Received: 7 August 2018 | Revised: 31 January 2019 | Accepted: 1 February 2019

DOI: 10.1111/mec.15051

SPECIAL ISSUE: THE ROLE OF GENOMIC STRUCTURAL VARIANTS IN ADAPTATION AND DIVERSIFICATION

WILEY MOLECULAR ECOLOGY

The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*

Andrew Catanach[1] | Ross Crowhurst[2] | Cecilia Deng[2] | Charles David[1] | Louis Bernatchez[3] | Maren Wellenreuther[4,5]

[1]The New Zealand Institute for Plant & Food Research Ltd, Lincoln, New Zealand

[2]The New Zealand Institute for Plant & Food Research Ltd, Auckland, New Zealand

[3]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec City, Quebec, Canada

[4]The New Zealand Institute for Plant & Food Research Ltd, Nelson, New Zealand

[5]School of Biological Sciences, University of Auckland, Auckland, New Zealand

Correspondence
Maren Wellenreuther, The New Zealand Institute for Plant & Food Research Ltd, Nelson, New Zealand.
Email: maren.wellenreuther@plantandfood.co.nz

Funding information
MBIE, Grant/Award Number: C11X1603

Abstract

Recent studies have highlighted an important role of structural variation (SV) in ecological and evolutionary processes, but few have studied nonmodel species in the wild. As part of our long-term research programme on the nonmodel teleost fish Australasian snapper (*Chrysophrys auratus*), we aim to build one of the first catalogues of genomic variants (SNPs and indels, and deletions, duplications and inversions) in fishes and evaluate overlap of genomic variants with regions under putative selection (Tajima's D and π), and coding sequences (genes). For this, we analysed six males and six females from three locations in New Zealand and generated a high-resolution genomic variation catalogue. We characterized 20,385 SVs and found they intersected with almost a third of all annotated genes. Together with small indels, SVs account for three times more variation in the genome in terms of bases affected compared to SNPs. We found that a sizeable portion of detected SVs was in the upper and lower genomic regions of Tajima's D and π, indicating that some of these have an effect on the phenotype. Together, these results shed light on the often neglected genomic variation that is produced by SVs and highlights the need to go beyond the mere measure of SNPs when investigating evolutionary processes, such as species diversification and adaptation.

KEYWORDS

CNV, indels, sex differences, SNP, standing variation, structural variation, whole-genome sequencing

1 | INTRODUCTION

Fundamental questions in evolutionary genetics concern the extent to which different mutational sources contribute to intraspecific genomic variation, and how this variation affects selection. In animals, the first wave of genomic information from the analysis of model species, such as human, mouse or *Drosophila*, revealed SNPs to be an important source of genetic variation (1000 Genomes Project Consortium, 2010; Langley et al., 2012; Wade et al., 2002). However, the ability to sequence multiple whole genomes of almost any species has changed this view recently and revealed that structural variation (SV) is not only taxonomically ubiquitous, but also more common than previously thought (Dobigny, Britton-Davidian, & Robinson, 2015; Feuk et al., 2005; Kirkpatrick, 2010). Indeed, recent comparative genomic studies in humans have shown that a large proportion of genetic differences are structural in nature, rather than substitutions involving SNPs, often outnumbering variation caused by SNPs in terms of total bases affected (1000 Genomes Project Consortium, 2015).

Recognized classes of SVs include rearrangements of DNA regions affecting the length or orientation of sequences in a genome, as well as sequences that are deleted, duplicated (copy number variants, CNV), inserted, inverted in orientation or translocated (Baker, 2012; Scherer et al., 2007). These variants are commonly subdivided into balanced SVs, with no losses or gains of genetic material, such as inversions or translocations within or between chromosomes, or unbalanced SVs, where part of the genome is lost or duplicated (i.e., a CNV). When SVs occur within the coding sequence of genes, they can affect the protein sequence, function and stability. SVs that encompass one or several coding units, as can be the case in deletions or duplications, can lead to overall changes in gene dosage. Furthermore, SVs affect gene expression in multifarious ways by disrupting the regulatory landscape (Spielmann & Mundlos, 2013).

We are undertaking a long-term research programme on the nonmodel marine teleost *Chrysophrys auratus* (Sparidae), commonly referred to as snapper, with the aim to develop this species for aquaculture. Species from this family are already successfully cultured in other parts of the world, including the Mediterranean region (*Sparus aurata)* and Japan (*Pagrus major*). Selective breeding of *C. auratus* started in 2017 at the Institute of Plant and Food Research in Nelson (New Zealand) with the aim to improve the growth rate and other traits of economic interest (Ashton, Hilario, Jaksons, Ritchie, & Wellenreuther, 2019). Prior to the selective breeding research programme, only a handful of genetic markers were available. Our group has been developing diverse resources for snapper to facilitate rapid breeding progress in this species, including a linkage map, a transcriptome and genome-wide sequence information of pedigreed snapper (Ashton et al., 2019; Ashton, Ritchie, & Wellenreuther, 2018; Wellenreuther, Luyer, Cook, Ritchie, & Bernatchez, 2018). The development of genome-wide data sets for this species is important to explore the genetics of growth and sex determination, something that is not well understood in the family Sparidae, as well as to describe the genotype–phenotype map for additional phenotypes and to aid broodstock selection. As part of this endeavour, we here characterize the genome-wide catalogue of SVs and SNPs in wild-caught males and females of *C. auratus*. Specifically, we investigate three objectives: first, characterize the prevalence and locations of the genomic variation in 12 genomes of wild-caught snapper. This enabled us to gain a deeper understanding of questions such as: how much of the variation is shared, and what is the size and location of SVs? Second, we inferred the potential functional impact of this variation using gene annotation information to identify how many of the SVs intersected with genes. Third, we investigated whether some of the SVs were under selection. These analyses of the genomic variation will enable us to assess the impact of the genomic differentiation on the functional portions of the genome.

2 | MATERIALS AND METHODS

2.1 | Genome assembly

A reference genome for *C. auratus* based on a male individual and assembled to the pseudo-chromosome level (hereafter referred to as chromosomes) was used for all sequence alignments. Genomic DNA was prepared by BGI Tech Solutions (BGI), Shenzhen, China, from white muscle tissue samples collected from a single male fish (collected on the 25/07/2014 at the Plant and Food Research Seafood Research Facility in Nelson, New Zealand) and used to construct five Illumina libraries with average insert sizes of 170 bp, 3 kb, 8 kb, 10 kb and 20 kb. Libraries were sequenced to yield read pairs of 125 bases in length. Postsequencing raw paired end reads from the 170 base insert library were filtered by BGI using a proprietary pipeline to remove adaptor sequences, contamination and low-quality reads generating a total of 420,559,488 cleaned paired end read pairs of 125 bases in length from the BGI sequencing. Following quality score analysis using FASTQC v0.11.2 (Andrews, 2010), residual adapters not removed by the BGI pipeline were filtered using fastq-mcf EA-UTILS v1.1.2-537 (Aronesty, 2013) using command line options: "-l 50 -q 20 -t 0.00001 -C 3000000" yielding 419,940,157 read pairs. For mated paired end libraries, the reads were trimmed to 36 bases in length using an in-house PERL script (C. Deng, unpublished) and redundant read pairs removed using in-house PERL script (R. Crowhurst, unpublished) yielding ~84.7, ~66.6, ~20.4 and ~10.8 million unique read pairs for the 3 kb, 8 kb, 10 kb and 20 kb insert libraries, respectively.

The filtered and processed reads from all five libraries were assembled using ALLPATHS-LG v50191 (Gnerre et al., 2011) yielding 5,998 scaffolds containing 739.7 Mb, with a longest scaffold length of 7.53 Mb, N50 of 1.427 Mb, N90 of 182 kb and 10.66% N content. These scaffolds were composed of 65,560 contigs containing 660.9 Mb with longest contig length of 61,709 bp, N50 of 30.3 kb and N90 of 3,485 bp. Scaffolds were subject to three iterations of gap closure with GAPCLOSER v1.12 (Luo et al., 2012) using the input paired end reads. Gap closure yielded a reduction in N% content to 4.29%.

To further improve the gap filled ALLPATHS-lg assembly, BioNano genome mapping was performed by the BioNano Genome Mapping Service at Kansas State University. This process yielded a super scaffold assembly composed of 5,634 scaffolds and containing 772.3 Mb with a longest scaffold of 19.53 Mb, N50 of 4.46 Mb, N90 of 220 kb and 8.5% N.

