072-8.

[36] T.S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: analysis of next generation sequencing data, BMC Bioinform. 15 (1) (2014) 356, https://doi.org/10.1186/s12859-014-0356-4.

[37] H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, Bioinformatics 32 (14) (2016) 2103–2110, https://doi.org/10.1093/bioinformatics/btw152.

[38] Y. Lin, et al., quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification, Hortic. Res. 10 (8) (2023), https://doi.org/10.1093/hr/uhad127 p. uhad127.

[39] J. Lu, et al., Chromosome-level genome assembly of *Acanthopagrus latus* provides insights into salinity stress adaptation of Sparidae, Mar. Biotechnol. 24 (3) (2022) 655–660, https://doi.org/10.1007/s10126-022-10119-x.

[40] M. Mahmoud, et al., Structural variant calling: the long and the short of it, Genome Biol. 20 (1) (2019) 246, https://doi.org/10.1186/s13059-019-1828-7.

[41] M. Manni, et al., BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, Mol. Biol. Evol. 38 (10) (2021) 4647–4654, https://doi.org/10.1093/molbev/msab199.

[42] G. Marçais, et al., MUMmer4: A fast and versatile genome alignment system, PLoS Comput. Biol. 14 (1) (2018) e1005944, https://doi.org/10.1371/journal.pcbi.1005944.

[43] E. Normandeau, M. de Ronne, D. Torkamaneh, SNPLift: Fast and Accurate Conversion of Genetic Variant Coordinates Across Genome Assemblie, 2023, https://doi.org/10.1101/2023.06.13.544861.

[44] S. O'Donnell, et al., Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in Saccharomyces cerevisiae, Nat. Genet. 55 (8) (2023) 1390–1399, https://doi.org/10.1038/s41588-023-01459-y.

[45] S. Ou, J. Chen, N. Jiang, Assessing genome assembly quality using the LTR assembly index (LAI), Nucleic Acids Res. 46 (21) (2018) e126, https://doi.org/10.1093/nar/gky730.

[46] Y. Papa, et al., Genetic stock structure of New Zealand fish and the use of genomics in fisheries management: an overview and outlook, NZ J. Zool. 48 (1) (2021) 1–31, https://doi.org/10.1080/03014223.2020.1788612.

[47] D. Parsons, et al., Snapper (Chrysophrys auratus): a review of life history and key vulnerabilities in New Zealand, N. Z. J. Mar. Freshw. Res. 48 (2) (2014) 256–283, https://doi.org/10.1080/00288330.2014.892013.

[48] J.V. Peñalba, J.B.W. Wolf, From molecules to populations: appreciating and estimating recombination rate variation, Nat. Rev. Genet. 21 (8) (2020) 476–492, https://doi.org/10.1038/s41576-020-0240-1.

[49] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (6) (2010) 841–842, https://doi.org/10.1093/bioinformatics/btq033.

[50] R Core Team (n.d.) 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing.

[51] J.E. Ramos, et al., Population genetic signatures of a climate change driven marine range extension, Sci. Rep. 8 (1) (2018) 9558, https://doi.org/10.1038/s41598-018-27351-y.

[52] A. Rhie, et al., Towards complete and error-free genome assemblies of all vertebrate species, Nature 592 (7856) (2021) 737–746, https://doi.org/10.1038/s41586-021-03451-0.

[53] M. Ruigrok, et al., The relative power of structural genomic variation versus SNPs in explaining the quantitative trait growth in the marine teleost *Chrysophrys auratus*, Genes 13 (7) (2022) 1129, https://doi.org/10.3390/genes13071129.

[54] G. Samuels, et al., Generational breeding gains in a new species for aquaculture, the Australasian snapper (*Chrysophrys auratus*), Aquaculture 586 (2024) 740782, https://doi.org/10.1016/j.aquaculture.2024.740782.

[55] J. Sandoval-Castillo, L.B. Beheregaray, M. Wellenreuther, Genomic prediction of growth in a commercially, recreationally, and culturally important marine resource, the Australian snapper (*Chrysophrys auratus*), G3 Genes|Genomes|Genetics 12 (3) (2022), https://doi.org/10.1093/g3journal/jkac015.

[56] F. Santini, G. Carnevale, L. Sorenson, First multi-locus timetree of seabreams and porgies (Percomorpha: Sparidae), Ital. J. Zool. 81 (1) (2014) 55–71, https://doi.org/10.1080/11250003.2013.878960.

[57] M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging, BMC. Res. Notes 9 (1) (2016) 88, https://doi.org/10.1186/s13104-016-1900-2.

