Widespread deviant patterns of heterozygosity in whole-genome sequencing due to autopolyploidy, repeated elements, and duplication

Xavier Dallaire, Raphael Bouchard, Philippe Hénault, Gabriela Ulmo-Díaz, Eric Normandeau, Claire Mérot, Louis Bernatchez, Jean-Sébastien Moore

Widespread Deviant Patterns of Heterozygosity in Whole-Genome Sequencing Due to Autopolyploidy, Repeated Elements, and Duplication

Xavier Dallaire [1,2,*], Raphael Bouchard[1,3], Philippe Hénault[1,3], Gabriela Ulmo-Diaz[1,3], Eric Normandeau[1,3,4], Claire Mérot[5], Louis Bernatchez [1,3,†], and Jean-Sébastien Moore[1,2,3]

[1]Institut de biologie intégrative et des systèmes, Université Laval, Québec, Canada
[2]Centre d'Études Nordiques, Université Laval, Québec, Canada
[3]Ressources Aquatique Québec, Université de Rimouski, Rimouski, Canada
[4]Plateforme de bio-informatique de l'IBIS, Université Laval, Québec, Canada
[5]CNRS, UMR 6553 ECOBIO, Université de Rennes, Rennes, France

[†]Louis Bernatchez passed away on September 28th 2023.

*Corresponding author: E-mail: xavier.dallaire.2@ulaval.ca.

Accepted: November 30, 2023

Abstract

Most population genomic tools rely on accurate single nucleotide polymorphism (SNP) calling and filtering to meet their underlying assumptions. However, genomic complexity, resulting from structural variants, paralogous sequences, and repetitive elements, presents significant challenges in assembling contiguous reference genomes. Consequently, short-read resequencing studies can encounter mismapping issues, leading to SNPs that deviate from Mendelian expected patterns of heterozygosity and allelic ratio. In this study, we employed the ngsParalog software to identify such deviant SNPs in whole-genome sequencing (WGS) data with low (1.5×) to intermediate (4.8×) coverage for four species: Arctic Char (*Salvelinus alpinus*), Lake Whitefish (*Coregonus clupeaformis*), Atlantic Salmon (*Salmo salar*), and the American Eel (*Anguilla rostrata*). The analyses revealed that deviant SNPs accounted for 22% to 62% of all SNPs in salmonid datasets and approximately 11% in the American Eel dataset. These deviant SNPs were particularly concentrated within repetitive elements and genomic regions that had recently undergone rediploidization in salmonids. Additionally, narrow peaks of elevated coverage were ubiquitous along all four reference genomes, encompassed most deviant SNPs, and could be partially associated with transposons and tandem repeats. Including these deviant SNPs in genomic analyses led to highly distorted site frequency spectra, underestimated pairwise F_{ST} values, and overestimated nucleotide diversity. Considering the widespread occurrence of deviant SNPs arising from a variety of sources, their important impact in estimating population parameters, and the availability of effective tools to identify them, we propose that excluding deviant SNPs from WGS datasets is required to improve genomic inferences for a wide range of taxa and sequencing depths.

Key words: heterozygosity, salmonid, whole-genome sequencing, paralog, autopolyploid, repetitive DNA.

Significance

Genomes can be very repetitive and hard to assemble into a reference, which can lead to biases when genotyping genetic markers in complex genomic regions. Here, we draw attention to this issue in various whole-genome datasets and validate a method to identify problematic single nucleotide polymorphisms (SNPs). We also explore processes creating such SNPs and their consequences on common population genomics analyses.

Introduction

Single nucleotide polymorphisms (SNPs) are now the most commonly used genetic markers in the field of genomics due to their widespread distribution within genomes, rich information content, and their detectability using a diverse range of sequencing technologies (e.g. restricted site associated DNA sequencing [RAD-seq], Andrews et al. 2016; whole-genome sequencing [WGS], Fuentes-Pardo and Ruzzante 2017; low-coverage whole-genome sequencing [lcWGS], Therkildsen and Palumbi 2017). Their utilization, however, relies on several assumptions, such as their biallelic nature, adherence to Mendelian inheritance, and, in some cases, independent segregation (absence of linkage disequilibrium). Multiple studies highlight the ongoing need to validate these assumptions (Hurles 2002; Chen et al. 2014; Jaegle et al. 2023), especially in complex genomes such as those that experienced polyploidization.

Polyploidization is the process by which a complete set of chromosomes is multiplied within a single nucleus and then passed on to progenies (Zhang et al. 2019; Todesco et al. 2020). It can happen following the hybridization of different species (allopolyploidy), in which case pairs of chromosomes descended from each species conserve preferential bivalent pairing (Cifuentes et al. 2010; Mason and Wendel 2020). In other cases, polyploidy can result from whole-genome duplication (WGD) events within a single genome (autopolyploidy), which creates groups of chromosomes in multivalent recombining pairing during meiosis. This tetrasomic inheritance can last until mutations create enough sequence divergence to reestablish bivalent pairing in a now doubled number of chromosomes, a process known as rediploidization (Ohno 1970; Weiss and Maluszynska 2004; Lien et al. 2016).

Due to a recent WGD in their common ancestor around 88 to 103 Mya, salmonids have been central to the study of autopolyploidy in animals (Macqueen and Johnston 2014). This WGD is the fourth (hence referred to as Ss4R) to occur along the lineages ancestral to salmonids, after two WGDs in early vertebrates (Simakov et al. 2020) and another one in the ancestor of all teleost fishes, around 320 to 350 Mya (Jaillon et al. 2004; Glasauer and Neuhauss 2014). Following Ss4R, rediploidization occurred at different times across duplicated regions, leading to either ancestral or lineage-specific (LORe) ohnolog resolution (Robertson et al. 2017). This is apparent when comparing levels of sequence divergence between pairs of syntenic ohnologs in salmonid assemblies. These ohnologs display from 85% to nearly 100% nucleotide identity (Lien et al. 2016; Gundappa et al. 2022; Smith et al. 2022; Mérot et al. 2023).

An additional source of complexity in salmonid genomes is their relatively high content in transposable elements (TE), which can make up 50% to 60% of assembled reference sequences (Minkley 2018). The correspondence in the timing of Ss4R and the proliferation of TE in salmonid genomes suggests that the WGD event may have disrupted TE regulation processes (Lien et al. 2016). Similar to what is observed in hybrids (Hénault et al. 2023; Laporte et al. 2019), TE expansion may, in turn, have contributed to the sequence divergence and chromosomal rearrangements that enabled rediploidization (Lien et al. 2016). In ray-finned fishes, the abundance of TEs is also closely correlated to genome size (Chalopin et al. 2015; Gao et al. 2016). Given the large genome size of salmonids (2.5 to 3.0 Gb), such abundance of TE, in combination with high sequence identity between pairs of recently rediploidized chromosomal regions, has considerably hindered efforts to assemble quality reference genomes in salmonids (Lien et al. 2016; Smith et al. 2022).

Due to their socio-economic importance and their scientific relevance for the investigation of many evolutionary and ecological processes (e.g. local adaptation and speciation), salmonid species have been extensively studied using short-read resequencing technologies like RAD-seq (Elmer 2016). By analyzing data from salmonids and other taxa derived from an ancestral WGD, it has become apparent that collapsed assemblies and under-splitting can bias the genotyping of SNPs on highly similar duplicated loci when they are considered as a single region (Harvey et al. 2015). When overlooked, these biases can have an impact on biological and management interpretations derived from genomic data (O'Leary et al. 2018), for example by creating a false signal of differentiation (Larson et al. 2021).

Identifying SNPs on paralogous or other multicopy sequences has been a long-lasting challenge in genomics. A variety of methods have been applied for their detection, such as the identification of apparent heterozygotes in haploid samples (Sánchez et al. 2009; Hecht et al. 2013) or increased depth of coverage compared to single-copy loci (Dou et al. 2012; Davey et al. 2013). The HDplot method (McKinney et al. 2017) is now commonly used to detect SNPs that do not conform to expected patterns of Mendelian inheritance by identifying deviations from Hardy–Weinberg Equilibrium (H) or the expected 1:1 allelic ratio in heterozygotes (D). We hereafter refer to such variants as "deviant" SNPs, as opposed to "canonical" SNPs that conform to expected patterns of heterozygosity and allelic ratio for nonduplicated loci (after Karunarathne et al. 2022).

In Chinook Salmon (*Oncorhynchus tshawytscha*; McKinney et al. 2017), deviant SNPs identified by the HDplot method (17% of all SNPs) predominantly matched paralogs identified by haploid mapping (McKinney et al. 2016) and were especially dense in chromosome arms with ongoing residual tetrasomy, suggesting that paralogy stemming from the ancestral WGD is the main source of deviant SNPs in salmonids. This method has since been applied to other salmonid species (*Oncorhynchus mykiss*:

22% of deviants, Fraik et al. 2021; *Oncorhynchus kisutch*: Xuereb et al. 2022; *Salvelinus alpinus*: 24%, Dallaire et al. 2021) as well as other taxonomic groups including polyploid trees (e.g. *Pinus cembra*: 85% of deviants, Rellstab et al. 2019), crustaceans (e.g. *Homarus americanus*: 22 to 40% of deviants, Dorant et al. 2020, 2022), and ranids (e.g. *Rana luteiventris*: 16% of deviants, Cayuela et al. 2022). In the last two examples, the HDplot method was adapted to identify and genotype copy number variants suspected to be associated with TE (Dorant et al. 2020; Cayuela et al. 2021). Altogether, these recent studies highlight the fact that deviant SNPs can also be found in significant proportions in datasets from non-polyploid species, but specific causal processes have yet to be examined in such cases.

In recent years, lcWGS has emerged as an alternative to reduced-representation sequencing for population genomic studies in non-model species (Therkildsen and Palumbi 2017). This cost-effective method offers the opportunity for unprecedented sample sizes of whole-genome sequences by reducing the per-sample depth of coverage to as low as 0.1×. At low and medium coverages (under 10×), the lack of confidence around individual genotypes can be circumvented by using a genotype likelihood (GL) framework, as implemented in ANGSD (Kornelissen et al. 2014). A growing variety of tools explicitly account for genotype uncertainty, allowing such data to be used in many common population genomics applications (Lou et al. 2021).

ngsParalog (Linderoth 2018) is an openly available bioinformatic software that uses signals similar to HDplot to test the hypothesis that the mismapping of reads creates deviant SNPs at specific genomic positions. By using a probabilistic approach that avoids genotype calling and takes into account the uncertainty inherent to low-coverage approaches, *ngsParalog* is reported to be able to detect deviant SNPs in next-generation sequencing datasets with coverage as low as 2× (Linderoth 2018). This tool has been used both on RAD-seq (Saglam et al. 2017; Benjamin et al. 2018; Hemstrom et al. 2022) and WGS (Márquez et al. 2020; Pope et al. 2023) datasets. However, none of these studies reported the number of SNPs filtered out by this approach.

