Investigating structural variant, indel and single nucleotide polymorphism differentiation between locally adapted Atlantic salmon populations

Abstract

Genomic structural variants (SVs) are now recognized as an integral component of intraspecific polymorphism and are known to contribute to evolutionary processes in various organisms. However, they are inherently difficult to detect and genotype from readily available short-read sequencing data, and therefore remain poorly documented in wild populations. Salmonid species displaying strong interpopulation variability in both life history traits and habitat characteristics, such as Atlantic salmon (*Salmo salar*), offer a prime context for studying adaptive polymorphism, but the contribution of SVs to fine-scale local adaptation has yet to be explored. Here, we performed a comparative analysis of SVs, single nucleotide polymorphisms (SNPs) and small indels (< 50 bp) segregating in the Romaine and Puyjalon salmon, two putatively locally adapted populations inhabiting neighboring rivers (Québec, Canada) and showing pronounced variation in life history traits, namely growth, fecundity, and age at maturity and at smoltification. We first catalogued polymorphism using a hybrid SV characterization approach pairing both short (16X) and long-read sequencing (20X) for variant discovery with graph-based genotyping of SVs across 60 salmon genomes, along with characterization of SNPs and small indels from short reads. We thus identified 115,907 SVs, 8,777,832 SNPs and 1,089,321 short indels, with SVs covering 4.8 times more base pairs than SNPs. All three variant types revealed a highly congruent population structure and similar patterns of F_{ST} and density variation along the genome. Finally, we performed outlier detection and redundancy analysis (RDA) to identify variants of interest in the putative local adaptation of Romaine and Puyjalon salmon. Genes located near these variants were enriched for biological processes related to nervous system function, suggesting that observed variation in traits such as age at smoltification could arise from differences in neural development. This study therefore demonstrates the feasibility of large-scale SV characterization and highlights its relevance for salmonid population genomics.

1 Introduction

Differences in DNA sequence and structure among individuals within species, referred to as genetic variation, serve as the basis for key evolutionary mechanisms such as speciation and local adaptation (Barrett & Schluter, 2008). Genetic variation can be described as a wide spectrum of variants of various sizes, ranging from single nucleotide polymorphisms (SNPs) to larger structural variants (SVs), which may span megabase-long stretches of DNA or even whole chromosomes (Feuk, Carson & Scherer, 2006; Wellenreuther & Bernatchez, 2018; Mérot et al., 2020). SVs such as insertions, deletions, duplications and inversions are now recognized as the main component of genetic variation, as they affect at least two to eight times more bases in genomes than SNPs (Catanach et al., 2019; Mérot et al., 2020; Hämälä et al., 2021). This estimate tends to increase as our ability to detect SVs from high-throughput sequencing data is constantly improving (Ho, Urban & Mills, 2020; Mérot et al., 2020).

SVs are also known to have a broad range of consequences at various biological levels. At the molecular scale, they may influence gene dosage, gene expression, DNA interactions and tridimensional structure by altering genetic elements' proximity and copy number (Feuk et al., 2006; Gamazon & Stranger, 2015; Spielmann, Lupiáñez & Mundlos, 2018). SVs that disrupt collinearity between homologous chromosomes, especially large inversions, are also likely to restrict or suppress recombination (Crown et al., 2018; Rowan et al., 2019). This may result in an apparent reduced gene flow around SVs, which may link co-adapted alleles, thus promoting the formation of supergenes that may underlie complex and adaptive phenotypes (Rieseberg, 2001; Kirkpatrick & Barton, 2006; Thompson & Jiggins, 2014).

A growing body of evidence also suggests that SVs can be involved in evolutionary mechanisms in various species (reviewed in Wellenreuther & Bernatchez, 2018). For instance, supergenes arising from large inversions have been linked to adaptive variation in wing color patterns in *Heliconius* butterfly (Joron et al., 2011), to migratory behavior in rainbow trout (*Oncorhynchus mykiss*; Pearse et al., 2014; Pearse et al., 2019) and Atlantic cod (*Gadus morhua*; Kirubakaran et al., 2016; Berg et al., 2017), as well as to reproductive strategies in the ruff (*Philomachus pugnax*; Küpper et al., 2015) and the white-throated sparrow (*Zonotrichia albicollis*; Tuttle, 2003). Other key examples of inversion polymorphism involved in ecotype divergence and local adaptation have been documented in the seaweed fly (*Coelopa frigida*;

Mérot et al., 2018) and in three-spined stickleback (*Gasterosteus aculeatus*; Jones et al., 2012). Besides large inversions, copy number variants (CNVs) have also been linked to adaptation to local temperature regimes in American lobster (*Homarus americanus*; Dorant et al., 2020) and to glacial lineage divergence in capelin (*Mallotus villosus*; Cayuela et al., 2021). Industrial melanism in peppered moths (*Biston betularia*), a textbook example of rapid adaptation to environmental change, has been associated with an intronic insertion in the *cortex* gene (Van't Hof et al., 2016). Similarly, a 2.25-kb intronic insertion would explain color pattern divergence among lineages in the *Corvus* genus, promoting reproductive isolation and thus leading to speciation (Weissensteiner et al., 2020).

Despite such well-documented cases of adaptive genomic rearrangements, most SVs other than large inversions remain understudied in a population genomics context. Relative to SNPs, very little is known about how such a large component of genetic variation is distributed within and between wild populations. Indeed, while standard procedures and pipelines are available for population-scale SNP calling, SV detection and genotyping, on the other hand, involve significant challenges for large, multisample datasets (Mahmoud et al., 2019; Ho, Urban & Mills, 2020).

Calling SVs requires specialized software due to their complexity and diversity in type and length. SV callers rely on various signals of discordance in read mapping relative to a reference genome in order to infer SVs in a given sample (Lin et al., 2015; Mahmoud et al., 2019). Because short-read sequencing is widely available and affordable, it is an appropriate technology for population-scale study of genetic variation. However, the performance of short-read-based SV callers is highly variable (Cameron, Di Stefano & Papenfuss, 2019). They are known to lack sensitivity, as true positive detection rates can be as low as 10% (Huddleston et al., 2017; Sedlazeck et al., 2018a). They also show low precision, with high false discovery rates reaching 89% for some datasets (Mills et al., 2011), especially for calls near SNPs, indels, low-complexity regions and repeats (Cameron, Di Stefano & Papenfuss, 2019). In fact, short reads are hard to map to the reference genome owing to their small length, especially when they include numerous sequencing errors, repeats (Sedlazeck et al., 2018b) or sequences differing considerably from the reference, such as SVs. Spurious mapping may result in underreporting of variation, an issue known as reference allele bias (Nielsen et al., 2011; Brandt et al., 2015). Calling SVs in a given dataset using multiple callers (ensemble calling), may increase the range of SV types and sizes detected or reduce false

discovery rate compared to single-tool SV calling. For instance, SV callsets may be merged across callers, then filtered for calls supported by a given minimum number of tools (Auton et al., 2015). However, the improvement in sensitivity and/or precision strongly depends on the callers used in combination (Kosugi et al., 2019; Mahmoud et al., 2019).