[58] F.V. Seersholm, et al., Subsistence practices, past biodiversity, and anthropogenic impacts revealed by New Zealand-wide ancient DNA survey, Proc. Natl. Acad. Sci. 115 (30) (2018) 7771–7776, https://doi.org/10.1073/pnas.1803573115.

[59] G.-H. Shin, et al., First draft genome for Red Sea bream of family Sparidae, Front. Genet. (2018) 9, https://doi.org/10.3389/fgene.2018.00643.

[60] A. Shumate, S.L. Salzberg, Liftoff: accurate mapping of gene annotations, Bioinformatics 37 (12) (2021) 1639–1643. Available at: https://doi.org/10.1093/bioinformatics/btaa1016.

[61] A.F.A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, Available at: http://www.repeatmasker.org, 2013.

[62] P. Sunnucks, D.F. Hales, Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae), Mol. Biol. Evol. 13 (3) (1996) 510–524. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a025612.

[63] P.B. Talbert, S. Henikoff, What makes a centromere? Exp. Cell Res. 389 (2) (2020) 111895 https://doi.org/10.1016/j.yexcr.2020.111895.

[64] B. Te Aika, et al., Aotearoa genomic data repository: an āhuru mōwai for taonga species sequencing data, Mol. Ecol. Resour. (2023), https://doi.org/10.1111/1755-0998.13866.

[65] The Darwin Tree of Life Project Consortium, Sequence locally, think globally: the Darwin tree of life project, Proc. Natl. Acad. Sci. 119 (4) (2022) e2115642118. Available at: https://doi.org/10.1073/pnas.2115642118.

[66] S.C. Urlich, S.J. Handley, From "clean and green" to "brown and down": a synthesis of historical changes to biodiversity and marine ecosystems in the Marlborough sounds, New Zealand, Ocean Coast. Manag. 198 (2020) 105349. Available at: https://doi.org/10.1016/j.ocecoaman.2020.105349.

[67] R. Vaser, et al., Fast and accurate de novo genome assembly from long uncorrected reads, Genome Res. 27 (5) (2017) 737–746, https://doi.org/10.1101/gr.214270.116.

[68] A. Voleníková, et al., Fast satellite DNA evolution in Nothobranchius annual killifishes, Chromosom. Res. 31 (4) (2023) 33, https://doi.org/10.1007/s10577-023-09742-8.

[69] M. Wellenreuther, et al., Domestication and temperature modulate gene expression signatures and growth in the Australasian snapper *Chrysophrys auratus*, G3: Genes|Genomes|Genetics 9 (1) (2018) 105–116, https://doi.org/10.1534/g3.118.200647.

[70] M. Wellenreuther, et al., Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification, Mol. Ecol. 28 (6) (2019) 1203–1209, https://doi.org/10.1111/mec.15066.

[71] J. Wold, et al., Expanding the conservation genomics toolbox: incorporating structural variants to enhance genomic studies for species of conservation concern, Mol. Ecol. 30 (23) (2021) 5949–5965, https://doi.org/10.1111/mec.16141.

[72] K. Ye, et al., Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads, Bioinformatics 25 (21) (2009) 2865–2871, https://doi.org/10.1093/bioinformatics/btp394.

[73] E. Young, et al., Comprehensive analysis of human subtelomeres by whole genome mapping, PLoS Genet. 16 (1) (2020) e1008347, https://doi.org/10.1371/journal.pgen.1008347.

[74] J.-X. Yue, et al., Contrasting evolutionary genome dynamics between domesticated and wild yeasts, Nat. Genet. 49 (6) (2017) 913–924, https://doi.org/10.1038/ng.3847.

[75] M. Zaharia, et al., Faster and more accurate sequence alignment with SNAP, arXiv (2011), https://doi.org/10.48550/arXiv.1111.5572.

[76] Y. Zhou, et al., The population genetics of structural variants in grapevine domestication, Nat. Plants 5 (9) (2019) 965–979. Available at: https://doi.org/10.1038/s41477-019-0507-8.

[77] Y. Zhou, et al., Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history, Genome Res. 32 (8) (2022) 1585–1601, https://doi.org/10.1101/gr.276550.122.

[78] K. Zhu, et al., A chromosome-level genome assembly of the yellowfin seabream (Acanthopagrus latus; Hottuyn, 1782) provides insights into its osmoregulation and sex reversal, Genomics 113 (4) (2021) 1617–1627, https://doi.org/10.1016/j.ygeno.2021.04.017.