In this study, we aimed to answer three main questions: (i) How can we reliably detect deviant SNPs in low- to intermediate-coverage WGS datasets? (ii) What are the main causes for deviant SNPs and how are they distributed in the studied genomes? (iii) What are the consequences of including deviant SNPs in common population genomic analyses? We developed and applied a common variant calling and filtering pipeline to new WGS datasets (1.5 to 2× of coverage) from two salmonid species (*Salvelinus alpinus* and *Coregonus clupeaformis*) and reanalyzed data on Atlantic Salmon (*Salmo salar*; 4.8×) published in Bertolotti et al. (2020) and on American Eel (*Anguilla rostrata*; 4×) published in Ulmo-Diaz et al. (2023). These datasets vary in sampling size, depth of coverage, population structure, and duplication history, in order to allow general conclusions that should apply to a wider variety of realistic situations. We used both ngsParalog and HDplot to compare their capacities in classifying deviant SNPs in low- to intermediate-coverage data. We then mapped canonical and deviant SNPs and compared their genomic distributions in relation to peaks of elevated coverage, repetitive regions, and regions inferred to have experienced delayed rediploidization in salmonids. Finally, we compared the results of common population genomics statistics and analyses before and after filtering for deviant SNPs to assess the impacts of these filters on inferences that are commonly drawn from WGS datasets.

Results

Identification and Validation of Paralog SNPs at Low Coverage

SNPs flagged as deviant by *ngsParalog* represent different proportions across datasets but they showed consistent characteristics such as low F_{IS} and high coverage. After subsampling, we obtained five datasets with an average depth of coverage of around 1.5× and a mode between 1.82× and 1.97× (Table 1; see supplementary fig. S1, Supplementary Material online for distribution of depth). We identified SNPs using ANGSD and calculated the likelihood of reads being misaligned at the positions of those SNPs using ngsParalog ($P < 0.001$). While deviant SNPs were in majority in both Lake Whitefish datasets (49.9% and 61.8%) and in Arctic Char (62.3%), they represented 22.6% of the SNPs in Atlantic Salmon and 10.6% in the American Eel (Table 1). Nearly all deviant SNPs had F_{IS} ranging from −1 to −0.05, while canonical SNPs had an F_{IS} distribution centered around 0 (Fig. 1A), except for the Arctic Char dataset where strong population structure (max pairwise $F_{ST} = 0.45$) created a deficit of heterozygotes (positive F_{IS}) putatively due to a Wahlund effect (Wahlund 1928). In datasets with depths of coverage around 1.5×, the average coverage was consistently higher at the positions of deviant SNPs than canonical SNPs (Fig. 1D). For example, 37% to 58% of the deviant SNPs had coverage higher than 3×, while less than 0.7% of canonical SNPs reached that depth in all species studied.

The characterization of deviant SNPs showed high consistency between the different methods employed. When visualizing SNPs using *HDplot* (McKinney et al. 2017), we observed that the vast majority of SNPs categorized as deviant by *ngsParalog* either had a high proportion of

Table 1
Summary of datasets reanalyzed in this study. See main text for the treatment of each dataset

Dataset	Family	GenBank assembly accession (species; Contig N50)	Origin	Number of sampling sites	Sample size	Sequencer	Original depth of coverage by sample; mean (SD)	Mode of depth in subsampled data	Number of SNP	Number of deviant SNP	% of deviant SNPs
American Eel (*Anguilla rostrata*)	*Anguillidae*	GCA_018555375.2 (*Anguilla rostrata*; N50 = 5.02 Mb)	East Coast of North and Central America	21	460	NovaSeq 6000 S4 PE150	3.68 (0.9)	1.97	16,379,954	1,729,598	10.6
Arctic Char (*Salvelinus alpinus*)	*Salmonidae*	GCA_002910315.2 (*Salvelinus sp.*; N50 = 55 kb)	Arctic Canada	16	520		1.65 (0.51)	1.88	5,408,454	3,371,454	62.3
Lake Whitefish (*Coregonus clupeaformis*)	*Salmonidae*	GCA_018398675.1 (*Coregonus clupeaformis*; N50 = 5.73 Mb)	Great Slave Lake, Canada James Bay, Canada	9 13	298 470		1.48 (0.34) 1.71 (1.90)	1.63 1.88	4,491,496 7,318,227	2,777,770 3,651,681	61.8 49.9
Atlantic Salmon (*Salmo salar*)	*Salmonidae*	GCA_905237065.2 (*Salmo salar*; N50 = 28.06 Mb)	Coastal Norway	55	126 228	HiSeq 2500 PE125 HiSeq 3000 PE100	7.46 (1.72) 4.79 (0.75)	1.82	4,392,892	990,766	22.6

heterozygotes or deviated from the 1:1 expected allelic ratio in heterozygotes (Z further from zero than canonical SNPs) (Fig. 1B and C). Among SNPs in excess of heterozygotes ($F_{IS} < 0$ and $P < 0.05$ for a Hardy–Weinberg equilibrium test in ANGSD), most were categorized as deviant by *ngsParalog*: 96.9% to 99.6% in the Lake Whitefish and Arctic Char datasets, 89.7% in the American Eel, and 78.3% in the Atlantic Salmon. For SNPs with lower minor allele frequency (MAF), the distribution of the average allelic ratio (in heterozygotes with more than 4× of coverage) for deviant SNPs had an obvious mode around 0.25 (1:3) and smaller one at 0.75 (3:1). For canonical SNPs, the average allelic ratio was centered on 0.50 (1:1). 0.8% to 4.3% of the SNPs were categorized as deviant only by *ngsParalog*, most of which had low values of MAF (average MAF = 0.15). In contrast, 0.5% to 3.7% of SNPs were categorized as deviant only by *HDplot* (average MAF = 0.33), and they largely overlapped the canonical SNP distribution for observed heterozygosity and allelic ratio (supplementary fig. S1, Supplementary Material online).

Prevalence of Deviant SNPs at Different Depths of Coverage

Deviant SNPs were not only observed in low-coverage sequencing, since we inferred their presence from low (1×) to intermediate (4.8×) depth of coverage in the Atlantic Salmon dataset. We categorized SNPs from the Atlantic Salmon dataset before subsampling (4.8×) using *ngsParalog*, then compared the list of SNPs retained in subsampled datasets at decreasing depths of coverage (Fig. 2). First, 49.8% of the SNPs (a total of 4.52 million) were characterized as deviants in the original dataset sequenced at intermediate coverage (4.8×). The number of canonical SNPs and the depth of coverage had a plateau-like relation: it was nearly stable between 4.8× and 3× (93.3% of SNPs left) and decreased between 3× and 1.5× (69.2% of SNPs left). On the other hand, the number of deviant SNPs (categorized based on the original dataset) did not reach a plateau between 1.5× and 4.8×, as the number of deviant SNPs was almost directly proportional to the average depth of coverage (Pearson's $R^2 = 0.98$).

We repeated the ngsParalog analysis on the subsampled Atlantic Salmon datasets to compare if the categorization of SNPs found in the 4.8× dataset changed at lower depths. Assuming the categories derived from the 4.8× dataset were closer to reality, subsampling the data led to increasing rates of deviant SNPs categorized as canonical (putative false negative) that reached 17.9% of deviant SNPs (5.9% of all SNPs) in the 1.5× dataset (hatched portion in Fig. 2). However, canonical SNPs categorized as deviant (putative false positive) were much rarer and peaked at 0.4% of canonical SNPs (0.2% of all SNPs) in the 4× dataset before decreasing in lower coverage datasets.

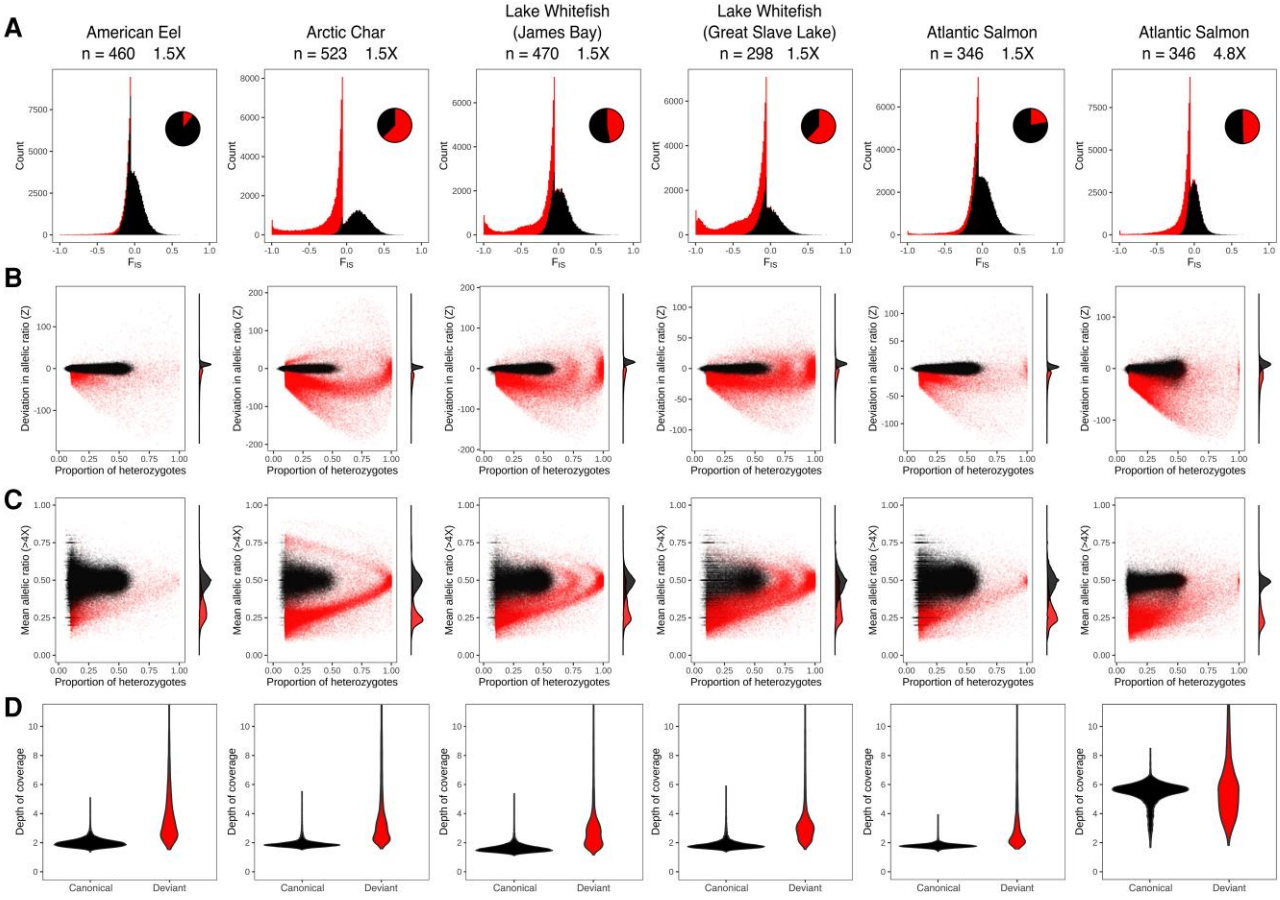

Fig. 1.—All investigated datasets harbor SNPs in deviation of Hardy–Weinberg equilibrium and allelic ratio. Summary of canonical (black) and deviant (red) SNPs as categorized by ngsParalog ($P < 0.001$) for 100,000 randomly selected SNPs in the American Eel, Arctic Char, Lake Whitefish (James Bay and Great Slave Lake), and Atlantic Salmon (1.5× and 4.8× coverage) datasets. A) Histograms of F_{IS} with an inset pie chart showing the proportion of SNPs by category. B) HDplot showing the proportion of heterozygotes in relation to deviation in allelic ratio (Z-score). C) Proportion of heterozygotes in relation to mean allelic ratio in heterozygous samples with at least 4× coverage. D) Distribution of depth of coverage by SNP category. The y axis was restricted to depth under 10× for clarity, but deviant SNPs had maximum depths that greatly exceeded 10×.