By contrast, recent advances in third generation sequencing platforms (Oxford Nanopore and Pacific Biosciences' technologies) have brought significant improvements regarding SV calling. Long reads span kilobase-long segments of DNA, thus fully overlapping SVs and their breakpoints, which considerably facilitates read mapping (Sedlazeck et al., 2018b) and improves sensitivity of SV detection (Mahmoud et al., 2019), especially for novel insertions (Ho, Urban & Mills, 2020). Specialized algorithms and pipelines have been developed to process long reads and account for their length and higher sequencing error rate (Rang, Kloosterman & de Ridder, 2018; Delahaye et al., 2021). However, high costs prevent long-read sequencing from becoming a routine tool for population-scale SV studies for species with large genomes such as salmonid fishes (usually around 3 Gb), which requires the sequencing of many genomes, namely for accurate estimates of allele frequency.

To provide an adequate balance between accurate SV characterization and genotyping in large datasets, emerging hybrid approaches can be considered, such as pairing affordable short-read sequencing for all samples with high performance third generation sequencing for a small subset of genomes only. Candidate SVs called from long reads can then be genotyped in all samples from short-read data using pangenome graphs, which offer considerable advantages over conventional, linear reference-based methods. Indeed, in a reference pangenome, the reference genome is represented as a base graph structure where known variants and alternate alleles are encoded as alternate paths, i.e., series of nodes and links (Paten et al., 2017). The integration of known genetic variation within the reference greatly facilitates mapping of reads that overlap such variants, thus improving both SV detection and genotyping, and reducing reference allele bias (Ameur, 2019). This approach has shown promising results for genome-wide population-scale SV detection in human (*Homo sapiens*; Yan et al., 2021), soybean (*Glycine max*; Lemay et al., 2022), lake whitefish (*Coregonus clupeaformis*; Mérot et al., 2023) and in kākāpō parrots (*Strigops habroptilus*; Wold et al., 2023).

Knowledge pertaining to SVs remains minimal in salmonid fishes, despite their genomes being extensively studied for aquaculture applications. The first comprehensive catalog of genome-wide SVs for Atlantic salmon (*Salmo salar*) was produced by Bertolotti et al. (2020) by calling putative SVs using short-read-based caller LUMPY (Layer et al., 2014) in 492 wild and domestic salmon from various populations in Europe and North America. SV calls were then manually curated with SV-plaudit (Belyeu et al., 2018) in order to eliminate false positives, yielding 15,483 high confidence SVs matching the expected population structure. This study also revealed a subset of outlier SVs overlapping genes enriched for brain expression, suggesting an implication in salmon domestication. Other population-scale SV catalogs were published for the rainbow trout (*Oncorhynchus mykiss*; Liu et al., 2021), and two sympatric sister species of lake whitefish (*Coregonus* sp.; Mérot et al., 2023).

Further work is required in order to fully appreciate SVs' relevance in the genomics and biology of Atlantic salmon, which could serve as an ideal candidate species for studying adaptive SVs and developing an efficient population-scale SV discovery pipeline. SVs and larger chromosomal rearrangements are likely a key feature of salmonid genomes, as they are critical to the reploidization process following the salmonid-specific fourth vertebrate whole-genome duplication that occurred at least 60 million years ago (Ss4R) (Allendorf & Thorgaard, 1984; Crête-Lafrenière, Weir & Bernatchez, 2012; Lien et al., 2016). Sequence repeats, which account for 50 to 60% of the Atlantic salmon genome (de Boer et al., 2007), are also known to promote SV formation (Levy-Sakin et al., 2019). Moreover, Atlantic salmon display considerable life history trait variation both within and between wild populations (Klemetsen et al., 2003). Consequently, there is considerable interest in understanding the genetic architecture of traits such as growth rate and disease resistance for aquaculture (Gjedrem & Rye, 2018), but also in the context of local adaptation, which usually involves such life history trait variation (Taylor, 1991; Lu & Bernatchez, 1999; Fraser & Bernatchez, 2005). Local adaptation is expected to be a major driver of population structure in Atlantic salmon, given its homing behavior (Allendorf & Waples, 1996) and the variability in habitat conditions (Taylor, 1991; Kawecki & Ebert, 2004). Indeed, the association between the genetic structure of seven groups of local salmon populations in Eastern Canada and regional rivers' environmental parameters suggests adaptive SNP divergence among these groups (Dionne et al., 2008; Bourret et al., 2013). Previous studies have also highlighted a few adaptive large

chromosomal rearrangements in wild Atlantic salmon populations (Wellband et al., 2019; Watson et al., 2022), as well as divergent SVs between domestic and wild populations (see Bertolotti et al., 2020). However, SVs' contribution to local adaptation remains poorly documented among North American populations, especially at a finer geographic scale (e.g., within neighboring rivers).

Two parapatric Atlantic salmon populations from the Romaine and Puyjalon rivers (Québec, Canada; 50.306337, -63.795602; Figure 1B) represent a prime case of putative fine-scale local adaptation. Indeed, admixture analysis and fixation index calculation (F_{ST} = 0.036) based on microsatellite markers showed moderate differentiation between Romaine (RO) and Puyjalon (PU) salmon, despite their geographical proximity and habitat connectivity (Albert & Bernatchez, 2006). Furthermore, they exhibit different trade-offs in major life history traits: earlier age at smoltification and sexual maturity have been reported among wild Romaine salmon (Fontaine et al., 2000; Belles-Isles et al., 2004; WSP Global, 2019), as well as in wild-born Romaine salmon reared in a hatchery environment at the LARSA (Laboratoire de Recherche en Sciences Aquatiques; Université Laval, Québec), whereas wild-born Puyjalon salmon have shown higher growth rates over several cohorts in the same hatchery conditions (Therrien et al., 2017; Langlois-Parisé et al., 2018; T. Dion et al., 2020; T. Dion, Langlois-Parisé & Proulx, 2020). The persistence of such life history trait variation among cohorts in both wild and controlled environments strongly suggests heritable genetic variation likely linked to local adaptation, as the Romaine and Puyjalon rivers differ in spawning habitat quality, substrate and hydrological parameters (Schieffer, 1975; Fontaine et al., 2000; GENIVAR, 2002; Belles-Isles et al., 2004; WSP Global, 2019). However, the genetic basis of this putative local adaptation has yet to be investigated.