The assembly units resulting from BioNano genome mapping were assigned to linkage groups using a GBS map (Ashton, Ritchie, & Wellenreuther, 2018) as follows. A 101 base genome region centred on each SNP marker was extracted from the ALLPATHS-lg assembly scaffolds using an in-house PERL script (R. Crowhurst, unpublished) and aligned to the super scaffold assembly using BOWTIE2 v2.2.5 (Langmead & Salzberg, 2012) using the command line options: "--end-to-end --sensitive –all." Regions aligning to a single location were retained and used as input to the reference bases genome arrangement tool CHROMONOMER v1.07 (Catchen, 2016). An in-house PERL script (R. Crowhurst, unpublished) was then used to construct the linkage group sequences from the AGP file produced by Chromonomer and the super scaffolds assembly.

Assembly correctness was visualized using "hagfish blockplots" (https://github.com/mfiers/hagfish) as follows: unique read pairs from the 20 kb long insert library were aligned to individual chromosome sequences using BOWTIE v1.0.0 (Langmead, Trapnell, Pop, & Salzberg, 2009) and command line options "-k 5 --best --strata -I 1 -X 100000" (https://github.com/mfiers/hagfish/wiki/ReadMappers). For an overview, alignments to individual chromosomes were visualized using "hagfish blockplots" using the command line options "-s 1 -e $LG_LEN -n $LG_LEN -f svg --dpi=220 -W 1600" where LG_LEN is the length (in bases) of each individual chromosome and individual plots were then manually merged into a single overview figure (Supporting Information Figure S1A). Additionally, "hagfish blockplots" for individual chromosomes were visualized at higher resolution using hagfish blockplots using the command line options "-s 1 -e $LG_LEN -n 2000000 -f svg --dpi=220 -W 1600" where LG_LEN is the length (in bases) of each individual chromosomes. Representative individual hagfish plots are presented in Supporting Information Figure S1B (LG2, LG5 and LG11).

Assembly completeness was estimated using BUSCO v3 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) with the settings "--m genome --p zebrafish" to search against the actinopterygii database (http://busco.ezlab.org/v2/datasets/actinopterygii_odb9.tar.gz). Repeat masking was completed using REPEATMASKER-OPEN v4.0.5 (Smit, Hubley, & Green, 2013) and REPEATMODELER-OPEN v1.0.11 (Smit, Hubley, & Green, 2008). First, a database was built based on the LG and un-anchored contig sequences using the command "BuildDatabase." A de novo repeat library was then constructed and annotated using RepeatModeler based on the database. The *C. auratus* genome was initially masked with the newly built library, followed by a second round of masking with "-species Opisthokonta," the common ancestor of fungi and animals.

2.2 | Sample information and DNA extraction

Fin tissue samples were collected from 12 wild-caught *C. auratus* from Tasman Bay (41.0458°S, 173.2934°E), Manukau Harbour (36.5054°S, 174.763336°E) and Hauraki Gulf (36.4263°S, 175.1894°E) in New Zealand. These locations are separated by several 100 km each; however, little is currently known about the extent of population differentiation or gene flow in this species, meaning that the spatial effect on genomic differentiation is yet to be established. At each location, two males and two females were sampled. DNA was isolated from the fin tissue samples using a modified salt extraction method (Aljanabi & Martinez, 1997). Quantification of DNA was carried out using Hoescht 33258 fluorescent dye on a BMG LABTECH CLARIOstar plate reader. Fragmentation of the extracted DNA was checked by gel electrophoresis using a Thermo Fisher 1 kbp DNA extension ladder to ensure that only samples with high molecular weight genomic DNA were sequenced.

2.3 | Genome resequencing and data processing

A short insert library (~250 bp insert size) was constructed for each sample following the standard Illumina protocol (Supporting Information Table S1). The 12 libraries were barcoded, pooled and sequenced across two lanes on Illumina HiSeq4000 at BGI to a targeted average read-depth of 30×. Sequencing data quality was checked using FASTQC v0.11.7 (Andrews, 2010) and MULTIQC v1.5 (Ewels, Magnusson, Lundin, & Käller, 2016). Read data were subsequently cleaned through adaptor removal, end trimming (9 bp from 5′ and 10 bp from 3′) and the elimination of low quality reads, using TRIMMOMATIC v0.36 (Bolger, Lohse, & Usadel, 2014). Reads were filtered for a minimum length of 50 bp and only paired reads were retained for further analysis (Supporting Information Table S1). The trimmed fastq files were converted into uBAM files using Picard-Tools v2.10.1 "FASTQTOSAM" (Broad Institute, 2015), while simultaneously adding read group, sample and library IDs. Finally, any remaining adaptor sequences were marked using Picard-Tools "MARKILLUMINAADAPTERS" (Broad Institute, 2015).

The reference genome was indexed using BWA v0.7.15 (Li & Durbin, 2009). Alignment of the uBAM files was achieved by converting them to interleaved Fastq files using Picard-Tools "SAMTOFASTQ" and then by piping the output to bwa-MEM using the switches -M (mark shorter split hits as secondary) and -p for interleaved Fastq files. The aligned bam files were subsequently merged with unaligned bam files using Picard-Tools "MERGEBAMALIGNMENT" to incorporate the metadata from the unaligned bam files. Duplicate reads were removed using Picard-Tools "MarkDuplicates." GATK v3.8.0 (McKenna et al., 2010) "IndelRealigner" was used to realign reads around indels.

2.4 | Short-variant calling (SNPs and indels)

Short variants were called using GATK v3.8.0. Specifically, GATK "HaplotypeCaller" was used in GVCF mode with the following settings: "useNewAFCalculator," "dontUseSoftClippedBases," "mbq 10" and "-mmq 0." Output g.vcf files were then jointly genotyped using GATK "GenotypeGVCFs" to generate a single vcf file of variants of all 12 samples. To estimate the extent of polymorphism composed of short variants, a variant set was generated by filtering output from GATK with "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum <−12.5 || ReadPosRankSum <−8.0" for SNPs and "QD < 2.0 || FS > 200.0 || ReadPosRankSum <−20.0" for indels. To avoid variants being counted twice, 118 indels that were called by both GATK and LUMPY v0.2.13 (Layer, Chiang, Quinlan, & Hall, 2014) were eliminated from the GATK call set using bedtools intersect of the BEDTOOLS SUIT v2.27.1 (Quinlan & Hall, 2010) with a required overlap of 80% or more. The resulting vcf files (SNPs and indels) were filtered for variants contained in linkage groups 1 to 24 (LG1-LG24). Chromosomal SNP frequencies were estimated from counts of filtered SNP sites at each linkage group. To estimate the extent of indels within the genome, the mean indel length was calculated from a histogram of indel length (vcftools -hist_indel_length v0.1.14) and multiplied by the chromosomal and genomic indel frequencies.

2.5 | Large variant calling and genotyping

Large structural variants (duplications, deletions and inversions) were called for each sample in parallel using alignment (bam) files produced above as input and LUMPY Express in LUMPY v0.2.13 (Layer et al., 2014) as the caller, following the workflow in the GitHub repository (https://github.com/arq5x/lumpy-sv, as at 17/5/2018). The resulting vcf files were then sorted and merged into a single vcf file using the l_sort.py and l_merge.py scripts. The l_merge script was run using the product option and 20 bp of slop (the slop category setting allows for some uncertainty by extending the confidence interval in both directions by a specified bp amount). The combined call file was then used to genotype the samples together using SVTyper (Chiang et al., 2015) from the original bam files. The resulting vcf files were filtered for variants contained in linkage groups 1–24 (LG1-LG24) and sorted using VCFTOOLS v0.1.14 (Danecek et al., 2011) and merged using vcf-merge of the vcftools package. For downstream analysis, we included duplications, deletions and inversions greater than or equal to 50 bp less than or equal to 50 kb, and with at least 6 supporting reads (SU).

Further filtering steps were taken to avoid false variants due to potential assembly anomalies. Variants that were genotyped as homozygous in all 12 samples or heterozygous in all 12 samples were likely due to sequence anomalies and were removed from the analysis. There was also a subset of variants that were called by LUMPY but were either not genotyped, or were genotyped as homozygous reference in all samples, and these were also removed from the analysis. Finally, any variants that spanned scaffold-to-scaffold junctions were removed to prevent false variants caused by scaffold misordering or misorientation. However, we kept rare variants, the rarest being those present as a single haplocopy in one individual. We did this because variants are expected to be present at low frequencies in the population and the visual inspection of a subset of rare variants showed that these were genuine and we thus think that discarding these variants would result in a skewed representation of the structural variant status in the species. Allele frequency spectra of variants of the three variant classes are given in Supporting Information Figure S2.