Distribution of Deviant SNPs in the Genome

The distribution of deviant SNPs was at least partly consistent with the hypothesis that they are caused by ancient polyploidization in salmonids. In all three salmonid species, the density of canonical SNPs decreased as the percentage of identity increased between the self-syntenic blocks (ohnolog pairs of sequence resulting from ancient WGD), while the density of deviant SNPs increased (Fig. 3, supplementary table S1, Supplementary Material online). In Arctic Char, Lake Whitefish, and Atlantic Salmon respectively, 7.9%, 28.6%, and 30.5% of the deviant SNPs were found in self-syntenic blocks with an identity above 95%, while 3.9%, 9.3%, and 12.7% of canonical SNPs were found in those same blocks.

The distribution of deviant SNPs was also associated with the presence of repetitive regions (i.e. interspersed and tandem repeats). For American Eel, Arctic Char, and Atlantic Salmon, both canonical and deviant SNPs were found in higher density in repetitive regions than in nonrepetitive regions, but the effect was strongest for deviant SNPs in American Eel and Arctic Char (Fig. 3, supplementary table S1, Supplementary Material online). In Lake Whitefish, canonical SNPs were less dense in repetitive regions, while deviant SNPs were denser.

The association between deviant SNPs and repeated elements was evident when looking at peaks of elevated coverage found in all datasets that varied in size in the order of tens to hundreds of bp (Table 2). While these high-coverage regions covered a relatively small portion of each genome, deviant SNPs were strongly over-represented in those peaks of coverage ($\chi^2 > 493{,}507$; ddf = 1) while canonical SNPs were under-represented ($\chi^2 > 3{,}230$). We explored the interaction between depth of coverage, deviant SNPs, and repetitive elements (Fig. 4A). Depending on the dataset, the peaks of elevated coverage were enriched in certain types of repetitive elements, namely long-terminal

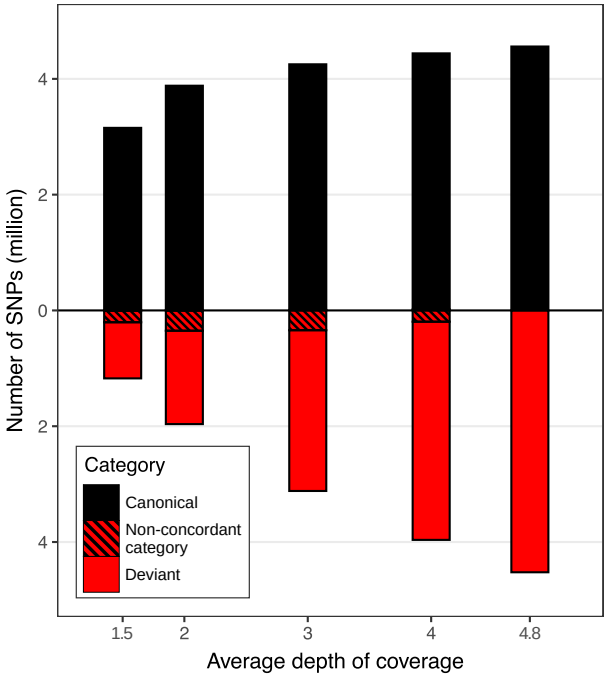

Fig. 2.—Deviant SNPs are found in low- to intermediate-coverage datasets. Number of canonical (black, above) and deviant (red, below) SNPs as categorized by ngsParalog in the 4.8× Atlantic Salmon dataset. Deviant SNPs categorized as canonical in the subsampled datasets are represented by the hatched portion of bars, but canonical SNPs categorized as deviant were too rare to be visualized. SNPs absent from the 4.8× dataset (less than 1.5% of all SNPs) were not shown.

repeat (LTR) retrotransposons, satellites, simple repeats, and low complexity regions, as defined by RepeatMasker (Fig. 4B, supplementary table S2, Supplementary Material online). This enrichment was especially strong and extended to most repeat types in the American Eel dataset.

Consequences of Deviant SNPs on Population Genomics Analysis

Including deviant SNPs or excluding them had a strong impact on population genomic statistics, such as shared polymorphism, genetic differentiation, and nucleotide genetic diversity. When comparing SNPs found for Whitefish in James Bay (JB; 7,318,277 SNPs) and Great Slave Lake (GSL; 4,491,496 SNPs), we found 4,037,742 SNPs at common positions, 3,714,622 of which had the same alternative alleles. Those common polymorphisms thus represented 82.7% of those found in GSL and deviant SNPs were found in similar proportions in the common list (65.8%) and in GSL (67.0%). However, 3.7% of the common SNPs were categorized as deviant only in the GSL dataset, and 2.7% were deviant only according to the JB dataset (supplementary fig. S2A, Supplementary Material online). Common SNPs categorized as deviant in both datasets showed highly correlated MAF ($R^2 = 0.90$, supplementary fig. S2B, Supplementary Material online) and F_{IS} ($R^2 = 0.90$, supplementary fig. S2C, Supplementary Material online), compared to other SNPs (canonical: $R^2 = 0.63$ and 0.07; SNPs with nonconcordant categories: $R^2 = 0.56$ and −0.31).

In the Arctic Char dataset, pairwise F_{ST} values estimated between 16 populations (120 unique pairs) with all SNPs and only canonical SNPs were highly correlated ($R^2 = 0.88$), but estimates were 2.02 times higher when using only canonical SNPs (Fig. 5A). When inspecting two-dimensional site frequency spectra (2dSFS) between sampling sites with different levels of divergence, we found that deviant SNPs were consistently located along the diagonal and concentrated at low (<0.1) and high (0.5) MAF (Fig. 5B and C). To summarize, failure to filter out deviant SNPs leads to underestimating genetic differentiation indices between populations by increasing the number of shared polymorphisms at similar frequencies.

We masked regions around deviant SNPs to measure their impact on nucleotide genetic diversity estimation. This masked 5.2% of the *Salvelinus* sp. genome (encompassing 9.1% and 98.1% of canonical and deviant SNPs, respectively) and 3.6% of the American Eel genome (3.4% and 97.6% of canonical and deviant SNPs, respectively). In Arctic Char, masking deviant regions in various populations led to a 39.9% to 51.6% decrease in per-site Watterson's estimator (Θ_W; Fig. 6A) and a 36.4% to 61.8% decrease in per-site nucleotide diversity (Θ_π; Fig. 6B). The effect on Tajima's D (Θ_W/Θ_π) ranged from −0.46 to +0.32 (Fig. 6C). In American Eel, masking deviant regions in subsamples of individuals led to a 3.1% to 3.5% decrease in per-site Θ_W, a 6.6% to 6.8% decrease in per-site Θ_π, and the effect on Tajima's D ranged from −0.054 to +0.060 (Fig. 6). Despite weaker effects in American Eel, nucleotide genetic diversity estimates decreased across more than 99% of the genome length (in 100 Mb windows).

Discussion

In this study, we aimed to investigate the prevalence of deviant SNPs in WGS data by applying a common variant calling pipeline to both new and previously published datasets covering four different fish species. We found a significant proportion of SNPs to be in deviation from expected patterns of heterozygosity and allelic ratio in salmonid datasets, and to a lesser extent in the American Eel. We attribute most of these deviant SNPs to collapsed assembled genomic regions, which is frequent in salmonid assemblies because of a recent rediploidization, as well as to repetitive sequences. Considering the widespread occurrence of deviant SNPs arising from a variety of sources, their important