Here, we address this lack of knowledge by proposing a multiplatform, graph-based SV discovery pipeline across numerous genomes (Figure 1A) in order to catalog genetic polymorphism in Romaine and Puyjalon salmon, allowing us to investigate candidate adaptive variation within these populations. With this approach, we primarily targeted small (50 - 1000 bp) to intermediate-sized SVs (< 5kb), as direct SV calling based on short reads and long reads is more accurate and powerful in this range of length (Mahmoud et al., 2019). This study thus served as an unprecedented opportunity to characterize SVs, SNPs and small indels in North American Atlantic salmon, as well as to explore the relative contribution of various forms of genetic variation to fine-scale adaptation and population differentiation.

2 Materials and Methods

2.1 Sampling, DNA extraction and sequencing

2.1.1 Short reads

Manipulations involving fish were authorized by the Comité de protection des animaux de l'Université Laval (permit number: 2021-783). Adipose fin clips were sampled from 60 wild-born adult salmon raised as broodstock at Université Laval's Laboratoire de Recherche en Sciences Aquatiques (LARSA) and stored in ethanol until use. The samples comprised 31 Puyjalon (16 males and 15 females) and 29 Romaine (14 males and 15 females) individuals.

Spin column DNA extractions were performed using Qiagen's DNeasy blood and tissue kit according to the manufacturer's protocol, with the exception of the elution step, which was done twice per sample with 50 µl of water. DNA quality was assessed by concentration measurement and migration on 1% agarose gel. DNA samples were then diluted to 10 ng/µl and sent to Génome Québec's Centre d'expertise et de services (Montréal, Canada) for library preparation and whole genome sequencing on an Illumina NovaSeq6000, using four S4 PE150 lanes for an anticipated depth of 16X per sample.

2.1.2 Long reads

Among the 60 fish sampled for whole genome short-read sequencing, four (one male and one female for each population) were used for Nanopore long-read sequencing. In order to provide intact high molecular weight DNA, whole blood was extracted from live fish using EDTA-prefilled syringes, followed by humane euthanasia by decapitation. Blood samples were flash-frozen in liquid nitrogen, transferred to storage tubes and stored at -80°C until use.

High molecular weight DNA extraction was performed twice for each fish using Circulomics' CBB protocol for nucleated blood (EXT-NBH-001; Circulomics, 2021), from 6 µl of blood mixed with 194 µl of ice-cold PBS. DNA quality was assessed by measuring concentration with Qubit and migrating DNA on a 0.5% agarose gel. DNA samples were then sent to the Centre for Integrative Genomics (CIGENE) at the Norwegian University of Life Science (NMBU) for sequencing. DNA fragments shorter than 25 kb were removed by size selection with Circulomics' Short Reads Eliminator kit, and seven libraries were prepared for each sample using the SQK-LSK110 kit (Oxford Nanopore Technologies).

Sequencing was performed on a PromethION24 in short serial runs following protocol NFL_9076_v109_revA. Each sequencing run was terminated after a few hours, when the number of active pores dropped to below 10%, in order to recover pores by nuclease-flushing flow cells, which were then refilled with the same DNA preparation for a next short sequencing run. Two FLO-PRO002 flow cells were used for each sample, which were each filled with six and five loadings, respectively, in order to obtain an approximate coverage of 20X. Basecalling was done with Guppy version 5.0.13 (high-accuracy basecalling model) and raw reads were filtered for a minimum qscore of nine. The average yield for the four samples was 47.1 Gb of DNA, while the mean N50 was 39.5 kb.

2.2 Characterization of genetic variation

2.2.1 Raw sequencing data preprocessing

2.2.1.1 Short reads

The Ssal_Brian_v1.0 assembly, derived from a North American wild salmon from Newfoundland (Norwegian University of Life Sciences, 2022; GenBank assembly accession: GCA_923944775.1; project accession: CAKLZZ000000000.1), was used as the reference genome for all downstream analyses. This genome features 28 chromosomes with two known polymorphic rearrangements, i.e., the translocation of chromosome ssa01's p arm (ssa01p) to ssa23 (ssa01-23) (Lehnert et al., 2019), and the fusion of chromosomes ssa26 and ssa28 (Brenna-Hansen et al., 2012).

Raw Illumina data was processed using the wgs_sample_preparation pipeline (https://github.com/enormandeau/wgs_sample_preparation). Adapters and low-quality ends were first trimmed from raw reads by running fastp 0.20.0 (Chen et al., 2018) with default parameters. Trimmed reads were then mapped to the indexed reference genome (samtools *faidx* command, version 1.8; Danecek et al., 2021) using BWA MEM (Li, 2013), allowing a minimum mapping quality of 10 (-q 10). Duplicate reads were filtered out of the alignment with *MarkDuplicates* (Picard 1.119; Broad Institute, 2019). After indexing the resulting bam files with Picard *BuildBamIndex*, mapping was refined around candidate indels using GATK 3.6-0 *RealignerTargetCreator* and *IndelRealigner* (McKenna et al., 2010) and overlapping read pairs were clipped to preserve read regions with the highest average quality using bamUtil 1.0.14 *clip overlap* (Jun et al., 2015). Finally, we used samtools *addreplacerg* to add unique read group names for each sample's bam file, which is a requirement for some variant calling tools we used.

2.2.1.2 Long reads

Since each sequencing run produced multiple raw read files, all fastq files obtained for a given sample were first concatenated to yield a single fastq file per sample. Raw reads were filtered for an average minimum quality of 10 and a minimum read length of 1,000 bp using NanoFilt 2.0.8 (DeCoster et al., 2018). We mapped filtered reads to the Ssal_Brian_v1.0 assembly with Winnowmap version 2.03 (Jain et al., 2020; Jain et al., 2022) using default parameters and a *k*-mer size of 15 (-k 15). The complete preprocessing pipeline (ONT_data_processing v1.0.0) can be found at https://github.com/LaurieLecomte/ONT_data_processing.