To estimate the extent that variants affected the genome, for both total variants and variants within each sample, variants of each class were merged using bedtools merge of BEDTOOLS v2.27.1 to account for those that overlapped with one another and the resulting merged variants were summed.

2.6 | Identification of genes that overlap with structural variants

The *C. auratus* transcriptome assembly (Wellenreuther et al., 2018) was accomplished using TRINITYRNASEQ v2.2.0 (Grabherr et al., 2011) and annotated with Trinotate (Bryant et al., 2017). The sequenced tissue types included replicated white muscle and brain samples, as well as whole larvae, preserved in RNAlater and extracted using the Trizol LS Reagent (Life Technologies) according to the manufacturer's instructions. RNA samples were individually prepared (mRNA was isolated from total RNA via poly(A) pull-down) for sequencing using the Illumina Tru-Seq kit on an Illumina HiSeq 2000 sequencer (paired end 100 bp sequencing, 160 bp insert length) at the Beijing Genomics Institute Shenzhen, China. The transcripts were mapped to the *C. auratus* genome using GMAP (Wu & Watanabe, 2005) to create gene models in gff3 format. Redundant gene entries (multiple transcripts mapped to the same locus) were excluded when calculating gene frequencies and overlaps between genes and SVs using bedtools merge of BEDTOOLS v2.27.1.

To determine the impact of structural variants on genes, bedtools intersect of BEDTOOLS v2.27.1 was used to determine genes that overlapped with variants. Settings were adjusted to determine genes that were entirely affected by a variant (−f 1), by 20% (−f 0.2) or affected by any intersection (default).

2.7 | Estimation of Tajima's *D* and nucleotide diversity

Duplicated regions that are under-represented in the reference genome are likely to result in high read-depths of reads from more than one copy of the duplication. These duplicated regions can introduce spurious variants that affect estimated nucleotide diversity statistics. Steps were therefore taken to eliminate duplicated regions from the analysis. Tajima's *D* and π were calculated for 20 kb windows using the following procedure. First, a list of accessible sites and regions was generated by eliminating sites that had one of the following: N-sequence, repeat sequences (based on the repeat masking), mean read-depths across all 12 samples >70 or <10 (not including reads with mapping qualities of <30); and regions spanned by duplications or deletions. Tajima's *D* was estimated for 20 kb windows using VCFtools and filtered only to include 20 kb windows that had >50% of sites accessible. For π, calculations were based on VCFtools site-pi. Output was filtered to include only accessible sites which were used to calculate mean (windowed) π of accessible sites within 20 kb windows. Tajima's *D* and π were plotted, along with the upper and lower 1% quantiles as outliers, using the R package Circlize (Gu, Gu, Eils, Schlesner, & Brors, 2014).

Clusters consisting of at least three outlier windows (i.e., windows of upper or lower 1% quantiles) with no more than 20 kb between each of them were assessed for the presence of transcribed genes and variants using bedtools intersect of BEDTOOLS v2.27.1 (Quinlan & Hall, 2010). A Welch two sample *t* test was used to test for the differences in means of feature (gene and variant) counts between the outlying 1% quantiles windows and all other 20 kb windows across the genome. Specifically, the *t* tests were performed by comparing windows of the outlier quantiles with all 20 kb windows used for calculation of π and Tajima's *D* across the genome. Variants that intersected with regions by at least half were counted. Finally, a Grubb's test was conducted to estimate if any LGs harbour more SVs than predicted based on their size.

2.8 | General analysis

Analysis was done using bash, Perl and R (R Core Team, 2008) in Jupyter Notebook (Kluyver et al., 2016). Circos plots were generated from bed files containing genomic coordinates and sizes of variants, and genomic coordinates and values of π and Tajima's *D* output using the R package CIRCLIZE v0.4.1 (Gu et al., 2014). Data manipulation for analysis was undertaken using Perl and the R package DPLYR v0.7.4 (Wickham & Francois, 2017). Data have been deposited on the New Zealand repository hosted by the national infrastructure platform Genomics Aotearoa (https://www.genomics-aotearoa.org.nz/data/).

3 | RESULTS

3.1 | Genome assembly statistics

A three-level genome assembly approach was taken to assemble the Snapper genome to chromosome level. The first level of assembly (ALLPATHs-lg based) yielded 5,998 scaffolds containing 739.7 Mb of sequence with an N50 of 1.427 Mb. The second level of assembly employed BioNano genome mapping and yielded a small reduction in scaffold number but an overall increase in assembly metrics based on N50 (4.46 Mb) and an increase in longest assembly unit length from 7.53 Mb to 19.53 Mb. Of the 5,634 super scaffolds from level 2, 1,049 were able to be assigned to 24 linkage groups ranging in size from 17.2 to 38.6 Mb. A further 37 scaffolds (3.82 Mb) were assigned to a 25th orphan linkage group. The remaining 4,548 scaffolds (33.88 Mb) with longest scaffold of 748 kb and N50 of 17.5 kb could not be assigned by this process. Evaluation of incorporation of paired end input reads into the final assembly was assessed by mapping these reads back to the final assembly using Bowtie2 (Langmead & Salzberg, 2012) using "–end-to-end" alignment which yielded an overall alignment rate of 92.27%.

Assembly correctness was assessed using alignment distances for read pairs from the 20 kb long insert library and graphically visualized using "hagfish blockplots" (https://github.com/mfiers/hagfish). These plots indicate regions where mates align within expected distances (green), at greater than expected distances (red) and at less than the expected distance (blue) (note that black indicates missing data). Visualization (Supporting Information Figure S1A,B) of alignments to each chromosome assembly suggests the majority of the assembly is assembled within expected distances for mate pairs from the 20 kb library, but that some regions not optimally assembled (potential miss-assembly points) were occasionally present on chromosomes (e.g., see LG2, LG5, LG7, LG4, LG14, LG15, LG17, LG24), suggesting that the availability of new data might lead to further improvement.

Evaluation of genome completeness estimated using BUSCO v3 (Simão et al., 2015) indicated the genome assembly was reasonably complete. Of 4,584 BUSCOs, 4,435 (96.8%) were considered complete, of which 4,327 (94.4%) were single copy BUSCOs and 108 (2.4%) were duplicated. Fifty-three (1.2%) BUSCOs were considered fragmented while 96 (2.0%) were missing from the genome assembly.

3.2 | Number of variants across the genome and impact on the total genomic sequence variation

We generated a filtered call set of short variants (SNPs and indels) across the 24 *C. auratus* linkage groups using GATK HaplotypeCaller and GenotypeGVCFs. This set consisted of a total of 6,547,716 SNPs, and 1,301,743 indels ranging from deletions as large as 269 bp to insertions as large as 338 bp. For details, see Supporting Information Table S2 for a breakdown of SNPs and SVs for individual genomes and Supporting Information Table S3 for a list of variants detected with both LUMPY and GATK. The size distribution was heavily skewed towards small variants (<=50 bp), for both insertions and deletions. Based on a mean indel length of 4.8 bp, indels affect approximately 6,248,366 bp of the genome (Supporting Information Tables S4 and S5). In terms of the genome-wide percentages of sequence affected by SNPs and indels, we estimated that SNPs affect 0.93% of the genomic sequence variation and similarly, that indels affect 0.89% of the genome (Figure 1a,b and Supporting Information Table S2).

A call set of larger structural variants using LUMPY and genotyped using SVTyper, filtered for duplications, deletions and inversions between 50 bp and 50 kb, was also generated. Supporting Information Table S5 shows counts of variants following each filtering step. The final call set of a total of 20,385 SVs included 2,427 duplications, 17,599 deletions and 359 inversions affecting 1.09%, 0.66% and 0.04% of the genome content, respectively (Table 1 and Supporting Information Figure S2). Therefore, while deletions outnumbered duplications by 7.25X, duplications were generally larger (minimum = 82, maximum = 49,834 bp, mean = 4,830 bp, median = 1510 bp) than deletions (minimum = 51 bp, maximum = 29,161 bp, mean = 269 bp, median = 104 bp) and inversions (minimum = 51 bp, maximum = 24,993 bp, mean = 959 bp, median = 73 bp). Overall, while genome-wide sequence variation caused by SNPs affected 0.93% of the genome, genomic variation caused by indels and the large structural variants (SV) was together 2.9X higher.

Analysis of the frequency spectra of duplications, deletions and inversions shows that most variants are rare, in line with theoretical expectations, but also shows that a large amount of variants are shared by a number of individuals, and are thus polymorphic in the population (Figure 2). The bubble plots also show that some variants are more common in males compared to females (Figure 2); however, the vast majority of variants showed no strong sex-specific pattern.