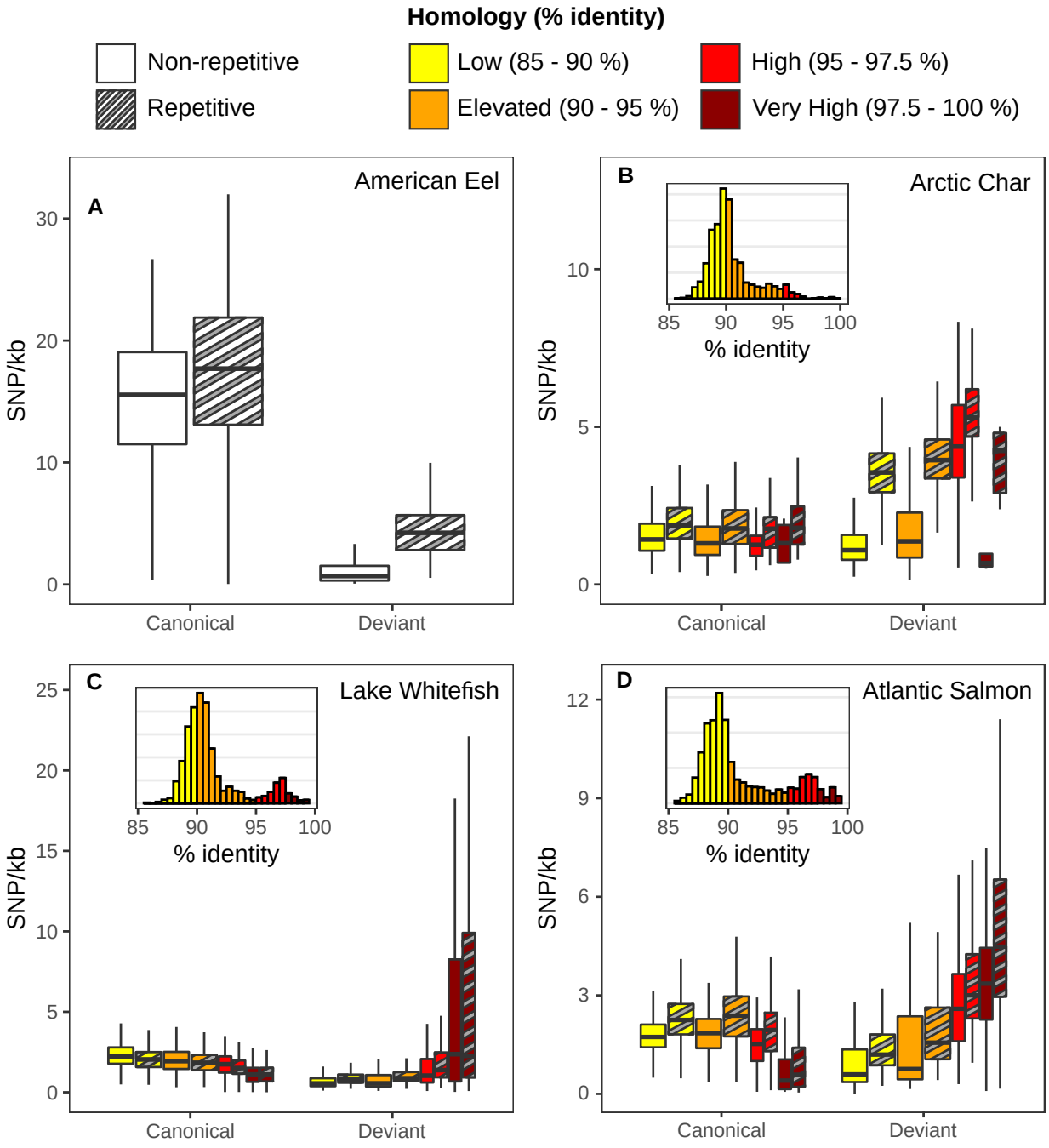

Fig. 3.—Deviant SNPs are more common in repetitive DNA and recently rediploidized regions. Density of canonical and deviant SNPs (by kb) in 1 Mb nonoverlapping windows for the A) American Eel, B) Arctic Char, C) James Bay Lake Whitefish, and D) 4.8× Atlantic Salmon datasets. SNPs are split based on whether they were found in repetitive regions identified by Repeat Masker (hatched) or not (plain boxplot). Windows are categorized based on their percentage of identity with their ohnolog, color-coded from yellow to dark red. For salmonid datasets, the inset histograms show the relative frequency of percentages of identity in windows. Only base pairs with an average depth above 0.75× were considered for the calculation of SNP densities.

impact in estimating population parameters, and the availability of effective tools to identify them, we suggest that excluding deviant SNPs from WGS datasets is required to improve genomic inferences for a wide range of taxa and sequencing depths.

Detection of Paralogs in lcWGS Datasets

We applied *ngsParalog* on a variety of datasets for salmonid species where we expected high numbers of deviant SNPs, as well as one nonsalmonid fish. By using the *HDplot* as a visual validation, we observed that deviant SNPs detected

Table 2
Characteristics of four datasets regarding the sufficiently covered part of the genome (depth of coverage above 0.75×) and peaks of elevated coverage (above 3× in American Eel, Arctic Char, and Lake Whitefish; above 8× in Atlantic Salmon). The percentage of canonical and deviant SNPs (as categorized by ngsParalog) inside peaks of coverage is shown

Species	% of the genome sufficiently covered	Median peak width (bp)	95th quantile of peak width (bp)	% of sequenced genome in peaks	% of canonical SNPs in peaks	% of deviant SNPs in peaks
American Eel	92.5	30	329	2.1	1.0	61.7
Arctic Char	83.4	76	398	3.2	1.3	59.7
Lake Whitefish (James Bay)	83.3	39	821	4.6	1.9	64.4
Atlantic Salmon (4.8×)	90.2	65	457	1.8	0.4	27.6

by *ngsParalog* displayed observed heterozygosities and allelic ratios in perfect accordance with the theoretical and simulated distributions presented in McKinney et al. (2017). Deviant SNPs were very common in the Arctic Char and Whitefish datasets, representing >50% of all SNPs. While we cannot accurately evaluate false positive and negative rates, we can infer that increased sample size ($n = 298$ and 470 in whitefish datasets) led to better-defined groups of SNPs (canonical vs deviant) in HDplot. This is likely reflected in the detection power for deviant SNPs when using *ngsParalog*, as this method uses diagnostic factors similar to those used with the *HDplot* method.

Despite similar results for the two methods used, we found *ngsParalog* to be more suited to deviant SNP detection when genotype calls are not available, as it explicitly tests a two-loci model using sequencing data (Linderoth 2018). In contrast, our implementation of *HDplot* relies on an imperfect identification of heterozygotes using a single-locus model in ANGSD which poorly fit the data in the case of deviant SNPs. Here, we opted for a straightforward calculation of *P*-values across whole datasets without accounting for a Wahlund effect, since *ngsParalog* simultaneously tests for (i) deviations from expected genotype distribution between individuals and (ii) allelic ratios within individuals, which should not be affected by population structure. An alternative would be to repeat the *ngsParalog* process at population level and combine *P*-values through Fisher's method. However, this led to the same conclusions as the whole-dataset approach for at least 98.8% of SNPs even in our most structured dataset (results not shown).

The subsampled Atlantic Salmon dataset (1.5×) had a much lower proportion of deviant SNPs than other salmonid datasets, but this was not the case in the original higher-coverage data (4.8×). In fact, subsampling the Atlantic Salmon dataset led to most deviant SNPs no longer passing ANGSD coverage filters, while canonical SNPs were mostly conserved. However, it is difficult to predict how an increased sequencing effort would affect the already high proportions of deviant SNPs in the Arctic Char and Lake Whitefish datasets because those differ in many ways from the Atlantic Salmon data. First, we used a recent version of the Atlantic Salmon reference genome (Ssal_v3.1), assembled with more resources and of better overall quality than the *Salvelinus* sp. and *Coregonus clupeaformis* references (Table 1). Second, while all other datasets were generated by a common protocol using Nextera libraries and Illumina NovaSeq 6000 S4 (paired-end reads of 150 bp), the Atlantic Salmon data were pieced together from multiple batches on Illumina HiSeq (read length of 100 or 125 bp). Since deviant SNPs are above all caused by mismapping on the reference genome as shown above, these differences might influence the main source of deviant SNPs in the Atlantic Salmon data. This is exemplified by the fact that paralogs were not as predominantly associated with peaks of elevated coverage as in the other datasets. Nevertheless, our analyses at different depths of coverage in Atlantic Salmon strongly suggest that high proportions of deviant SNPs are not exclusive to low-coverage datasets and are likely pervasive in higher-coverage datasets.

For organisms without recent polyploidization (i.e. not expected to produce datasets harboring high levels of paralogs), as the American Eel included here, common filters include setting a maximum depth of coverage per SNP (e.g. four times the average depth). This should avoid SNP calling in collapsed regions where the alignment of reads from multiple loci leads to a localized increase in coverage. However, in the American Eel data, canonical and deviant SNPs had somewhat overlapping depth distributions, and only 3.2% of the deviant SNPs had extreme values for depth of coverage (>8×). Another common approach is to remove SNPs with a strong excess of heterozygotes (Hardy–Weinberg deviation) since few biological processes can explain such patterns in the distribution of genotypes. However, we found that we did not have the power to detect such excesses for most deviants in the American Eel datasets as they appeared to be at much smaller MAFs (and thus expected heterozygosities) than in Arctic Char and Lake Whitefish. This might be caused by panmixia in the American Eel (Côté et al. 2013; Ulmo Diaz et al. in prep) or by higher copy numbers for deviant SNPs created by repetitive DNA than for those arising from residual tetrasomy and delayed rediploidization like in salmonids. Based on these observations, we argue that simpler filtering steps might miss the majority of deviant SNPs in a wide variety

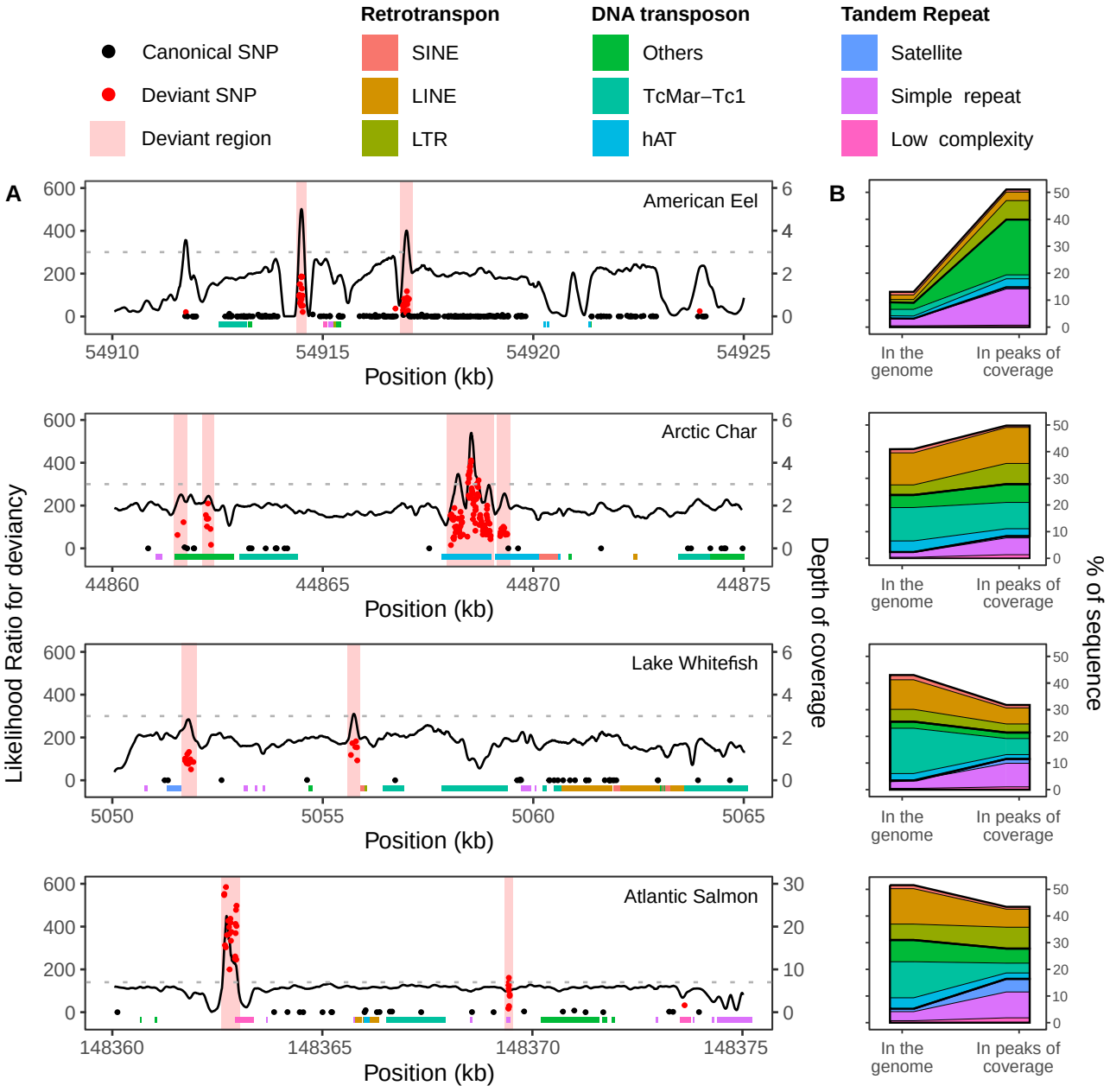

Fig. 4.—Peaks of elevated coverage are enriched in both deviant SNPs and certain classes of repetitive DNA elements. A) Depth of coverage in 15 kb windows on the first chromosome of the American Eel, Arctic Char, Lake Whitefish, and Atlantic Salmon (from top to bottom; the first chromosome was arbitrarily selected and is shown here as a representative illustration of patterns observed over the entire genome). The position of canonical (black) and deviant (red) SNPs are marked as points according to their likelihood ratio of being in a mismapped region, according to ngsParalog. The extent of repetitive elements is indicated by colored rectangles at the bottom of the plots, the depth threshold delimiting peaks of coverage is shown by the light gray dashed line, and deviant regions composed of 150 bp windows centered on each deviant SNPs are shadowed in light red. B) Proportion of sequence covered of the most frequent clades of transposable elements and other repeat types in the sufficiently covered portion of the genome (left) compared to peaks of elevated coverage between 20 and 1,000 bp (right).