2.2.2 SNP and short indel (1-50 bp) calling

SNPs and small indels were called exclusively from short-read data, as higher basecalling error rates in long-read data are likely to interfere with SNP detection (Rang, Kloosterman & de Ridder, 2018; Ahsan et al., 2021). Variant calling was performed in all 60 samples at once and for each chromosome separately, using bcftools *mpileup* and *call* (version 1.16) and requiring a minimum mapping quality of five at a given site (-q 5). The 28 single chromosome VCF files were then concatenated with bcftools *concat*.

In order to apply the same filtering criteria as SVs (described below), samples without at least four supporting reads and a minimum genotype quality of five for a given variant, or that had more than five times the anticipated whole genome short-read sequencing coverage (80) or an exceedingly high genotype quality (GQ = 127), were assigned the genotype "missing" ("./.") using bcftools *+set-GT* (version 1.15). Finally, we kept SNPs and small indels that had a minor allele frequency between 0.05 and 0.95 and that were genotyped in at least 50% of samples (i.e., population-scale filters), using bcftools *filter* (version 1.13). The full SNP and indel calling pipeline is available at https://github.com/LaurieLecomte/SNPs_indels_SR (version v1.0.0).

2.2.3 Structural variant calling

2.2.3.1 Short reads

In order to alleviate some of the challenges inherent to SV detection (e.g., low precision), we proposed an ensemble approach where SVs were first called independently with three separate tools, then merged across tools in order to obtain a union callset, which we filtered for calls supported by at least

two callers. We assumed that SVs confidently called by multiple tools are more likely to be true positives than SVs called by a single tool.

The three callers used in combination were chosen based on reported performance in previous studies and benchmarks (Cameron, Di Stefano & Papenfuss, 2019; Kosugi et al., 2019; Mérot et al., 2023; Stenløkk, 2023). Each caller was provided with the same 60 bam files, as well as the reference genome used for mapping reads. We first ran DELLY (version 1.1.6; Rausch et al., 2012) following guidelines for germline calling in high coverage genomes (https://github.com/dellytools/delly#germline-sv-calling). Putative SVs were first called separately in all samples and in each of the 28 chromosomes, then merged together in order to obtain a list of known SV sites to be genotyped by DELLY in each sample. Genotyped SVs were then merged across all samples into a unified, multisample VCF. We filtered for deletions (DEL), insertions (INS), duplications (DUP) and inversions (INV) labelled as PASS and PRECISE using bcftools 1.13 *filter*. Next, we used Manta version 1.6.0 (Chen et al., 2015) according to instructions for germline joint samples analysis (https://github.com/Illumina/manta/blob/master/docs/userGuide/README.md#germline-configuration-examples). We parallelized SV calling across the 28 chromosomes instead of across samples, since Manta has no built-in procedure for merging calls across samples. SVs tagged as BND (breakends) were converted into explicit inversions using the script convertInversion.py provided in Manta's installation directory. The 28 chromosome-specific VCFs were then concatenated into a single multisample file, which was filtered for PASS and PRECISE calls as well. The last short-read-based caller included in the pipeline was LUMPY (Layer et al., 2014), through smoove (version 0.2.7; Pedersen, Layer & Quinlan, 2020). Following recommendations for population-level calling (https://github.com/brentp/smoove#population-calling), we called SVs in the same manner as with DELLY. Only DEL, DUP, INV calls labelled as PRECISE were retained.

The three SV sets were then merged together using Jasmine version 1.1.5 (Kirsche et al., 2023), which integrates various information including chromosome, position, end, size and type to determine whether SV calls from different files or samples refer to the same SV or not. We ran Jasmine with parameters "--mutual_distance --max_dist_linear=0.25", so that the maximum allowed distance required between two SVs for them to be merged is correlated with their size. The merged VCF was then edited

with a custom R script (R version 4.1.2; R Core Team, 2021) in order to convert symbolic alternate alleles to explicit sequences and to standardize VCF fields. The formatted merged VCF was finally filtered for calls supported by at least two callers, with bcftools *filter*. Moreover, in accordance with the most prevalent definition of SVs (Feuk, Carson & Scherer, 2006), variants smaller than 50 bp were considered as small indels instead of SVs and were therefore filtered out. This first set of merged SVs will be referred to as the short-read SV set (SR SVs). The detailed short-read SV calling pipeline can be found at https://github.com/LaurieLecomte/SVs_short_reads (version v1.0.0).

2.2.3.2 Long reads

The SV calling procedure for Oxford Nanopore data is equivalent to the pipeline described above for SV detection from short reads, i.e., independent SV detection with three different tools, merging of SV calls across callers, and filtering for calls supported by at least two tools. However, since most long-read-based SV callers do not support multisample calling, SVs were first called separately for each sample using all three chosen tools, then merged across samples (across-sample merge) to obtain a single VCF per caller. The three multisample VCFs were then merged together (across-caller merge) to obtain the long-read SV set (LR SVs). The pipeline is available at https://github.com/LaurieLecomte/SVs_long_reads (version v1.0.0).

We ran Sniffles 2.0.7 (Sedlazeck et al., 2018a; Smolka et al., 2022) (default settings and "--output-rnames --combine-consensus" options) on each sample and filtered for PASS and PRECISE calls. We then refined alternate allele sequences and breakpoints for insertions, deletions and some duplications by running Iris (Kirsche et al., 2023): we first preprocessed each sample's VCF with Jasmine ("--dup_to_ins --preprocess_only"), and ran Iris with parameters "--keep_long_variants --also_deletions". The four samples' refined VCFs were merged together using Jasmine --ignore_strand --mutual_distance --allow_intrasample --output_genotypes, and refined SVs were then converted back to their original type with Jasmine --dup_to_ins --postprocess_only. The multisample Sniffles VCF was finally filtered again for PASS and PRECISE insertions, deletions, duplications and inversions. SVs were also called with SVIM 2.0.0 (Heller & Vingron, 2019) using parameters "--insertion_sequences --read_names --max_consensus_length=50000 --interspersed_duplications_as_insertions", following the same procedure as Sniffles to produce a multisample VCF with PASS calls. Last, we used NanoVar 1.4.1 (Tham et al.,

2020) with default settings. Supporting reads' names were added manually using a custom R script in order to allow for the refinement of SV breakpoints by Iris. The three multisample VCFs were finally merged together with Jasmine, formatted with custom R scripts and filtered, as described for the SR SV set.