3.3 | Genome-wide estimates of Tajima's *D* and π, and overlap with genes

Tajima's *D* and nucleotide diversity (π) scores of 20 kb window of the 24 LGs revealed a number of outlier windows that indicate genomic loci under putative selection (Figure 3 and Supporting Information Figure S3). We also calculated the numbers of genes and SVs to be present in the upper and lower 1% quantiles of Tajima's *D* and π (Table 2). An excess of genes was found within the upper 1% quantiles of Tajima's *D* ($p = 0.00518$) and a deficit of genes was found in the lower 1% quantile of Tajima's *D* ($p = 0.00003116$). A large

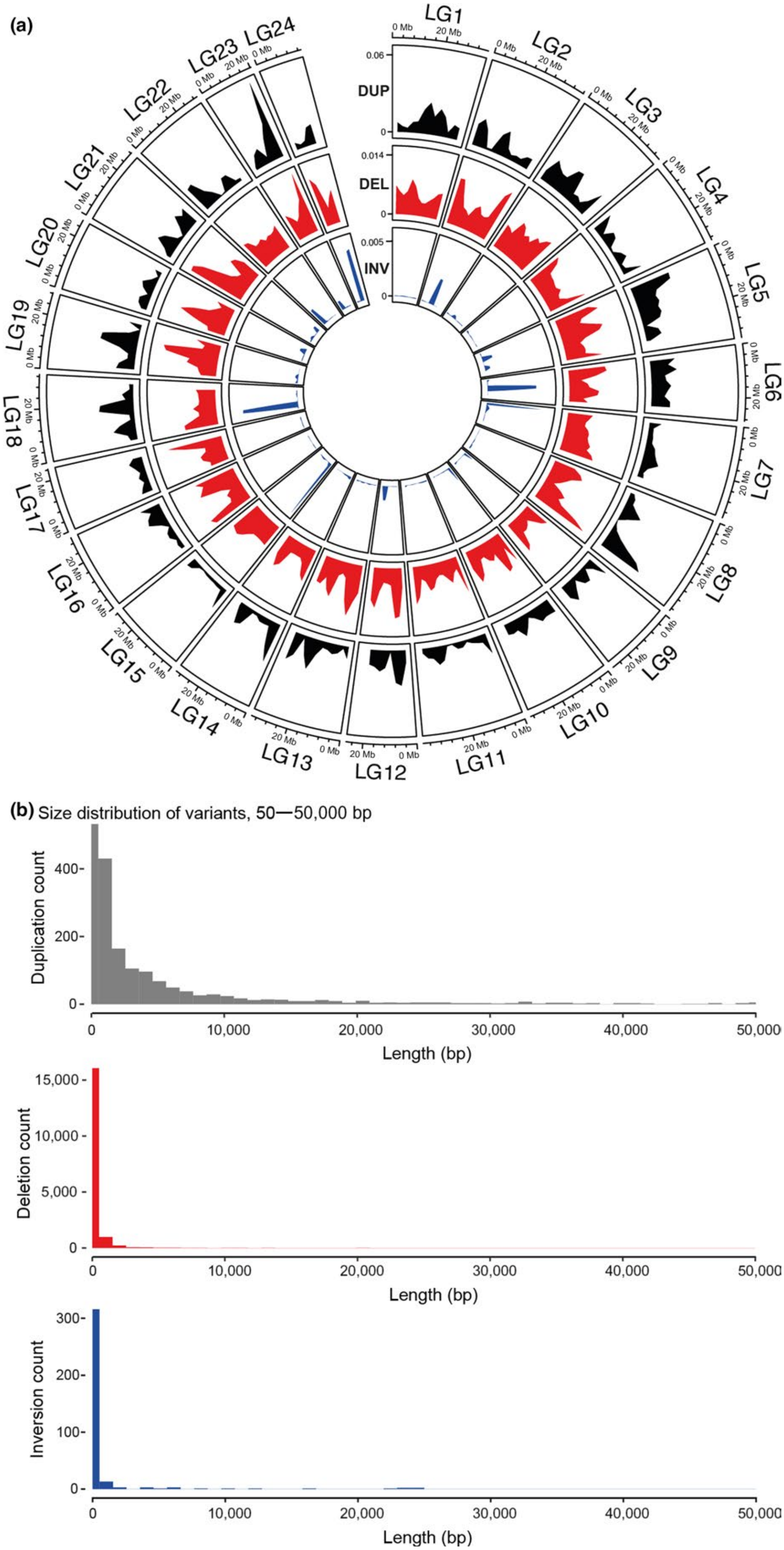

FIGURE 1 Panel (a) shows a Circos plot showing the genomic densities (as a fraction of 5 Mb windows) of different variant classes in *Chrysophrys auratus* investigated in the analysis. From the outside to the inside, the tracks are densities of duplications in black, deletions in red and inversions in blue. Panel (b) shows the size distributions of duplications, deletions and inversions [Colour figure can be viewed at wileyonlinelibrary.com]

excess of genes was found in the upper 1% quantile of π windows (p = 2.2E-16), but no significant excess nor deficit was seen in windows of the lower 1% quantile. With respect to variants, most notable are significant excesses of deletions in upper 1% quantiles of both Tajima's D (p = 0.003267) and π (p = 2.2E-16) and significant deficits within the lower 1% quantiles of both Tajima's D (p = 2.2E-16) and π (p = 2.2E-16).

TABLE 1 List of counts of detected SNPs and indels, as well as deletions, duplications and inversions >50 bp and <50 kb

Type of variant	Number of variants	Base pairs affected	% of genome
SNPs	6,547,716	6,547,716	0.93
Indels	1,301,743	6,248,366	0.89
Deletions	17,599	4,666,669	0.66
Duplications	2,427	7,699,239	1.09
Inversions	359	249,945	0.04
Total SVs	1,330,589	18,864,219	2.68

Note. The column denoted "base pairs affected" indicates the number of sequence bases affected by a particular type of genomic variant, whereas the column denoted "% of genome" indicates the percentage of all 24 linkage groups of the whole genome (determined bases) affected by each variant class. The per cent of the genome of indels was calculated based on a mean indel length of 4.8 bp. The "Total SVs" is based on indels in addition to deletions, duplications and insertions. For a breakdown of variants by linkage group, see Supporting Information, Table S3.

Particular attention was paid to clusters of outlier windows (a cluster being defined as at least three outlier 20 kb windows within 100 kb), of which 19 high and 16 low π clusters, and four high and six low Tajima's D clusters were found. Clusters also demonstrated higher and lower counts of SVs: the 1.44 Mb (0.20% of the genome) of high π 20 kb window clusters carry a total of 117 SVs (0.57%) at a density of 0.81 variants per 10 kb, while the 1.020 Mb (0.145% of the genome) of low π 20 kb window clusters carry only 11 SVs (0.037%), at a density of 0.11 variants per 10 kb. This indicates that high π clustered regions (those of the upper 1% quantile) carry more than 7.36 times the number of variants than low π clustered regions (those of the lower 1% quantile), and suggests that regions that harbour high SNP variation also harbour high SV variation.

Interestingly, regions with the highest gene densities were seen in the high π clusters, with two clusters of LG23 containing at least one gene/10 kb (Figure 3 and Supporting Information Table S6). LG23 has four clusters of high π within a 1.34 Mb region of the distal end of the chromosome. In addition to high nucleotide diversity, this 1.34 Mb region shows high densities of all SVs, with 56 deletions spanning 9,770 bp (0.7%), 33 duplications spanning 233,202 bp (17.4%) and 7 inversions spanning 47,583 bp (3.6%).

The Grubb test showed that LG23 is an outlier in terms of duplications, indels and the overall number of SVs (Supporting Information Table S7). Further, LG17 and LG20 are the furthest from all other LGs in deletions and inversions, respectively; however, they are not considered as significant outliers based on their p value > 0.05 (Supporting Information Table S7).