of datasets and that multiple factors (e.g. excess of heterozygotes, deviation from expected allelic ratio, and depth of coverage) should be jointly considered when filtering deviants. As such, we found *ngsParalog* to offer a resource-unintensive and multifactor program for reliable deviant SNP filtration that could easily be applied to all WGS datasets. Since our analyses show that despite filtering on sequencing and mapping quality, ANGSD is susceptible to call deviant SNPs, we suggest that the use of *ngsParalog* or equivalent programs flagging noncanonical

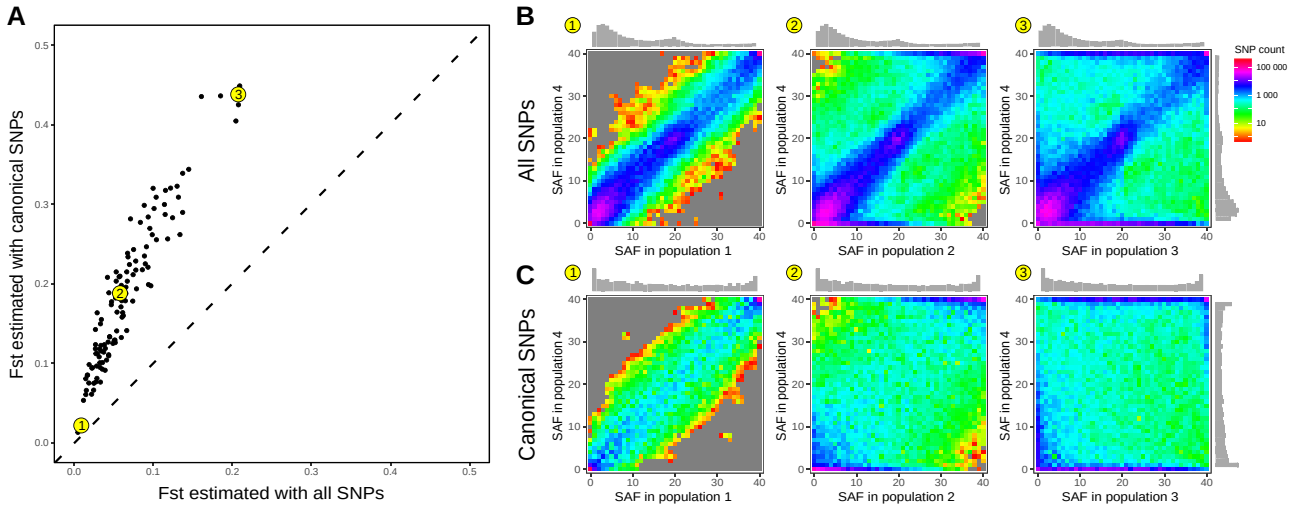

Fig. 5.—Population genetic differentiation is underestimated when deviant SNPs are not removed from the dataset. A) Relationship between pairwise Fst estimated between populations using all SNPs in the Arctic Char dataset and only SNPs categorized as canonical by ngsParalog. Three pairs of populations are highlighted in yellow, corresponding to population 1 (low Fst), 2 (medium Fst), and 3 (high Fst) paired with a fourth common population. Unfolded two-dimensional site frequency spectra (2dSFS) are shown for these pairs using B) all SNPs and C) only canonical SNPs. One-dimensional site frequency spectra (1dSFS) for the four highlighted populations are shown along the axes of the 2dSFS. For better visualization, SNPs fixed for either allele were hidden from the 1dSFS.

SNPs could also be beneficially integrated into pipelines using other variant callers (e.g. *bcftools*, *freebayes*).

Distribution of Deviant SNPs

Next, we investigated the main processes leading to the presence of deviant SNPs in the analyzed datasets by comparing the distribution of canonical and deviant SNPs along the genome. Our observations were consistent with the prediction that deviant SNPs were more frequent in (i) regions of elevated homology due to delayed rediploidization (LORe) and (ii) in repetitive elements. First, LORe regions either have very recently started to differentiate or are still experiencing tetrasomic recombination (Waples et al. 2016; Robertson et al. 2017). This creates important challenges for the linearization of those sequences in reference genomes, which might result in the collapse of both ohnologs into a single consensus sequence, the mismapping of reads, and the creation of deviant SNPs. This is in line with the RAD-seq data on Chinook Salmon shown in McKinney et al. (2017), where putative paralogs were found almost exclusively in chromosome arms expected to have experienced LORe.

However, we found deviant SNPs to be ubiquitously distributed along all four studied genomes. This suggests that delayed rediploidization following a WGD is not the only factor at play, since up to half of deviant SNPs in salmonid datasets were distributed outside of LORe regions. Moreover, we also found numerous deviant SNPs in the American Eel, for which the last WGD event is much older than in salmonids. We observed that deviant SNPs were more frequent inside repetitive sequences and that this effect was especially strong in the American Eel. This supports the idea that interspersed and tandem repeats are another significant source of collapsed assemblies and mismapping. This seemed to happen in narrow peaks of elevated coverage which were extremely dense in deviant SNPs (Fig. 4). Those peaks were of similar size or smaller than reads and were most often disproportionally associated with micro- and minisatellites, as well as LTR transposons. Both TE and tandem repeats, such as satellites, are notoriously challenging to assemble (Treangen and Salzberg 2012; Sotero-Caio et al. 2017; Tørresen et al. 2019), possibly even when using long-read technologies (Liljegren et al. 2016). Tandem repeats are sometimes referred to as genomic "dark matter" in that they are nearly impossible to assemble (Sedlazeck et al. 2018; Weissensteiner et al. 2020) and they hamper the contiguity of assemblies by creating gaps (Star et al. 2011; Peona et al. 2021).

The genome-wide distribution of deviant SNPs we observed here thus apparently arises from a multitude of sources. While some sources of deviant SNPs remain cryptic, we identified processes specific to organisms with recent polyploid ancestors, i.e. the lingering homology between ohnolog pairs of chromosomes, as well as some processes common to all organisms, i.e. repetitive elements. These observations suggest that the problems caused by deviant SNPs are not restricted to highly complex or recently duplicated genomes.

Consequences of Deviant SNPs

Our analyses support the idea that deviant SNPs, no matter their origin, should be removed from genomic datasets

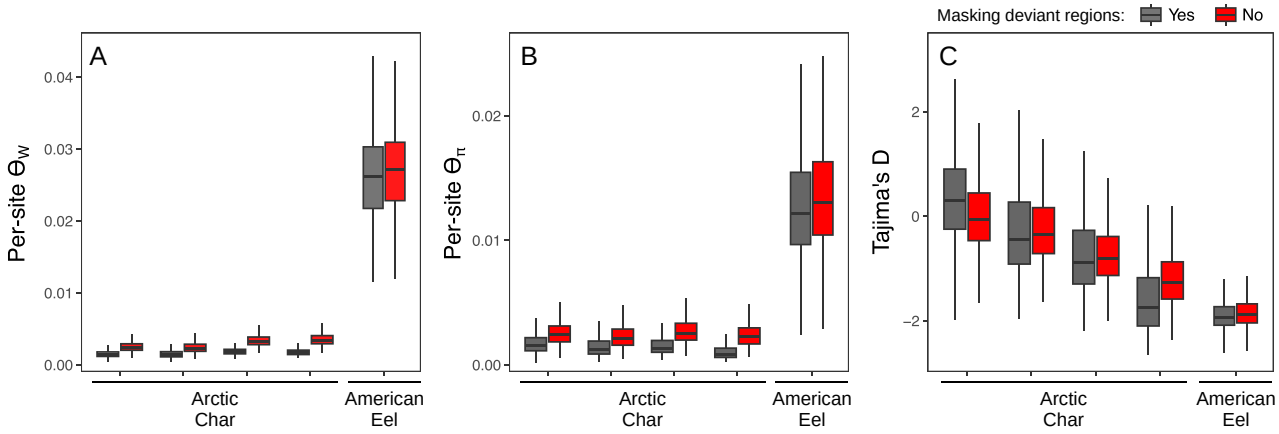

Fig. 6.—Failure to remove deviant SNPs leads to overestimation of genetic diversity in datasets with various deviant SNP densities. Distribution of A) Waterson's estimator (Θ_W), B) nucleotide diversity (Θ_π), and C) Tajima's D in windows of 100 Mb (window step of 20 kb) for four populations ($n = 30$ individuals) of Arctic Char and a random sample of 30 individuals in the panmictic population of American Eel. Diversity estimation was performed before (red) and after (gray) masking a region of 150 bp centered on each deviant SNP.

before proceeding with any population-level analyses. Indeed, we showed that deviant SNPs can create noise or biased signals when quantifying and interpreting the extent of nucleotide polymorphism within and between populations. Similar to Verdu et al. (2016), we reported an overestimation of genetic nucleotide diversity when deviant SNPs were not removed, as is to be expected considering the inclusion of a high proportion of SNPs with inflated heterozygosities. Diverged duplicates, i.e. spurious SNP resulting from two collapsed loci fixed for different alleles, would also increase the apparent genetic diversity as they falsely appear polymorphic. Genetic diversity was much less inflated in the presence of deviant SNPs in the American Eel dataset than in Arctic Char. This could be due to the low content in repetitive elements in the Eel genome, as well as the huge effective size and genetic diversity in the panmictic American Eel population. In other less diverse nonpolyploid species, we could anticipate variable degrees of overestimation of nucleotide genetic diversity when forgoing deviant masking, which underscores the need for caution. As a side note, while treating deviant SNPs as regular nucleotide substitutions inflates canonical nucleotide diversity, they may nevertheless be the reflection of structural variation, an aspect of genetic diversity that remains challenging to quantify.

An abundance of deviant SNPs in the dataset could also obscure fine-scale population structure, as reported in Atlantic Salmon when comparing F_{ST} with unfiltered and filtered data using PMERGE (Nadukkalam Ravindran et al. 2018). Here, we observed that deviant SNPs displayed shared frequencies, even over genetically diverged groups of individuals (e.g. the two geographically distinct Whitefish datasets or distant Arctic Char populations), leading to distorted joint sites frequency spectra. This would create an apparent genetic homogenization of populations, as exemplified by the underestimation of F_{ST} in the Arctic Char dataset before filtering for deviant SNPs.