2.2.3.3 Combination of SV datasets

A final merging step was performed for combining the short-read and long-read SV sets using Jasmine with parameters "--ignore_strand --ignore_merged_inputs --normalize_type --output_genotypes", resulting in a large union set of candidate SVs to be genotyped in the 60 Atlantic salmon genomes (https://github.com/LaurieLecomte/merge_SVs_SRLR; version v1.0.0).

2.2.4 SV genotyping

We implemented a graph-based genotyping pipeline (https://github.com/LaurieLecomte/genotype_SVs_SRLR, version v1.0.0) using the vg toolkit version 1.46.0 (Hickey et al., 2020), following recommendations from https://github.com/vgteam/vg/wiki/SV-genotyping-with-vg#sv-genotyping-with-vg-call. We first built an indexed reference graph structure from the reference genome fasta and the SV VCF file using *vg autoindex*, then computed snarls, i.e., sites of known variation in the genome graph, with the *vg snarls* command. We then remapped short reads to the variant-aware reference graph for all samples separately using *vg giraffe* (Sirén et al., 2021), computed read support for variation sites (*vg pack*), then genotyped these sites (*vg call*). We used bcftools *+set-GT* on each sample's VCF to set the genotype as missing ("./.") for calls that were not supported by at least four reads and that had a quality score lower than five, or that had an extreme quality score (GQ = 256) or an extreme depth (DP = 80), as such calls tend to be false positives (Cameron, Di Stefano & Papenfuss, 2019). All 60 sample VCFs were merged together with bcftools *merge*. The genotyped SV set was finally filtered for variants with a minor allele frequency between 0.05 and 0.95, and less than 50% missing data. The 50% missingness threshold was arbitrary and based on Mérot et al. (2023). Comparison with both less and more stringent missing data proportion thresholds showed that the choice of threshold did not impact the post-filtering variant count differently for SVs than for SNPs or indels (Suppl. Table 1).

In an effort to better link the genomic context and the confidence in SV calling, we compared the frequency distributions of both high-quality SVs that passed all filtering steps and low-quality, filtered out

SVs in two genomic features known to interfere with SV calling: highly similar regions resulting from whole-genome duplication events (e.g., syntenic regions) and repeated content (repeats and transposable elements). To identify syntenic regions, we followed the steps described by Dallaire et al. (2023): in summary, we aligned the genome to itself with nucmer (built-in mapper in MUMmer version 4.0.0; Marçais et al., 2018), then performed synteny analysis with SyMAP (Soderlund, Bomhoff & Nelson, 2011), et re-mapped syntenic blocks to the genome with LASTZ (version 1.04.15, Harris, 2007) to get the homology percentage. We identified repeats and transposable elements using RepeatMasker (version 4.0.8). To distinguish between probable false positive SVs and probable true positive SVs, we relied on the filtering criteria we applied on the sample level (read depth and genotype quality) and on the population level (on minor allele frequency and missing data proportion) during the genotyping procedure. Using bedtools *window*, we then extracted excluded and filtered SVs overlapping with either a syntenic region or a repeat region (within a 100-bp window), or both.

In addition, because information on putative SVs is lost at the genotyping step, we applied the procedure described in Supplementary Methods 1 to match genotyped SVs with a known putative SV based on the correspondence of position and allele length, in order to retrieve information on variant type, length and platform support (e.g., short- and/or long-read). This information allowed us to perform additional analyses on the set of genotyped and matched SVs. Indeed, in order to see if long-read SVs could be reliably genotyped from short-read data and a pangenome, we examined the concordance between the genotypes called by vg for long-read SVs and the genotypes called by long-read-based SV callers prior to merging datasets across platforms. First, for the four samples sequenced with both short- and long-read platforms, we extracted the three genotypes (from Sniffles, SVIM and NanoVar) for each long-read SV. We then determined the consensus genotype when possible, e.g., if at least two callers called the same genotype in a given sample. Each SV was labelled as concordant when its consensus genotype matched the corresponding genotype outputted by vg, or as non-concordant if its consensus genotype differed from the vg genotype. Alternatively, when all three callers outputted different genotypes for a given SV in a given sample, no consensus genotype could be inferred, and therefore the concordance between callers and vg could not be determined. We also performed this procedure for

short-read SVs for comparison purposes. This concordance analysis is detailed in the scripts compare_GTs_LR_vs_vg.sh and compare_GTs_LR_vs_vg.R from the genotype_SVs_SRLR pipeline.

2.3 Population genomics analyses

2.3.1 Differentiation between the Romaine and Puyjalon populations

We used ANGSD version 0.937 (Korneliussen, Albrechtsen & Nielsen, 2014) for performing various population and comparative genomics analyses on SVs, SNPs and small indels separately by adapting a previous pipeline designed for SNPs (https://github.com/clairemerot/angsd_pipeline). To investigate population structure, we first performed principal component analysis (PCA) on a normalized covariance matrix produced from input VCF files using VCFtools (version 0.1.16; Danecek et al., 2011), pcangsd (Meisner & Albrechtsen, 2018) and custom R scripts. From input VCF files, we then estimated average genome-wide fixation index (F_{ST}; Weir & Cockerham, 1984) from each population's allele frequency spectrum using ANGSD's -doSaf and realSFS functions (version 0.937). We also computed F_{ST} along the genome by sliding windows of 100 kb (per-window F_{ST}), as well as for each variant (per-variant F_{ST}).

We employed two complementary approaches for identifying candidate variants likely involved in local adaptation (Figure 1C). We first extracted the most highly differentiated variants falling within the upper 97% per-variant F_{ST} quantile (F_{ST} outliers). We also performed Fisher's exact tests on per-population allelic counts at each site and identified outliers with a corrected *p*-value (*q*-value) lower than 0.01 (Benjamini & Hochberg, 1995). We then extracted common outliers between F_{ST} and Fisher exact tests to yield a set of strongly differentiated variants used for further analysis.