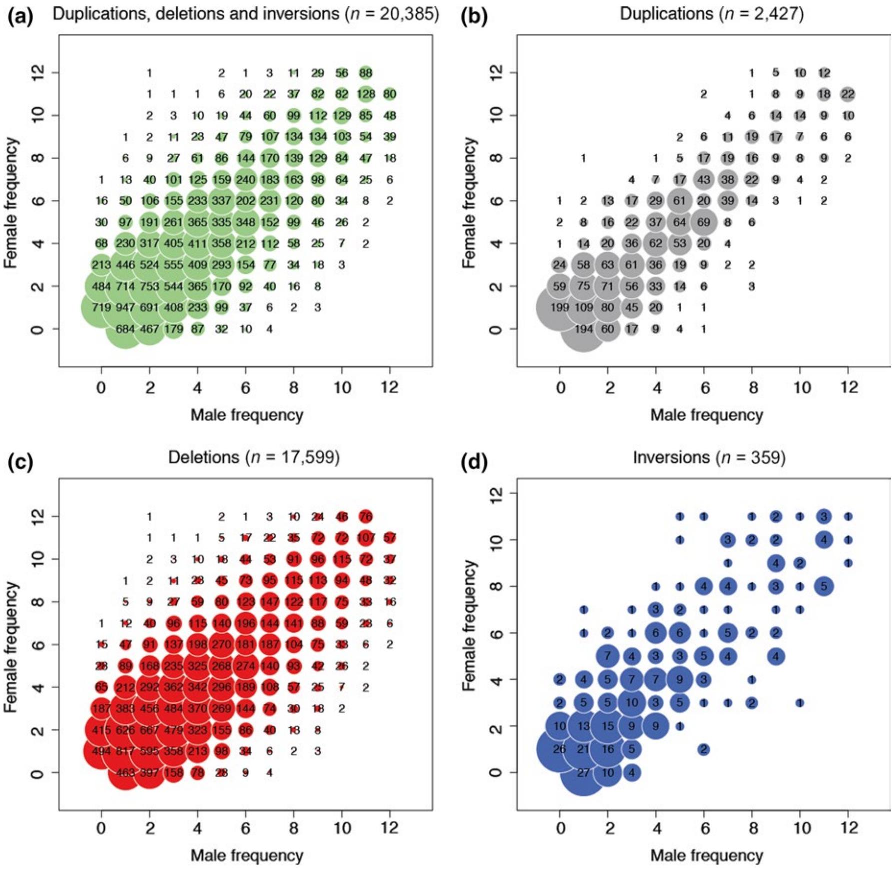

FIGURE 2 Bubble plots showing the frequency distribution of different variant classes between males and females. Each variant was counted for its presence within males and females. The total number of variants is equal to the sum of counts of variants within each state. Counts are composed of either observation of the variant whether heterozygous or homozygous. Panel (a) shows the frequency or all SVs, panel (b) for duplications, panel (c) for deletions and panel (d) for inversions [Colour figure can be viewed at wileyonlinelibrary.com]

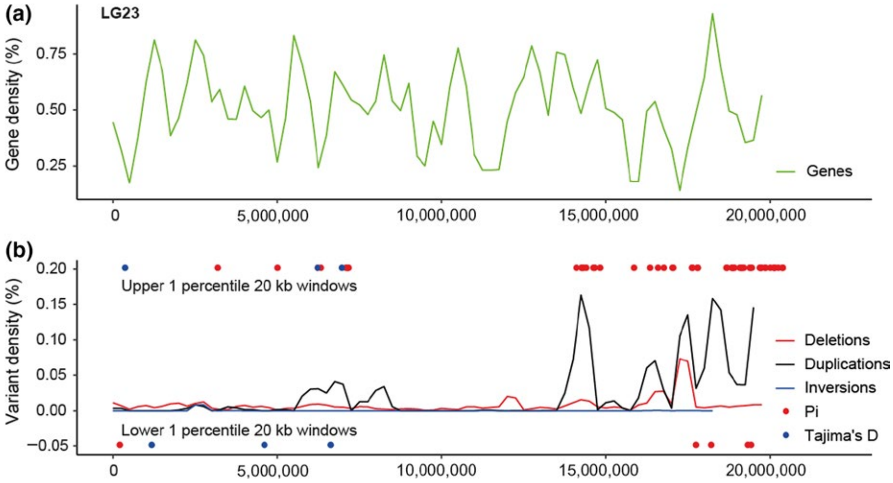

FIGURE 3 Gene density and variant density of LG23 estimated in 500 kb windows. Variant densities were plotted along with the locations of outlying 20 kb windows of π and Tajima's *D* [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Counts of genes and variants along with means, standard deviations (*SD*) and *p* values (generated from Welch two sample *t* tests), within 20 kb outlier windows of (A) Tajima's *D* and (B) π

	Tajima's *D*										
	Upper 1 percentile (*N* = 320)				Total (*N* = 32,007)			Lower 1 percentile (*N* = 320)			
Feature type	Total	Mean	*SD*	*p*-Value	Total	Mean	*SD*	Total	Mean	*SD*	*p*-Value
(A)											
Genes	241	0.7531	1.0132	0.00518**	18,983	0.5931	0.8822	133	0.4156	0.7464	3.12E-05***
Duplications	37	0.1156	0.4065	0.01488*	1,913	0.0598	0.3574	14	0.0438	0.2589	0.2738
Deletions	208	0.6500	0.8506	0.003267**	16,276	0.5085	0.7682	69	0.2156	0.4688	2.2E-16***
Inversions	3	0.0094	0.1248	0.915	324	0.0101	0.1047	2	0.0063	0.0789	0.3849
Total SVs	248				18,513			85			
	π										
	Upper 1 percentile (*N* = 352)				Total (*N* = 35,165)			Lower 1 percentile (*N* = 352)			
Feature type	Total	Mean	*SD*	*p*-Value	Total	Mean	*SD*	Total	Mean	*SD*	*p*-Value
(B)											
Genes	419	1.1903	1.0785	2.2E-16***	20,353	0.5788	0.8753	178	0.5057	0.8372	0.1041
Duplications	48	0.1364	0.5160	0.009132*	2,254	0.0641	0.3520	71	0.2017	0.7014	0.000274***
Deletions	466	1.3239	1.1336	2.2E-16***	17,549	0.4990	0.7635	61	0.1733	0.4543	2.2E-16***
Inversions	8	0.0227	0.1492	0.1174	359	0.0102	0.1060	2	0.0057	0.0753	0.2646
Total SVs	522				20,162			134			

Note. Outlier windows were taken as the upper and lower 1% quantiles. Numbers in brackets refer to the number of windows of each outlier quantile, or the total number of windows used for the analysis.
The symbol * after a *P* - value denotes the significance threshold that was applied, with one asterisk denoting a threshold of 0.01-0.05, two asterisk 0.001-0.01and three asterisk < 0.001.

3.4 | Genome-wide estimates of genes affected by with structural variants

We counted annotated genes that were fully encapsulated by SVs (i.e., genes that intersected with SVs by 100% as a percentage of the gene), or that partially intersected with SVs, by at least 20% or by any intersection of at least 1 bp (Figure 4a,b). A total of 294 genes were found to be fully encapsulated by variants. Most of these (208) were encapsulated by duplications, with fewer encapsulated by deletions (76) or inversions (10). Of genes encapsulated by duplications, nine of these were also encapsulated by deletions and three were also encapsulated by inversions. Of genes that intersected with SVs by at least 20%, there were again more intersections with duplications (447) than deletions (142) and inversions (18). However, when any form of intersection was taken into account, 965 genes were affected by duplications, 4,918 genes by deletions and 155 by inversions. Overall,