Apart from F_{ST} estimation, SFS and their joint form are the basis for many other analyses in population genomics, such as neutrality tests (Tajima 1989; Fu and Li 1993), selection scans (Andolfatto 2007; Begun et al. 2007), and demographic inferences (Gutenkunst et al. 2009; Excoffier et al. 2013; Rougeux et al. 2017). Hence, we expect the inclusion of large numbers of SNPs with biased allele frequencies to have more extensive impacts than the underestimation of population differentiation. For example, the enrichment in both low- and high-frequency sites might result in positive or close to zero values for Tajima's D, which could be interpreted as spurious signals for balanced selection when coupled with elevated heterozygosity and shared polymorphism across the studied system (Fijarczyk and Babik 2015). It is hard to predict how the disrupting impacts of deviant SNPs scale with their abundance. However, given the relatively low-effort options available to identify and filter such SNPs, the cautionary principle should be applied and they should be removed from all datasets, even when studying species whose genomes contain comparatively fewer repetitive regions.

Perspectives on Best Practices

Based on our analysis of multiple WGS datasets, we argue that most if not all next-generation sequencing datasets would benefit from a rigorous identification of SNPs deviating from expected patterns of heterozygosity and allelic ratios. Indeed, multiple and sometimes cryptic mechanisms can result in the misalignment of short-read sequences. Although our analyses were limited to low and

intermediate coverage, we can extrapolate that numerous deviant SNPs could also be found in higher-coverage WGS datasets. However, due to budgetary constraints, such datasets are usually characterized by smaller sample sizes, and the resulting decrease in the power to detect deviant SNPs might have obscured the magnitude of the issue until now.

A long-term objective should be to continuously aim for improved genome assemblies, but such concerns fall outside the scope of most resequencing studies, in which the most obvious course of action remains to identify and remove deviant SNPs. This might result in the need to plan for increased sequencing effort when aiming for a specific depth of coverage in complex genomes, as multicopy loci could gather more reads and decrease the effective coverage for canonical SNPs. Alternatively, some progress is being made to mitigate the problem at the source and improve the design of sequencing libraries. For example, duplex-specific nucleases have been used in library preparation to degrade repetitive sequences, increasing the concentration of single-copy fragments to be sequenced (Shagina et al. 2010; Matvienko et al. 2013; Todesco et al. 2020), thus improving the effective depth of coverage for a similar cost.

It is important to note, however, that the extensive and stringent filtering proposed here still retains millions of high-quality canonical SNPs. We show how appropriate filtration applied on rediploidized and other genomes contributes to making lcWGS a major improvement over high-coverage reduced-representation sequencing techniques. Indeed, a higher density of variants, coupled with substantial sample size, has the potential to lead to important advancement for a wide array of applications, such as the identification of peaks of differentiation (genomic regions putatively under divergent selection), genome-wide association studies, and gene–environment interactions.

Materials and Methods

Sequencing and Preprocessing Data

To assess the presence of deviant SNPs in studies using low-coverage sequencing, we applied a common detection pipeline to five datasets, four from salmonid species (Arctic Char; Atlantic Salmon; and two datasets from Lake Whitefish, *Coregonus clupeaformis*) and one from a nonsalmonid teleost fish (American Eel, *Anguilla rostrata*). The Arctic Char and Lake Whitefish datasets were produced by WGS (Illumina NovaSeq 6000 S4 PE150) on Illumina Nextera libraries made from tissues preserved in Ethanol 95% or RNAlater and extracted by Nucleomag kits. The Atlantic Salmon (Bertolotti et al. 2020) and American Eel (Ulmo-Diaz et al. 2023) datasets are available on NCBI SRA. To limit the potential impacts of transatlantic population structure in the Atlantic Salmon dataset, we downloaded raw sequences from Norwegian samples only, which were sequenced either on Illumina HiSeq 2500 PE125 or HiSeq3000 PE100 (accession numbers in supplementary table S3, Supplementary Material online). Samples on either batch were distributed across the study range on the coast of Norway (supplementary fig. S3, Supplementary Material online).

All data were prepared following the pipeline described at https://github.com/enormandeau/wgs_sample_preparation. In brief, raw sequences were trimmed using *fastp* (Chen et al. 2018) and aligned on their respective reference genome (Table 1) with *bwa mem* (minimum alignment quality of 10). Note that the reference genome used for the Arctic Char data was assembled using sequencing data from a Dolly Varden (*Salvelinus malma*) or a *Salvelinus* sp. hybrid (Christensen et al. 2021). Duplicate reads were then removed using *picard*, indels were realigned, and overlapping ends of paired reads were clipped. The average per-base depth of coverage was then estimated using *samtools depth*. To mitigate batch effects in the Atlantic Salmon dataset, sequences obtained on HiSeq 2500 PE125 were randomly subsampled using *samtools view -s* with a factor of 0.64 to normalize both sequencer batches around an average coverage of 4.8× while maintaining variation in coverage between individuals. Three of these samples (Alta_12_0228, Arga_12_0089, Naus_12_0059) that had initial coverages over 12× were instead subsampled to a target coverage of 4.8×. The normalized (4.8×) Atlantic Salmon dataset was further subsampled to create datasets of decreasing average coverages (4×, 3×, 2×, and 1.5×). Reads from the American Eel data were also randomly subsampled with a factor of 0.5 to reach a similar depth as the Arctic Char and Lake Whitefish datasets.

Identification and Characterization of SNPs

For each dataset, SNPs were detected using ANGSD v0.931 (Korneliussen et al. 2014) with the GATK genotype likelihood framework (*-GL 2*) and the following parameters. Only reads in a properly mapped pair, with a sequencing quality over 20, and a mapping quality over 30 were considered for SNP calling. Biallelic SNPs were kept when sequenced with a coverage of at least 1× in 75% of all samples and with a MAF above 0.05 *(-doMaf 1 -minMaf 0.05*). Hardy–Weinberg equilibrium was assessed based on the global MAF *(-doHWE 1*) and individual read counts were extracted for each allele *(-doCounts 1 -dumpCounts 4*).

To categorize SNPs as either canonical or deviant, we used the calcLR function in *ngsParalog* (https://github.com/tplinderoth/ngsParalog) that compares the likelihood that reads at the position of SNPs come from either one or multiple copies of a loci. These hypotheses are tested assuming Hardy–Weinberg expectations for nonduplicated loci and using a genotype likelihood framework to account for

low-coverage data (Linderoth 2018). The likelihood ratios of both hypotheses were then compared to a χ^2 distribution (1 degree of freedom) and SNPs with a *P*-value (adjusted by applying the Benjamini–Hochberg procedure) under the conservative threshold of 0.001 were considered as deviant. We summed the number of reads for each SNP to compare depths of coverage between canonical and deviant SNPs.

To validate deviant SNPs, we used an alternative to ngsParalog, consisting of two tests adapted from the HDplot method (McKinney et al. 2017). We first assessed excesses of heterozygotes based on the ANGSD -doHWE 1 output ($P < 0.05$ and $F_{IS} < 0$), then computed the deviation from the expected allelic ratio in heterozygotes (1:1), following Karunarathne et al. (2022)'s implementation of HDplot. In brief, we used individual read ratios for the alternative allele in heterozygotes (genotype probability >0.8) to compute a Z-score for each SNP and compared it to a probability density function with a standard deviation of $\sqrt{n}$, where n is the number of heterozygotes. SNPs displaying an excess of heterozygotes or outside of the 0.025 and 0.975 quantiles for the probability density function were considered deviant.

Distribution of Deviant SNPs in the Genome

We ran RepeatMasker v.4.0.8 (Smit et al. 2013) on all four reference genomes to soft-mask TE and other repeats based on the combined DFam (Hubley et al. 2016) and Repbase (Bao et al. 2015) databases for teleost fishes. We masked 45.5% of the *Salvelinus* sp. genome, 52.0% of the Lake Whitefish genome, 49.5% of the Atlantic Salmon genome, and 12.7% of the American Eel genome (supplementary table S4, Supplementary Material online). We then defined "sufficiently covered" regions of the genome by measuring the average depth of coverage of every base pair in each dataset using *samtools depth* and delimiting all segments with an average coverage above 0.75× (i.e. the minimum coverage allowing SNP calling according to our ANGSD parameters). In nonoverlapping 1 Mb windows, we separately counted the number of canonical and deviant SNPs in soft-masked and unmasked sequences, then converted this number in a density of SNPs by dividing it by the total length of the sufficiently covered soft-masked or unmasked sequences in the window, respectively.

To estimate the remaining level of homology following the rediploidization of duplicated chromosomes in salmonids, we identified blocks of synteny, i.e. ohnolog pairs of genomic regions descending from the WGD. To do so, we hard-masked the repeats identified above and then aligned each salmonid reference genome used in this study on itself using *MUMmer v3.23* (Kurtz et al. 2004), implemented in *SyMap v4.2* (Soderlund et al. 2011). We found syntenic blocks (supplementary table S5, Supplementary Material online) concordant with those reported in the original reference for those assemblies (Lien et al. 2016; Christensen et al. 2018; Mérot et al. 2023). We then used *lastz v1.04.15* (Harris 2007) to realign each ohnolog pair (with arguments *–gfextend –chain –nogapped*), and averaged the percentage of identity in nonoverlapping windows of 1 Mb on all chromosomes. Only windows where the *lastz* anchors covered at least 1% of the sequence were kept.

To assess the impact of delayed rediploidization (increased homology) and repeated elements on the risk of mismapping reads and creating deviant SNPs, we created two negative binomial models per dataset, for canonical and deviant SNPs, respectively. The number of SNPs in either the soft-masked or unmasked fraction of the window was used as the response variable, and the length of the sufficiently covered soft-masked or unmasked fraction of the window was set as an offset to model the density of SNPs rather than the absolute count. The repetitive status (soft-masked or unmasked) and the average percentage of identity in the window were treated as fixed effects for the salmonid datasets. For the American Eel dataset, similar models were built with the repetitive status as the only fixed effect. The negative binomial model was selected after checking for overdispersion of the data in Poisson regressions (c-hat > 1). We checked the goodness-of-fit of the eight negative binomial models using hanging rootograms (supplementary fig. S4, Supplementary Material online).

Peaks of Coverage and TE

To better understand the interaction between deviant SNPs, sequencing depth, and repetitive elements in the reference genome of each species, we cataloged "peaks of coverage" where the average depth was elevated compared to the rest of the genome. We set the threshold at 1.5 times over the mode of depth in each dataset, corresponding to 3× for the Arctic Char, Lake Whitefish (James Bay), and American Eel datasets, and 8× for the Atlantic Salmon (4.8×) dataset. We measured the width of the peaks of coverage and then counted canonical and deviant SNPs occurring inside and outside peaks.

To test the hypothesis that peaks of coverage are enriched in repetitive DNA (which we hypothesized to be one possible cause of peaks of coverage), we used a χ^2 test to compare the proportion of base pairs covered by different clades of TEs and other repeats in the sufficiently covered portion of the genome and in peaks of coverage.