Second, we ran a redundancy analysis (RDA) on the imputed genotype matrix, with the population as the only explanatory variable using the R package vegan (Oksanen et al., 2022). While F_{ST} and Fisher's exact test are more likely to detect outlier loci of large effect, RDA allows identifying covarying markers with individually weak effect that may be involved in polygenic control of phenotypic expression (Rellstab et al., 2015; Forester et al., 2018), as previously documented for life history traits such as age at sexual maturity or growth rate (Sinclair-Water et al., 2020; Debes et al., 2021). We defined RDA candidates as variants with loadings falling over the three standard deviations threshold (Forester, Laporte

& Manel, 2018). We thus obtained a set of outlier variants and a set of candidate variants for each of the three variant types studied.

2.3.2 Functional analysis of candidate genomic variation

In order to assess the potential functional impact of candidate variants on life-history trait variation observed in Romaine salmon, we investigated the overlap between variants of interest and known genes. We first annotated the Ssal_Brian_v1.0 assembly using the pipeline GAWN v0.3.5 (https://github.com/enormandeau/gawn) based on the transcriptome of the Ssal_v3.1 assembly (GenBank assembly accession: GCF_905237065.1) and filtered out possible duplicate annotations, which produced a list of 36,697 known genes.

For each set of variants of interest, we ran bcftools *window* (version 2.30.0; Quinlan et al., 2010) to identify a set of overlapping genes located within 10 kb of at least one variant. We then performed Gene Ontology (GO) enrichment analysis on each gene set with goatools 1.2.3 (Klopfenstein et al., 2018), using the list of 36,697 of genes from GAWN as the background (population) set and the go-basic database version 1.2 (2022-07-01; http://release.geneontology.org/2022-07-01/ontology/go-basic.obo). Only enriched terms that referred to a biological process (BP) and with a corrected *p*-value (Benjamini & Hochberg, 1995) under 0.1 were considered significant and preserved. We then used REVIGO (Supek et al., 2011) to cluster significant GO terms by semantic similarity with a cutoff value of 0.5 ("small list"), for easier interpretation. All scripts used for population genomics analyses can be found at https://github.com/LaurieLecomte/SVs_SNPs_indels_compgen (version v1.0.0).

3 Results

3.1 Long reads revealed more variants while short reads allowed population-scale SV genotyping

SVs were identified through our multistep calling procedure involving both short- and long-read data, where different callers and datasets showed high variability in the number, types and sizes of SVs detected. Indeed, short reads revealed mostly deletions, whereas long reads allowed the detection of many more SVs, especially deletions, insertions and duplications. Among short-read-based callers, Manta reported the most SVs of various types and sizes (151,103), while smoove called the least (28,164),

almost exclusively deletions (Table 1 & Suppl. Fig. 1). In total, 238,492 SV calls were merged across the three callers, of which only 15.5% (37,041) were shared by at least two tools and were longer than 50 bp: this short-read SV set primarily consisted of deletions (34,761) smaller than 100 bp, and very few duplications (318) and insertions (849) (Table 2 & Suppl. Fig. 2).

For the long-read pipeline, SVIM called over 3.5 million SVs, mostly insertions and deletions, more than the two other callers combined (Table 3 & Suppl. Fig. 3). The NanoVar callset was the smallest (454,697 SVs) but featured the largest proportions of duplications (31.8%) and inversions (16.4%) (Table 3). The proportion of all merged long-read SV calls (3,832,032) supported by multiple tools was 12.5%, smaller than for the merged short-read SV set. The 345,695 long-read SVs supported by at least two tools and longer than 50 bp retained for subsequent steps were mostly deletions (57.7%) and insertions (40.0%) and included only 208 inversions (Table 4 & Suppl. Fig. 4).

Merging both short- and long-read SVs yielded a set of 361,107 putative SVs, mainly deletions (59.1%) and insertions (38.3%) (Table 5 & Figure 2). The vast majority of these merged SVs, i.e., 89.7%, were exclusively called from long reads, including almost all (99.4%) of the 138,404 insertions identified (Suppl. Table 2). Similarly, 96.0% of duplications and 83.7% of deletions were uniquely detected from long reads. By contrast, only 15,412 SVs, or 4.3% of the merged SV set (Table 5), were unique to the short-read SV callset, including nonetheless 83.3% of all inversions identified (Suppl. Table 2). Moreover, the 21,629 SVs called from both short and long reads accounted for 5.9% of all merged SVs (Table 5), meaning that 58% of all short-read SVs were also supported by long-read data.

The merged SVs were represented in a variant-aware genome graph on which we mapped short-read data to genotype SVs in all 60 individuals. On average, 333,031 SVs were genotyped in each sample using the graph-based pipeline, for a total of 344,468 distinct SVs. As expected owing to the complex nature of SVs, the average proportion of missing data per site was over four times larger than for raw SNPs (Suppl. Table 3). 40.8% of raw genotyped SVs did not meet the minimum coverage required in at least half of samples, while two thirds had a minor allele frequency under 5% (Suppl. Table 4). Consequently, filtering on both proportion of missing data and minor allele frequency considerably reduced the SV set to 115,907 high-confidence variants (Table 6), or about 33.6% of the 344,468 raw genotyped SVs. These 115,907 SVs were used for subsequent population genomics analyses.

3.2 SVs encompassed a more extensive fraction of the genome than SNPs and small indels

While SNPs were the most frequent type of variants, SVs impacted a much higher proportion of genome base pairs, with large heterogeneity along the genome. Indeed, in addition to the final 115,907 SVs, we identified 8,777,832 SNPs and 1,089,321 short indels (Table 6) that met the same filters on minor allele frequency and proportion of missing data. SVs added up to over 41.8 Mb (including insertions), or about 1.68% of total genome length, which was approximately 4.8 times more than SNPs (total length: 8.7 Mb; Table 6) and four times more than small indels (total length: 10.3 Mb; Table 6). Similarly, the proportion of base pairs covered by a given variant type per 100-kb window was much more variable for SVs, reaching as much as 94.0% for some regions (e.g., around a 94.1-kb deletion on chromosome ssa09), whereas the maximum observed proportion of base pairs covered by SNPs was only 2.9% and 10.2% for indels (Table 6 & Figure 3).