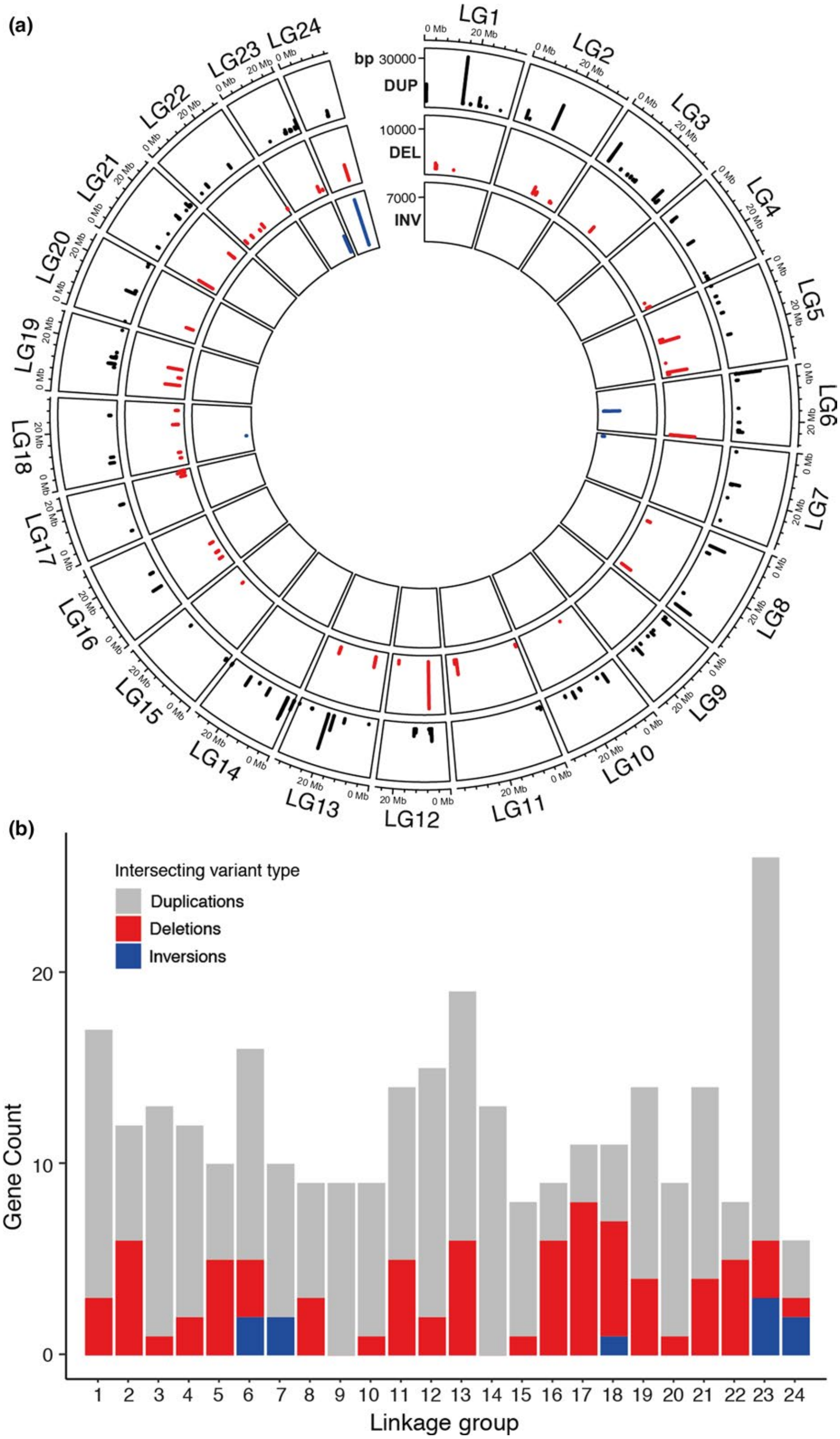

FIGURE 4 Panel (a) shows a Circos plot of the locations and relative sizes of genes affected by structural variants of the *Chrysophrys auratus* genome. Genes affected by duplications, deletions and inversions are represented in black, red and blue respectively. Only genes completely overlapped by a variant are shown. Panel (b) Counts of genes completely overlapped by the three structural variant classes on each linkage group (LG) [Colour figure can be viewed at wileyonlinelibrary.com]

6,038 genes of the 22,424 annotated genes assessed (27%) were affected in some way by a structural variant of at least 50 bp (Table 3).

Of genes that were entirely encapsulated by variants, putative functions based on sequence similarity were able to be ascribed to 118 of them, with 91 in duplications, 26 in deletions and 1 in an inversion (Supporting Information Table S8). Interestingly, only one of the flanking regions of the 91 genes bounded by duplications showed similarity to sequences of transposable elements, while 21 of the 26 genes bounded by deletions showed flanking regions with similarity to sequences of transposable elements.

4 | DISCUSSION

Structural variants were the first molecular markers used in the field of evolution, and classic examples include inversion polymorphism

TABLE 3 Counts of genes affected by structural variants (duplications, DUPs; deletions, DELs; and inversions, INVs, >50 bp and <50 kb, based on a genome size of 705 Mb)

	Amount of intersection					
	100%		20%		Any intersection	
	N	%	*N*	%	*N*	%
Duplications	208	0.93%	447	1.99%	965	4.30%
Deletions	76	0.34%	142	0.63%	4,918	21.93%
Inversions	10	0.04%	18	0.08%	155	0.69%
Total	294	1.31%	607	2.71%	6,038	26.93%

Note. An affected gene is one that is totally encapsulated (100% intersection), has 20% of its length intersecting a variants (20% intersection) or has any part of the gene intersecting a variant (any intersection)

work on *Drosophila* spp. by Sturtevant in the 1920s (Sturtevant, 1921), and work on inversions and transposable elements in *Zea mays* by McClintock shortly after (McClintock, 1931, 1950). In the decades to follow SVs became increasingly abandoned in favour for SNPs during the wave of sequencing technology improvements that have continued until today. It is thus no surprise that SNPs have for several decades constituted the main currency for investigations into the genotype–phenotype map and into evolutionary processes involved in the diversification and adaptation of species. Consequently, our understanding of SNPs is comprehensive for the euchromatic portion of the genome. In contrast, SVs have fared far worse partly because of their stronger association with repetitive DNA (Sudmant et al., 2015). Recently, however, a growing number of studies indicate that varied types of SV can be common (Feulner et al., 2013; Weischenfeldt, Symmons, Spitz, & Korbel, 2013), and even be key drivers shaping major processes in evolution (e.g., inversions Wellenreuther & Bernatchez, 2018). Here, we describe a structural variant map of 20,385 SVs that cover 18.9 Mb of the assembled genome. Together with short indels, the catalogued SVs affect 2.69% of the genome and are affecting three times (2.9×) more of the genome-wide sequence information compared to SNPs (0.93%). Large levels of genome-wide SV variation have also been recorded in a study on the three-spine stickleback (Feulner et al., 2013) and on African cichlids (Brawand et al., 2014), but direct comparisons are difficult owing to the different data sets, methods and structural variant types being assessed. Our study further highlighted that many of the genetic variants constitute potential candidates that may modulate phenotypic variation. Evidence for this comes from the finding that many SVs showed extensive overlap with genes (27% of all genes were affected by SVs), and the finding that several SVs are found in regions of elevated Tajima *D* and/or π, and as such may be able to provide insights into the location of functionally important polymorphisms.

4.1 | Increasing evidence that structural variation is ubiquitous and involved in adaptive diversification

Recent large-scale population-based analyses coupled with advances in sequencing technologies have demonstrated that the genome of model species is significantly more diverse than originally thought. For example, whole-genome inspections have provided detailed information on SVs in model species like the thale cress *Arabidopsis thaliana* (DeBolt, 2010), the common fruitfly *Drosophila melanogaster* (Chakraborty et al., 2018; Huang et al., 2014), zebrafish *Danio rerio* (Brown et al., 2012), mice *Mus musculus* (Keane et al., 2014; Quinlan et al., 2010) and humans *Homo sapiens* (Feuk, Carson, & Scherer, 2006; Sudmant et al., 2015). In contrast, work to understand the abundance and position of SVs within nonmodel organisms, and ultimately their role in evolution, remains largely unstudied, but some progress has been made. Of particular interest are these where a link to adaptive processes could be made, and we will mention some representative examples here. Adaptive structural variation has been identified in the genomes of the well-studied green anole (Alföldi et al., 2011), three-spine stickleback (Jones et al., 2012), cichlids (Brawand et al., 2014; Fan & Meyer, 2014), *Heliconius* butterflies (Dasmahapatra et al., 2012; Pinharanda, Martin, Barker, Davey, & Jiggins, 2017) and Darwin's finches (Lamichhaney et al., 2015; Zhang et al., 2014) by either intra- or interspecific genome comparisons. For example, in both the *Anolis* species and the *Heliconius* butterflies comparative genomic analyses found an increase in mobile genetic elements such as transposable elements compared to closely related genomes, and this variation is thought to be related to diversification in gene expression (Alföldi et al., 2011; Brawand et al., 2014). In another study on the hybridizing *Heliconius* butterflies *H. melpomene* and *H. cydno*, no impact of inversions on genome introgression was found, however, indicating that differences in inversion are not necessary to maintain the species barriers in this pair (Davey et al., 2017). In sticklebacks and cichlids, structural variation is abundant and appears to be linked to adaptive phenotypes (Fan & Meyer, 2014; Jones et al., 2012). Specifically, in sticklebacks, adaptive divergence between marine and freshwater ecotypes has been shown to involve chromosomal inversions on chr11 spanning 5.7 Mb (Jones et al., 2012) and appear to be linked with the generation of marine- and freshwater-specific KCNH4 isoforms. This is because the repeats flanking the chr11 inversion contain alternative 3′ exons for the voltage-gated potassium channel gene *KCNH4* and since the *KCNH4* transcription itself is initiated within the inversion, the alternative inversion orientations could lead to different isoforms (Jones et al., 2012). Moreover, analysis of structural genomic variation in five cichlid species from the Great Lakes in East Africa showed that while the overall level was generally high and much of it is shared between closely related species, the rates of inversions and deletions at the terminal taxa were substantially higher than the rates at the ancestral lineages, indicating that these variants may be related to the rapid diversification of this lineage (Fan & Meyer, 2014). In addition, much of the structural variation was located in functionally important regions of genes (e.g., regulatory regions) further indicating that they are forming some of the genomic substrate of the adaptive radiation of this group. A recently emerged additional explanation for this pattern comes from a phylogenomics study of this cichlid radiation showing that ancestral hybridization could have led to the observed pattern of shared SVs (Irisarri et al., 2018), underscoring that hybridization can be a creative force and facilitate speciation bursts and the partitioning of genetic variation among linages. These studies on nonmodel species are collectively demonstrating that the structural genome architecture of species can play a key role in the diversification of species.