Impact of Deviant SNPs on Population Genomics Analyses

We compared the list of SNPs in the two Lake Whitefish datasets and their categorization as either canonical or deviant in each dataset. For SNPs common to both datasets, we calculated the Pearson correlation between the MAF

and F_{IS} in one dataset and the other. For the Arctic Char dataset, we constructed an ancestral reference genome to polarize alleles. This was done by aligning WGS data from four closely related species (Atlantic Salmon; Lake Trout, *Salvelinus namaycush*; Rainbow Trout, *Oncorhynchus mykiss*; and Chinook Salmon, *Oncorhynchus tshawytscha*; SRA accession number in supplementary table S6, Supplementary Material online) on the *Salvelinus sp.* reference genome (ASM291031v2) and using the most common allele as ancestral. We constructed 2dSFS and calculated pairwise F_{ST} between each population using the argument *-dosaf 1* in ANGSD and the *realsfs fst index*, *print*, and *stats* functions. We repeated this using first the complete list of SNPs, then only the canonical SNPs.

To estimate nucleotide genetic diversity in the Arctic Char and American Eel populations, we first masked deviant regions by defining 150 bp windows centered on each deviant SNP (see Fig. 4A for examples). Regions around isolated deviant SNPs, i.e. not within 150 bp of another deviant SNP, were not masked. We randomly selected four populations of Arctic Char with $n = 30$ and four subsamples of 30 individuals in the panmictic population of American Eel and generated SFS for each as described above while including invariant sites and rare SNPs. Using the *-doTheta* and *thetaStat* functions in ANGSD, we calculated Waterson's estimator (Θ_W), nucleotide diversity (Θ_π), and Tajima's D in windows of 100 Mb (window step of 20 kb) along all chromosomes. We repeated this process before and after masking for deviant regions. Since the four subsamples for the American Eel were very similar, we only show one in the results.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by Genome Canada and Genome Quebec as part of "FISHES: Fostering Small-scale Fisheries for Health, Economy and Food Security". Sampling was made possible by the collaboration of Fisheries and Ocean Canada (Les N Harris, Simon Willey, Ross Tallman, Xinhua Zhu, David Boguski); Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs (Québec; Julien Mainguy); Government of Nunavut; Makivik Corporation (Nunavik); Eeyou Marine Region Wildlife Board (Natacha Louttit, Stephanie Varty); Cree Trappers Association (Sanford Diamond, George Natawapineskum, John Lameboy); and numerous Inuit, Cree, and Dene communities and local fishermen across Canada. Thanks to Anne Beemelmanns, Charles Babin, Bérénice Bougas, Alysse Perreault-Payette, and Gabriel Piette-Lauzière for their help with laboratory work and coordination. We also want to thank Tyler Linderoth for making *ngsParalog* openly available and discussing its use with us, Anne-Marie Dion-Côté for her advice on repeated elements, and Alicia C. Bertolotti, Daniel Macqueen, their coauthors, and the Norwegian University of Life Sciences for providing the Atlantic Salmon data.

Author Contributions

X.D. led the design of the study and the writing of the manuscript, produced the Arctic Char data, developed analytical tools, and analyzed the data. R.B., P.H., and G.U.-D. produced data for the James Bay Whitefish (Bouchard), the Great Slave Lake Whitefish (Hénault), and the American Eel (Ulmo-Diaz), and all three contributed to the design of the study and the writing of the manuscript. E.N. and C.M. contributed some analytical tools, to the design of the study and the writing of the manuscript. L.B. and J.S.M. contributed to the design of the study and the writing of the manuscript.

Data Availability

All new genetic data have been deposited on Short Read Archive for Arctic Char (PRJNA1031558) and Lake Whitefish (PRJNA1037535, PRJNA1051576). Existing data for the Atlantic Salmon (Bertolotti et al. 2020) and American Eel (Ulmo-Diaz et al. 2023) are available as part of project PRJEB38061 and PRJNA964587, respectively. A bioinformatical pipeline presenting our suggested best practices for lcWGS including deviant SNP masking is available on GitHub at: https://github.com/xav9536/angsd_pipeline.

Literature Cited

Andolfatto P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res. 2007:17(12):1755–1762. https://doi.org/10.1101/gr.6691007.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016:17(2):81–92. https://doi.org/10.1038/nrg.2015.28.

Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015:6(1):11. https://doi.org/10.1186/s13100-015-0041-9.

Begun DJ, Holloway AK, Stevens K, Hillier LDW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 2007:5(11):e310. https://doi.org/10.1371/journal.pbio.0050310.

Benjamin A, Sağlam ?K, Mahardja B, Hobbs J, Hung TC, Finger AJ. Use of single nucleotide polymorphisms identifies backcrossing and species misidentifications among three San Francisco estuary osmerids. Conserv Genet. 2018:19(3):701–712. https://doi.org/10.1007/s10592-018-1048-9.

Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent MP, Røsæg LL, Holen MM, et al.

The structural variation landscape in 492 Atlantic salmon genomes. Nat Commun. 2020:11(1):5176. https://doi.org/10.1038/s41467-020-18972-x.

Cayuela H, Dorant Y, Forester BR, Jeffries DL, Mccaffery RM, Eby LA, Hossack BR, Gippet JMW, Pilliod DS, Chris Funk W. Genomic signatures of thermal adaptation are associated with clinal shifts of life history in a broadly distributed frog. J Anim Ecol. 2022:91(6):1222–1238. https://doi.org/10.1111/1365-2656.13545.

Cayuela H, Dorant Y, Mérot C, Laporte M, Normandeau E, Gagnon-Harvey S, Clément M, Sirois P, Bernatchez L. Thermal adaptation rather than demographic history drives genetic structure inferred by copy number variants in a marine fish. Mol Ecol. 2021:30(7):1624–1641. https://doi.org/10.1111/mec.15835.

Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 2015:7(2):567–580. https://doi.org/10.1093/gbe/evv005.

Chen N, Van Hout CV, Gottipati S, Clark AG. Using Mendelian inheritance to improve high-throughput SNP discovery. Genetics. 2014:198(3):847–857. https://doi.org/10.1534/genetics.114.169052.

Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Christensen KA, Rondeau EB, Minkley DR, Leong JS, Nugent CM, Danzmann RG, Ferguson MM, Stadnik A, Devlin RH, Muzzerall R, et al. The Arctic charr (*Salvelinus alpinus*) genome and transcriptome. PLoS One. 2018:13(9):1–30. https://doi.org/10.1371/journal.pone.0204076.

Christensen KA, Rondeau EB, Minkley DR, Leong JS, Nugent CM, Danzmann RG, Ferguson MM, Stadnik A, Devlin RH, Muzzerall R, et al. Retraction: the Arctic charr (Salvelinus alpinus) genome and transcriptome assembly. PLoS One. 2021:16(2):e0247083. https://doi.org/10.1371/journal.pone.0247083.

Cifuentes M, Grandont L, Moore G, Chèvre AM, Jenczewski E. Genetic regulation of meiosis in polyploid species: new insights into an old question. New Phytol. 2010:186(1):29–36. https://doi.org/10.1111/j.1469-8137.2009.03084.x.

Côté CL, Gagnaire PA, Bourret V, Verreault G, Castonguay M, Bernatchez L. Population genetics of the American eel (*Anguilla rostrata*): FST = 0 and North Atlantic Oscillation effects on demographic fluctuations of a panmictic species. Mol Ecol. 2013:22(7):1763–1776. https://doi.org/10.1111/mec.12142.

Dallaire X, Normandeau É, Mainguy J, Tremblay JÉ, Bernatchez L, Moore JS. Genomic data support management of anadromous Arctic Char fisheries in Nunavik by highlighting neutral and putatively adaptive genetic variation. Evol Appl. 2021:14(7):1880–1897. https://doi.org/10.1111/eva.13248.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features of RAD sequencing data: implications for genotyping. Mol Ecol. 2013:22(11):3151–3164. https://doi.org/10.1111/mec.12084.

Dorant Y, Cayuela H, Wellband K, Laporte M, Rougemont Q, Mérot C, Normandeau E, Rochette R, Bernatchez L. Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. Mol Ecol. 2020:29(24):4765–4782. https://doi.org/10.1111/mec.15565.

Dorant Y, Laporte M, Rougemont Q, Cayuela H, Rochette R, Bernatchez L. Landscape genomics of the American lobster (Homarus americanus). Mol Ecol. 2022:31(20):5182–5200. https://doi.org/10.1111/mec.16653.

Dou J, Zhao X, Fu X, Jiao W, Wang N, Zhang L, Hu X, Wang S, Bao Z. Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. Biol Direct. 2012:7(1):17. https://doi.org/10.1186/1745-6150-7-17.

Elmer KR. Genomic tools for new insights to variation, adaptation, and evolution in the salmonid fishes: a perspective for charr. Hydrobiologia. 2016:783(1):191–208. https://doi.org/10.1007/s10750-015-2614-5.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013:9(10):e1003905. https://doi.org/10.1371/journal.pgen.1003905.

Fijarczyk A, Babik W. Detecting balancing selection in genomes: limits and prospects. Mol Ecol. 2015:24(14):3529–3545. https://doi.org/10.1111/mec.13226.

Fraik AK, McMillan JR, Liermann M, Bennett T, McHenry ML, McKinney GJ, Wells AH, Winans G, Kelley JL, Pess GR, et al. The impacts of dam construction and removal on the genetics of recovering steelhead (*Oncorhynchus mykiss*) populations across the Elwha river watershed. Genes (Basel). 2021:12(1):89. https://doi.org/10.3390/genes12010089.

Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993:133(3):693–709. https://doi.org/10.1093/genetics/133.3.693.

Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. Mol Ecol. 2017:26(20):5369–5406. https://doi.org/10.1111/mec.14264.

Gao B, Shen D, Xue S, Chen C, Cui H, Song C. The contribution of transposable elements to size variations between four teleost genomes. Mob DNA. 2016:7(1):4. https://doi.org/10.1186/s13100-016-0059-7.

Glasauer SMK, Neuhauss SCF. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 2014:289(6):1045–1060. https://doi.org/10.1007/s00438-014-0889-2.

Gundappa MK, To TH, Grønvold L, Martin SAM, Lien S, Geist J, Hazlerigg D, Sandve SR, Macqueen DJ. Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. Mol Biol Evol. 2022:39(1):msab310. https://doi.org/10.1093/molbev/msab310.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009:5(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Harris RS. Improved pairwise alignment of genomic data [Doctoral dissertation]. [State College (PA)]: Pennsylvania State University; 2007.

Harvey MG, Judy CD, Seeholzer GF, Maley JM, Graves GR, Brumfield RT. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ. 2015:3:e895. https://doi.org/10.7717/peerj.895.

Hecht BC, Campbell NR, Holecek DE, Narum SR. Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. Mol Ecol. 2013:22(11):3061–3076. https://doi.org/10.1111/mec.12082.