If we consider the occurrence of variants instead of the number of variable bases, there was no considerable difference in variant density along the genome, as SVs, SNPs and small indels all tended to be more frequent towards the extremities of chromosomes (Suppl. Fig. 5). The gap observed in the number of SVs by 100-kb window (SV density) on chromosome ssa10 could be attributable to a large 2.5-Mb deletion called from short reads, but not successfully genotyped by graphs (Suppl. Fig. 5); this gap was likely not as obvious with SNPs and indels, since numerous markers (over 61,000 SNPs and 8,000 indels) could still be genotyped in samples that did not have this putative deletion.

3.3 All genetic variants underlay a congruent population structure

Despite differences in amount and proportion of genome base pairs covered, SVs, SNPs and small indels displayed a consistent population structure and genetic differentiation between the Romaine and Puyjalon populations. Principal component analysis revealed an important differentiation between the two populations, with individual salmon clustering by river along PC1 while PC2 explained variation within Puyjalon samples (Figure 4). This pattern was strongly conserved across all variant types and confirmed anticipated population structure from a previous study (Albert & Bernatchez, 2006). Average genome-wide weighted F_{ST} values ranged from 0.044 for SVs to 0.065 for SNPs and 0.079 for small indels (Table

6). This observation, along with the strong clustering of samples in principal component analysis, indicates moderate levels of differentiation between Romaine and Puyjalon samples.

Differentiation along the genome was also highly variable, as regions of strong per-window F_{ST} (> 0.2) were numerous and dispersed on all chromosomes (Figure 5). These high F_{ST} peaks were overall consistent across SVs, SNPs and short indels. However, the correlation was the greatest between SNPs and indels (R^2 = 0.930, *p*-value < 0.001), which shared multiple peaks and tended to have higher F_{ST} than SVs. SVs displayed a weaker correlation with SNPs (R^2 = 0.612, *p*-value < 0.001) and indels (R^2 = 0.595, *p*-value < 0.001). Only a few per-window F_{ST} peaks were unique to SVs (e.g., on chromosomes ssa01q, 10, 29; Figure 5). Per-variant F_{ST} distribution also differed between variant types: values ranged from 0 to 0.981 for SNPs, whereas the maximum per-variant F_{ST} observed for SVs was 0.702 (Table 6).

3.4 Candidate variants for local adaptation overlapped genes involved in putatively important biological functions

Among all filtered variants of each type, we identified those that showed a possible relevance in the putative local adaptation of Romaine and Puyjalon salmon. For each variant type, we reported the most strongly differentiated variants between both populations, as these could include major-effect variants, as well as RDA candidates, which might instead reveal multiple, small-effect loci. We identified 1.62 times more outliers of differentiation than RDA candidates for SVs, whereas SNPs and small indels showed the opposite trend, with 1.59 and 2.23 times more RDA candidates than outliers for SNPs and indels, respectively (Table 7). These outlier and RDA candidate variants did not have more missing data than the others, non-outlier and non-candidate variants (Suppl. Fig. 6), meaning that the apparent differentiation and multilocus signals were not artificially driven by missing genotypes. For each of these six sets of variants of interest, we reported a set of overlapping genes within 10 kb, ranging from only 940 genes for RDA candidate SVs to 15,711 genes for RDA candidate SNPs (Table 8). GO enrichment analysis performed on each of these six gene sets revealed various biological processes, with a redundancy of terms associated with cellular structure and nervous system function. The 1,407 genes located near SV outliers were enriched for 108 terms mostly related to cellular adhesion and junction, as well as to synapse organization (Suppl. Table 5), whereas RDA candidate SV genes were enriched for only 27 GO terms clustered under "chemical synaptic transmission" (Suppl. Table 6). Many of the GO

terms that were enriched for outlier SV genes, such as cell adhesion and junction, synapse organization or developmental processes, were also among the 223 enriched terms for outlier SNP genes (Suppl. Table 7). A much wider range of biological functions were overrepresented in the 528 GO terms associated with RDA candidate SNP genes, including growth and nervous system development (Suppl. Table 8). Finally, genes that overlapped with outlier and RDA candidate indels were enriched for fewer GO terms (212 and 292, respectively), but enriched terms themselves were similar to those reported for the outlier and RDA candidate SNP gene sets (Suppl. Table 9; Suppl. Table 10). Raw GO enrichment result tables (prior to simplification with REVIGO) can be found in Supplementary Data 1.

4 Discussion

Structural variants now appear as major components of genetic polymorphism with increasingly recognized implications for phenotype and adaptation, but they are inherently difficult to characterize, especially at the population level. By proposing a hybrid pipeline combining short- and long-read sequencing with graph-based genotyping, we aimed to alleviate some of the challenges hampering the study of SVs in population genomics. Applying this pipeline to a set of 60 Atlantic salmon genomes allowed us to catalog a wide spectrum of polymorphism, from SNPs to multi-kilobase SVs. From this catalog, we described the population structure of salmon populations inhabiting two adjacent tributaries, the Romaine and Puyjalon rivers, estimated the level of differentiation between them, and uncovered putative variants underlying local adaptation.

4.1 Multiplatform variant detection and pangenome-based approaches facilitate population-scale analysis of SVs

The combination of long- and short-read sequencing with graph-based genotyping is a promising, yet incomplete solution to the challenges raised by analyzing SVs at the population scale. Our results indicate that long reads are a highly valuable asset for SV characterization, because the vast majority of SVs were identified from the only four genomes sequenced in Oxford Nanopore long reads, despite high-coverage (16X) paired-end Illumina short-read data being available for all 60 samples. Long-read data was particularly crucial for detecting putative insertions and characterizing their sequence, as expected from previous studies (Ho, Urban & Mills, 2020). Since insertions and deletions are the most prevalent

forms of SVs in genomes (Gamazon & Stranger, 2015), the incorporation of third generation sequencing enabled a considerable gain in the amount of genetic variation uncovered. However, elevated costs remain an obstacle for population-scale implementation of long-read sequencing, especially for species with large genomes such as salmonids, with prices ranging from 22 to 200 $ USD per gigabase depending on technology used (De Coster, Weissensteiner & Sedlazeck, 2021). Here, graph-based genotyping allowed us to exploit short reads to genotype putative SVs in a much higher number of samples whose genomes were only sequenced in short reads. Indeed, about 93.9% of the 115,907 final, filtered and genotyped SVs were initially called from long reads (or from both short and long reads; Suppl. Table 11) and could be genotyped in more than half of all samples, indicating that even though short reads are suboptimal for SV detection, they are highly relevant for genotyping SVs in populations. The hybrid approach thus offered a major increase in overall SV genotyping power at a much lower cost, which therefore represents a reasonable compromise between cost and efficiency.