Our analyses showed that SVs are not only widespread across the entire snapper genome (Figure 1a), but also that the size of these variants can be large (Figure 1b), particularly for duplications. We need to be cautious, however, with any firm conclusions at this point as our genome sample size was relatively modest (12 resequenced genomes) making any frequency distribution estimates preliminary at this stage. Additionally, our analyses excluded large SVs above 50 kb and precluded us from detecting SVs that would span across different scaffolds; hence, there might have been larger SVs that we simply could not detect, making our total bp estimate affected by SVs conservative. Our data further showed that large interchromosomal differences were apparent among the 24 chromosomes. For example, chromosome 23 showed the highest number of duplications of all of the chromosomes, and the number of chromosome specific inversions and deletions was also elevated. The spatially patchy distribution of variants across the genome has been detected previously, for example the heterogametic sex chromosomes commonly show similar pronounced differences.

The genomic regions high in variants are typically referred to as “hotspots.” Often, these hotspots are flanked on either side by areas of low recombination, which traditionally were inferentially detected by examining familial pedigrees for regions of the genome that showed recombination more frequently than expected. Nowadays, recombination rates can be estimated with high power with whole-genome sequencing data of pedigreed individuals; an area that our group is actively exploring in snapper. Furthermore, while we were able to use the data at hand to create an extensive catalogue of SVs for this species, we also need to stress that it is inherently complex to detect certain SVs using short-reads next-generation sequencing data and that, in particular, the identification of large and complex structural variants is extremely challenging (Ye, Hall, & Ning, 2015). On balance, we decided to include in the analysis rare variants that were not eliminated by the filtering steps. Analysis of the sensitivity and specificity of LUMPY to detect variants showed that sensitivity, at lower allele frequencies, is more of an issue than specificity (Layer et al., 2014). In the present study, rare variants required all evidence for the calling of the variant to be derived from the single rare-variant instance. Allele frequency spectra show no excess of rare variants, in fact, the allele frequency of deletions showed conservative calling of rare variants. While various sequencing technologies and bioinformatic pipelines have now been developed to identify structural variations, there is still no informatics method or algorithm that is capable of identifying the full range structural DNA variation (Ye et al., 2015). Instead, users are currently advised to combine calling results obtained from multiple complementary tools in order to increase sensitivity and specificity, thus to verify variants it may be advisable to apply different sequencing technologies as well as calling methods. With recent advances in single molecule sequencing such as that afforded by long read platforms PacBio (Pendleton et al., 2015) and Oxford Nanopore's MinION (Ip et al., 2015), substantial impact and changes to our ability to dissect the underlying structural variants in genomes can be expected in the genomics community.

4.2 | Structural variation effects on genes and evidence for selection

We detected that SVs intersected with 27% of all annotated genes, and that a substantial portion of the SVs was also in the upper and lower genomic regions of Tajima's D and π. Caution needs to be placed on this finding, however, as the sample size in our study was only modest, and sample size has been shown to have an effect on the measures of both Tajima's D and π (Subramanian, 2016), and some Tajima's D regions may also present false positives because calculations of these can also be affected and confounded by regions of reduced recombination (Thornton, 2005). Nevertheless, our analyses do indicate that at least some of the SVs are under selection and have an impact on the phenotype and are under selection. We are aware, however, that these analyses can only provide the first step in understanding which of the SVs may be involved in phenotypic evolution. Future work on the genotype–phenotype map of this species, for example, may be able to shed further light on this. Nevertheless, our findings are in line with the growing general awareness in the field of evolutionary ecology and genetics that SVs provide significant variation for selection to act on (Chain & Feulner, 2014).

Of these, inversion polymorphisms are prime candidates for rapid evolutionary change because they protect inverted sequences from recombination, allowing favourable allelic combinations to be maintained in the face of gene flow (Wellenreuther & Bernatchez, 2018). Our study detected 359 inversions which together affected 249,945 bp (0.04% of the genome). Some of these intersected with genes (Figure 4b), but inversions are also known to exhibit more indirect effects on the genome, for example, in the form of position variegation effects which can cause expression differences between the inverted and noninverted genomic region. The varied effects of inversions on the phenotype have recently been revisited by several studies due to the enhanced ability of new genomic technologies to screen for inversion variants. Indeed, mounting evidence shows that loci involved in local adaptation (Anderson, Hoffmann, Mckechnie, Umina, & Weeks, 2005; Coluzzi, Sabatini, della Torre, Di Deco, & Petrarca, 2002; Kirkpatrick & Barton, 2006; Lowry & Willis, 2010) and pre- or postzygotic isolation are commonly found on inverted regions of the genome (Ayala, Guerrero, & Kirkpatrick, 2013; Noor, Grams, Bertucci, & Reiland, 2001; Rieseberg, 2001). Moreover, inversions can also affect the gene expression by influencing expression profiles of the genes proximal to inversion breakpoint regions or by modifying expression patterns genome-wide due to rearranging large regulatory domains (Chan et al., 2010; Said et al., 2018). In addition to inversions, repeated deletions of the enhancer of the Pitx1 gene changed the gene expression patterns and are responsible for the repeated and independent loss of the pelvic spines in the freshwater stickleback populations (Chan et al., 2010). It is thus no surprise that structural variants like inversions and deletions have been recurrently linked to spectacular phenotypes and have a pervasive role in eco-evolutionary processes, from mating systems, environmental adaptation, reproductive isolation to speciation (Wellenreuther & Bernatchez, 2018).

5 | CONCLUSIONS

While the most abundant genomic variations are typically in the form of SNPs, which has made them prime candidates to mark genome fragments related to diseases or certain traits (e.g., for GWAS), we have shown that SVs can outnumber the genomic changes that they induce in terms of base pairs affected. However, we also found that regions rich in SNPs are also commonly rich in SVs. The co-occurrence of the high SNP and SV density at some regions in the genome may be caused by variation in recombination rate along the genome, which could increase the occurrence of both types in a somewhat similar manner. Moreover, it is conceivable that background selection and associated hitchhiking effects may modulate the local diversity of both types of SNP and SV variants in a similar way, but further empirical and theoretical data would be needed to investigate this in detail. Our analyses provide a first glance at the genomic variation of the marine teleost nonmodel species snapper and while SNPs were the most frequent type of variant, structural genomic variation affected three times more base pair-changes than SNPs. This underscores the emerging view that SVs are important to consider when studying genetic diversity, sex differentiation as well as genome evolution. Future improvements in SV detection and analysis should allow researchers to even better evaluate the impact of SVs in the generation of new diversity and to study the role they have in the adaptive evolution of species.

ACKNOWLEDGEMENTS

We would like to thank Elena Hilario, David Ashton and Peter Ritchie for their work on the snapper reference genome as well as Eric Normandeau for his contribution to the transcriptome annotation. We would like to also thank the MBIE platform "Genomics Aotearoa" for guidance with the publication of the snapper genome. This project was funded through the MBIE research programme grant "Accelerated breeding for enhanced seafood production" (contract number #C11X1603) to MW. We would like to acknowledge Ben Te Aika, Michael Black, Peter Dearden and David Chagne and Aotearoa Genomics for help with data management and hosting.

ORCID

Maren Wellenreuther https://orcid.org/0000-0002-2764-8291

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*, 1061. https://doi.org/10.1038/nature09534

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Alföldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., … Lindblad-Toh, K. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, *477*, 587–591. https://doi.org/10.1038/nature10390

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

Anderson, A. R., Hoffmann, A. A., Mckechnie, S. W., Umina, P. A., & Weeks, A. R. (2005). The latitudinal cline in the In (3R) Payne inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Molecular Ecology*, *14*, 851–858.