Hemstrom WB, Freedman MG, Zalucki MP, Ramírez SR, Miller MR. Population genetics of a recent range expansion and subsequent loss of migration in monarch butterflies. Mol Ecol. 2022:31(17):4544–4557. https://doi.org/10.1111/mec.16592.

Hénault M, Marsit S, Charron G, Landry CR. The genomic landscape of transposable elements in yeast hybrids is shaped by structural variation and genotype-specific modulation of transposition rate. bioRxiv 539935. https://doi.org/10.1101/2023.05.08.539935, 9 May 2023, preprint: not peer reviewed.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016:44(D1):D81–D89. https://doi.org/10.1093/nar/gkv1272.

Hurles M. Are 100,000 'SNPs' useless? Science. 2002:298(5598): 1509–1509. https://doi.org/10.1126/science.298.5598.1509a.
Jaegle B, Pisupati R, Soto-Jiménez LM, Burns R, Rabanal FA, Nordborg M. Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. Genome Biol. 2023:24(1):44. https://doi.org/10.1186/s13059-023-02875-3.
Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004:431(7011): 946–957. https://doi.org/10.1038/nature03025.
Karunarathne P, Zhou Q, Schliep K, Milesi P. A comprehensive framework for detecting copy number variants from single nucleotide polymorphism data: 'rCNV', a versatile r package for paralogue and CNV detection. Mol Ecol Res. 2022:23(8):1772–1789. https://doi.org/10.1101/2022.10.14.512217.
Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 2014:15(1): 356. https://doi.org/10.1186/s12859-014-0356-4.
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004:5(2):R12. https://doi.org/10.1186/gb-2004-5-2-r12.
Laporte M, Le Luyer J, Rougeux C, Dion-Côté AM, Krick M, Bernatchez L. DNA methylation reprogramming, TE derepression, and postzygotic isolation of nascent animal species. Sci Adv. 2019:5(10): eaaw1644. https://doi.org/10.1126/sciadv.aaw1644.
Larson WA, Isermann DA, Feiner ZS. Incomplete bioinformatic filtering and inadequate age and growth analysis lead to an incorrect inference of harvested-induced changes. Evol Appl. 2021:14(2): 278–289. https://doi.org/10.1111/eva.13122.
Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016:533(7602): 200–205. https://doi.org/10.1038/nature17164.
Liljegren MM, de Muinck EJ, Trosvik P. Microsatellite length scoring by single molecule real time sequencing—effects of sequence structure and PCR regime. PLoS One. 2016:11(7):e0159232. https://doi.org/10.1371/journal.pone.0159232.
Linderoth T. Identifying population histories, adaptive genes, and genetic duplication from population-scale next generation sequencing. Berkeley: University of California; 2018.
Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol Ecol. 2021:30(23):5966–5993. https://doi.org/10.1111/mec.16077.
Macqueen DJ, Johnston IA. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B Biol Sci. 2014:281(1778):20132881. https://doi.org/10.1098/rspb.2013.2881.
Márquez R, Linderoth TP, Mejía-Vargas D, Nielsen R, Amézquita A, Kronforst MR. Divergence, gene flow, and the origin of leapfrog geographic distributions: the history of colour pattern variation in Phyllobates poison-dart frogs. Mol Ecol. 2020:29(19): 3702–3719. https://doi.org/10.1111/mec.15598.
Mason AS, Wendel JF. Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. Front Genet. 2020:11: 1014. https://doi.org/10.3389/fgene.2020.01014.
Matvienko M, Kozik A, Froenicke L, Lavelle D, Martineau B, Perroud B, Michelmore R. Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. PLOS ONE. 2013:8(2): e55913. https://doi.org/10.1371/journal.pone.0055913.
McKinney GJ, Seeb LW, Larson WA, Gomez-Uchida D, Limborg MT, Brieuc MSO, Everett MV, Naish KA, Waples RK, Seeb JE. An integrated linkage map reveals candidate genes underlying adaptive variation in Chinook salmon (*Oncorhynchus tshawytscha*). Mol Ecol Resour. 2016:16(3):769–783. https://doi.org/10.1111/1755-0998.12479.
McKinney GJ, Waples RK, Seeb LW, Seeb JE. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol Ecol Resour. 2017:17(4):656–669. https://doi.org/10.1111/1755-0998.12613.
Mérot C, Stenløkk KSR, Venney C, Laporte M, Moser M, Normandeau E, Árnyasi M, Kent M, Rougeux C, Flynn JM, et al. Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (Coregonus sp.) with long and short reads. Mol Ecol. 2023:32(6):1458–1477. https://doi.org/10.1111/mec.16468.
Minkley DR. Transposable elements in the salmonid genome [master's thesis]. [Victoria (BC), Canada]: University of Victoria; 2018.
Nadukkalam Ravindran P, Bentzen P, Bradbury IR, Beiko RG. PMERGE: computational filtering of paralogous sequences from RAD-seq data. Ecol Evol. 2018:8(14):7002–7013. https://doi.org/10.1002/ece3.4219.
Ohno S. Evolution by gene duplication. New York: Springer; 1970.
O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you'e looking for: principles of effective SNP filtering for molecular ecologists. Mol Ecol. 2018:27(16):3193–3206. https://doi.org/10.1111/mec.14792.
Peona V, Blom MPK, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jønsson KA, Zhou Q, et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. Mol Ecol Resour. 2021:21(1): 263–286. https://doi.org/10.1111/1755-0998.13252.
Pope NS, Singh A, Childers AK, Kapheim KM, Evans JD, López-Uribe MM. The expansion of agriculture has shaped the recent evolutionary history of a specialized squash pollinator. Proc Natl Acad Sci. 2023:120(15):e2208116120. https://doi.org/10.1073/pnas.2208116120.
Rellstab C, Dauphin B, Zoller S, Brodbeck S, Gugerli F. Using transcriptome sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus cembra*. Mol Ecol Resour. 2019:19(2):536–551. https://doi.org/10.1111/1755-0998.12986.
Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biol. 2017:18(1):1–14. https://doi.org/10.1186/s13059-017-1241-z.
Rougeux C, Bernatchez L, Gagnaire P-A. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish Species pairs (Coregonus clupeaformis). Genome Biol Evol. 2017:9(8):2057–2074. https://doi.org/10.1093/gbe/evx150.
Saglam IK, Prince DJ, Meek M, Ali OA, Miller MR, Peacock M, Neville H, Goodbla A, Mellison C, Somer W, et al. Genomic analysis reveals genetic distinctiveness of the Paiute Cutthroat Trout *Oncorhynchus clarkii seleniris*. Trans Am Fish Soc. 2017:146(6):1291–1302. https://doi.org/10.1080/00028487.2017.1356373.
Sánchez CC, Smith TPL, Wiedmann RT, Vallejo RL, Salem M, Yao J, Rexroad CE. Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics. 2009:10(1):559. https://doi.org/10.1186/1471-2164-10-559.
Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev

Genet. 2018:19(6):329–346. https://doi.org/10.1038/s41576-018-0003-4.
Shagina I, Bogdanova E, Mamedov IZ, Lebedev Y, Lukyanov S, Shagin D. Normalization of genomic DNA using duplex-specific nuclease. BioTechniques. 2010:48:455–459. https://doi.org/10.2144/000113422.
Simakov O, Marlétaz F, Yue JX, O'Connell B, Jenkins J, Brandt A, Calef R, Tung CH, Huang TK, Schmutz J, et al. Deeply conserved synteny resolves early events in vertebrate evolution. Nat Ecol Evol. 2020:4:820–830. https://doi.org/10.1038/s41559-020-1156-z.
Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. http://www.repeatmasker.org.
Smith SR, Normandeau E, Djambazian H, Nawarathna PM, Berube P, Muir AM, Ragoussis J, Penney CM, Scribner KT, Luikart G, et al. A chromosome-anchored genome assembly for Lake Trout (*Salvelinus namaycush*). Mol Ecol Resour. 2022:22(2):679–694. https://doi.org/10.1111/1755-0998.13483.
Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 2011:39(10):e68. https://doi.org/10.1093/nar/gkr123.
Sotero-Caio CG, Platt RN II, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 2017:9(1):161–177. https://doi.org/10.1093/gbe/evw264.
Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011:477(7363):207–210. https://doi.org/10.1038/nature10342.
Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989:123(3):585–595. https://doi.org/10.1093/genetics/123.3.585.
Therkildsen NO, Palumbi SR. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour. 2017:17(2):194–208. https://doi.org/10.1111/1755-0998.12593.
Todesco M, Owens GL, Bercovich N, Légaré JS, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EBM, Imerovski I, et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. Nature. 2020:584(7822):602–607. https://doi.org/10.1038/s41586-020-2467-6.
Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019:47(21):10994–11006. https://doi.org/10.1093/nar/gkz841.
Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012:13(1):36–46. https://doi.org/10.1038/nrg3117.
Ulmo-Diaz G, Engman A, McLarney WO, Lasso Alcalá CA, Hendrickson D, Bezault E, Feunteun E, Prats-Léon FL, Wiener J, Maxwell R, et al. Panmixia in the American eel extends to its tropical range of distribution: biological implications and policymaking challenges. Evol Appl. 2023:eva.13599. https://doi.org/10.1111/eva.13599.
Verdu CF, Guichoux E, Quevauvillers S, De Thier O, Laizet Y, Delcamp A, Gévaudant F, Monty A, Porté AJ, Lejeune P, et al. Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in *Robinia pseudoacacia* L. Ecol Evol. 2016:6(20):7323–7333. https://doi.org/10.1002/ece3.2466.
Wahlund S. Zusammensetzung Von Populationen Und Korrelationserscheinungen Vom Standpunkt Der Vererbungslehre Aus Betrachtet. Hereditas. 1928:11(1):65–106. https://doi.org/10.1111/j.1601-5223.1928.tb02483.x.
Waples RK, Seeb LW, Seeb JE. Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (Oncorhynchus keta). Mol Ecol Resour. 2016:16(1):17–28. https://doi.org/10.1111/1755-0998.12394.
Weiss H, Maluszynska J. Chromosomal rearrangement in autotetraploid plants of *Arabidopsis thaliana*. Hereditas. 2004:133(3):255–261. https://doi.org/10.1111/j.1601-5223.2000.00255.x.
Weissensteiner MH, Bunikis I, Catalán A, Francoijs KJ, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. Discovery and population genomics of structural variation in a songbird genus. Nat Commun. 2020:11(1):3403. https://doi.org/10.1038/s41467-020-17195-4.
Xuereb A, Rougemont Q, Dallaire X, Moore JS, Normandeau E, Bougas B, Perreault-Payette A, Koop BF, Withler R, Beacham T, et al. Re-evaluating Coho salmon (*Oncorhynchus kisutch*) conservation units in Canada using genomic data. Evol Appl. 2022:15(11):1925–1944. https://doi.org/10.1111/eva.13489.
Zhang K, Wang X, Cheng F. Plant polyploidy: origin, evolution, and its influence on crop domestication. Hortic Plant J. 2019:5(6):231–239. https://doi.org/10.1016/j.hpj.2019.11.003.

Associate editor: Andrea Betancourt