Andrews, S. (2010). FASTQC: a quality control tool for high throughput sequence data.

Aronesty, E. (2013). Comparison of sequencing utility programs. *Open Bioinformatics Journal*, *7*(1), 1–8.

Ashton, D. T., Hilario, E., Jaksons, P., Ritchie, P. A., & Wellenreuther, M. (2019). Genetic diversity and heritability of economically important traits in the Australasian snapper (*Chrysophrys auratus*). *Aquaculture*, (in press).

Ashton, D. T., Ritchie, P. A., & Wellenreuther, M. (2018). High-density linkage map and QTLs for growth in snapper (*Chrysophrys auratus*). *Genes, Genomes, Genetics*, 376012.

Ayala, D., Guerrero, R. F., & Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution*, *67*, 946–958. https://doi.org/10.1111/j.1558-5646.2012.01836.x

Baker, M. (2012). Structural variation: The genome's hidden architecture. *Nature Methods*, *9*, 133–137. https://doi.org/10.1038/nmeth.1858

Bolger, A. M., Lohse, M., & Usadel, B. (2014). TRIMMOMATIC: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*, 375. https://doi.org/10.1038/nature13726

Broad Institute, (2015). Picard-tools. Retrieved from https://broadinstitute.github.io/picard/

Brown, K. H., Dobrinski, K. P., Lee, A. S., Gokcumen, O., Mills, R. E., Shi, X., … Lee, C. (2012). Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proceedings of the National Academy of Sciences*, *109*, 529–534. https://doi.org/10.1073/pnas.1112163109

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., … Whited, J. L. (2017). A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. *CellReports*, *18*, 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Catchen, J. (2016). CHROMONOMER. Retrieved from catchenlab.life.illinois.edu/chromonomer/

Chain, F. J. J., & Feulner, P. G. D. (2014). Ecological and evolutionary implications of genomic structural variations. *Frontiers in Genetics*, *5*, 326.

Chakraborty, M., VanKuren, N. W., Zhao, R., Zhang, X., Kalsow, S., & Emerson, J. J. (2018). Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nature Genetics*, *50*, 20. https://doi.org/10.1038/s41588-017-0010-y

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., … Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*, *327*, 302–305. https://doi.org/10.1126/science.1182213

Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., … Hall, I. M. (2015). SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nature Methods*, *12*, 966. https://doi.org/10.1038/nmeth.3505

Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M. A., & Petrarca, V. (2002). A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*, *298*, 1415–1418. https://doi.org/10.1126/science.1077769

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., … Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*, 94. https://doi.org/10.1038/nature11041

Davey, J. W., Barker, S. L., Rastas, P. M., Pinharanda, A., Martin, S. H., Durbin, R., … Jiggins, C. D. (2017). No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. *Evolution Letters*, *1*, 138–154.

DeBolt, S. (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biology and Evolution*, *2*, 441–453. https://doi.org/10.1093/gbe/evq033

Dobigny, G., Britton-Davidian, J., & Robinson, T. J. (2015). Chromosomal polymorphism in mammals: An evolutionary perspective. *Biological Reviews*, *92*, 1–21. https://doi.org/10.1111/brv.12213

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MULTIQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fan, S., & Meyer, A. (2014). Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Frontiers in Genetics*, *5*, 163. https://doi.org/10.3389/fgene.2014.00163

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*, 85–97. https://doi.org/10.1038/nrg1767

Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., … Scherer, S. W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics*, *1*, 489–498. https://doi.org/10.1371/journal.pgen.0010056

Feulner, P. G. D., Chain, F. J. J., Panchal, M., Eizaguirre, C., Kalbe, M., Lenz, T. l., … Bornberg-bauer, E. (2013). Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular Ecology*, *22*, 635–649. https://doi.org/10.1111/j.1365-294X.2012.05680.x

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, *108*, 1513–1518. https://doi.org/10.1073/pnas.1017351108

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*, 644–652. https://doi.org/10.1038/nbt.1883

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). CIRCLIZE implements and enhances circular visualization in R. *Bioinformatics*, *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ramia, M., Tarone, A. M., … Mackay, T. F. C. (2014). Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Research*, *24*, 1193–1208. https://doi.org/10.1101/gr.171546.113

Ip, C. L. C., Loose, M., Tyson, J., deCesare, M., Brown, B. L., Jain, M., … MinION Analysis and Reference Consortium. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. [version 1; Referees, approved]. *F1000Research*, *4*, 1075.

Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., … Meyer, A. (2018). Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications*, *9*, 3159. https://doi.org/10.1038/s41467-018-05479-9

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61. https://doi.org/10.1038/nature10944

Keane, T. M., Wong, K., Adams, D. J., Flint, J., Reymond, A., & Yalcin, B. (2014). Identification of structural variation in mouse genomes. *Frontiers in Genetics*, *5*, 192.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, *8*, 2040. https://doi.org/10.1371/journal.pbio.1000501

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*, 419–434. https://doi.org/10.1534/genetics.105.047985

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., … Ivanov, P. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows, 87–90.

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375. https://doi.org/10.1038/nature14181

Langley, C. H., Stevens, K., Cardeno, C.,Lee, Y. C., Schrider, D. R., Pool, J. E., … Begun, D. J. (2012). Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, *192*, 533–598.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with BOWTIE 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*, R84. https://doi.org/10.1186/gb-2014-15-6-r84

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, *8*, 2227.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18. https://doi.org/10.1186/2047-217X-1-18

McClintock, B. (1931). Cytological observations of deficiencies involving known genes, translocations and an inversion in *Zea mays*.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*, 344–355. https://doi.org/10.1073/pnas.36.6.344

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, *98*, 12084–12088. https://doi.org/10.1073/pnas.221274498

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*, 780. https://doi.org/10.1038/nmeth.3454

Pinharanda, A., Martin, S. H., Barker, S. L., Davey, J. W., & Jiggins, C. D. (2017). The comparative landscape of duplications in *Heliconius melpomene* and *Heliconius cydno*. *Heredity (Edinb)*, *118*, 78–87. https://doi.org/10.1038/hdy.2016.107

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. I., Zhang, Y., Hurles, M. E., … Hall, I. M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, *20*, 623–635. https://doi.org/10.1101/gr.102970.109

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

R Development Core Team, R. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org.

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution*, *16*, 351–358. https://doi.org/10.1016/S0169-5347(01)02187-5

Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C., & Corbett-Detig, R. (2018). Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proceedings of the National Academy Sciences*, *115*, 5492–5497.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., … Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, *39*, S7–S15.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A., Hubley, R., & Green, P. (2008). *RepeatModeler*. Seattle, WA: Institute for Systems Biology.

Smit, A., & Hubley, R., Green, P. (2013). 2013–2015. REPEATMASKER OPEN-4.0.

Spielmann, M., & Mundlos, S. (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. *BioEssays*, *35*, 533–543. https://doi.org/10.1002/bies.201200178

Sturtevant, A. H. (1921). A case of rearrangement of genes in *Drosophila*. *Proceedings of the National Academy of Sciences*, *7*, 235–237. https://doi.org/10.1073/pnas.7.8.235

Subramanian, S. (2016). The effects of sample size on population genomic analyses – implications for the tests of neutrality. *BMC Genomics*, *17*, 123. https://doi.org/10.1186/s12864-016-2441-8

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*, 75–81. https://doi.org/10.1038/nature15394

Thornton, K. (2005). Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics*, *171*, 2143–2148.

Wade, C. M., Kulbokas, E. J., Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., … Daly, M. J. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, *420*, 574. https://doi.org/10.1038/nature01252

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*, 125–138. https://doi.org/10.1038/nrg3373

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*, 427–440.

Wellenreuther, M., Le Luyer, J., Cook, D., Ritchie, P. A., & Bernatchez, L. (2018). Domestication and temperature modulate gene expression signatures and growth in the Australasian snapper *Chrysophrys auratus*. *Genes, Genomes, Genetics*, *9*, 105–116.

Wickham, H., & Francois, R. (2017). DPLYR: a grammar of data manipulation, 2013. Retrieved from https://github. com/hadley/dplyr. version 0.1.[p 1].

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Ye, K., Hall, G., & Ning, Z. (2015). Structural variation detection from next generation sequencing. *Journal of next Generation Sequencing & Applications*, *1*, S1:007. https://doi.org/10.4172/2469-9853.S1-007

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*, 1311–1320. https://doi.org/10.1126/science.1251385

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Catanach A, Crowhurst R, Deng C, David C, Bernatchez L, Wellenreuther M. The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Mol Ecol*. 2019;28:1210–1223. https://doi.org/10.1111/mec.15051