This multiplatform approach nevertheless does not address all issues pertaining to SV characterization. First, merging SV calls across tools relies on arbitrary thresholds and may be suboptimal, depending on the precision of SV characterization procedures. We made the arbitrary decision to retain only SVs that were supported by at least two callers, which might be an overly conservative filtering approach. Given the high false discovery rates previously reported for SV calling in the Atlantic salmon genome (Bertolotti et al., 2020), we chose to attempt improving precision over sensitivity. However, multitool calling is not guaranteed to improve precision nor sensitivity, depending on datasets and callers used in combination, and callers relying on the same evidence types are likely to call the same false positives (Chaisson et al., 2019; Kosugi et al., 2019; Mahmoud et al., 2019).

In addition, despite using Jasmine, an up-to-date graph-based minimum spanning forest algorithm that integrates numerous SV parameters, a surprisingly small number of all SVs called by multiple tools could be merged together. This is especially true for merged long-read SVs, for which the proportion of shared calls was even smaller than in the merged short-read SV set. We suggest that the lower basecalling accuracy inherent to third generation sequencing data might increase breakpoint imprecision (Lemay et al., 2022) and lead to undermerging of shared variants, especially for those that could not be refined by Iris. For example, long-read-based callers each reported a few thousand inversions, but only

208 were successfully merged across Sniffles and NanoVar. This likely results from imprecision of inversion breakpoints (Sudmant et al., 2015a), as the breakpoints refinement tool (Iris) cannot process inversions at this time, or from differences in how inversions are identified or reported by different tools. All these issues highlight the need for more efficient validation methods for large SV datasets. Since visualization remains the most robust SV validation method (Spies et al., 2015), emerging machine learning-based and/or automated methods show promising applications for this purpose, such as samplot-ML (Belyeu et al., 2021), MAVIS (Reisle et al. 2019) and DeepSVFilter (Liu et al., 2020), but also for SV detection and genotyping (Cue; Popic et al., 2023). We can expect that such tools, which currently primarily target short-read data, will eventually allow automated long-read SV validation as well, thus further improving SV analysis in the upcoming years.

Although graph-based genotyping has been essential for genotyping long-read SVs from short-read data, it is not immune to bias in SV characterization. We explicitly treated genotypes with insufficient read support (or genotype quality) as missing data, and up to 40% of SVs were filtered out due to missing genotypes in at least half of the samples, meaning that very few reads could be mapped to these SV regions in the pangenome. Since long-read SVs had a consistently higher proportion of missing genotypes than short-read SVs, both before and after filtering on genotype quality, depth and minor allele frequency (Suppl. Fig. 7), we speculate that some short reads still cannot be accurately mapped to certain SV regions where long reads could be confidently aligned. This might also explain the lower concordance between the genotypes outputted by vg and the genotypes provided by caller genotypes for candidate long-read SVs than for candidate short-read SVs (Suppl. Table 12). As stated above, the higher breakpoint imprecision for long-read SVs might also increase noise around SV positions and therefore contribute to the poorer mapping of short reads to the graph. Various features of the Atlantic salmon genome are known to promote spurious mapping of reads to the linear reference genome, such as residual tetrasomy (10 to 20%; Houston & Macqueen, 2019), highly similar duplicated regions (81 – 89%; Davidson et al., 2010) and a large proportion of repeats (50 – 60%; de Boer et al., 2007). We can expect that these features also impact pangenome-based mapping and genotyping to some extent, especially repeats (Chen et al., 2019; Outten et al., 2021). Indeed, low-confidence SVs that were filtered out after genotyping were more prevalent in regions with repeated contents (transposable elements and repeats)

and/or in syntenic regions with elevated levels of homology following a past whole-genome duplication event in salmonids (Suppl. Fig. 8). Second, very large putative SVs spanning considerable chromosomal regions, such as the 2.5-Mb deletion on chromosome ssa10, could not be successfully genotyped using graphs, likely because such large rearrangements cannot be reliably represented by complex graph structures (Hübner, 2022). This limitation is particularly problematic for the study of SVs in the context of population genomics, as larger rearrangements were found to play a key role in adaptive processes (Wellenreuther & Bernatchez, 2018; Wellenreuther et al., 2019). Alternatively, some of the very large candidate SVs that were not successfully genotyped could have been false positives, as over half of putative SVs larger than 30 kb were inversions and deletions exclusively supported by short-read callers (Suppl. Table 13). Our study would therefore benefit from the addition of complementary approaches, such as assembly comparison or chromatin conformation data, to identify, validate and genotype large SVs (Mérot et al., 2020), hence further expanding the range of SVs identified. Despite these limitations, the multiplatform strategy developed for this study represents a considerable improvement over short-read-only approaches, and the incorporation of novel automated curation approaches will undoubtedly lead to further advancements in population-scale SV characterization.

4.2 SVs are a key feature of the Atlantic salmon genome

Our findings showed that the contribution of SVs to standing genetic polymorphism is important in Atlantic salmon. High-confidence, genotyped SVs accounted for 4.8 times more genome base pairs than SNPs. This proportion is in the same order of magnitude as previously estimated using an equivalent approach in lake whitefish, a closely related species (e.g., five times; Mérot et al., 2023). The number of SVs identified in our study is over seven times larger than previously documented in rainbow trout (almost 14,000 SVs; Liu et al., 2021) and in Atlantic salmon (over 15,000 SVs; Bertolotti et al., 2020) in previous studies involving more samples, but relying only on short-read data. Similarly, we reported between 20 and 30 SVs per 100-kb window, whereas the median per-megabase SV count reported by Bertolotti et al. (2020) is under 10. SV counts described in our study might be inflated due to a certain number of false positives in our dataset, since we did not exclude calls located in problematic regions (e.g., high coverage regions, assembly gaps and low complexity regions) nor performed manual curation of SVs. However, such an important difference in SV count can most likely be explained by the integration of long-read

sequencing data. Indeed, in the human genome, over six times more high-confidence SVs were identified from long reads (27,662 SVs; Chaisson et al., 2019) than from short reads in another study (4,442 SVs; Abel et al., 2020).

4.3 SVs are informative markers relevant for population genomics studies

SVs also appear to reliably capture population structure and differentiation to the same extent as SNPs. The very high correspondence of population structure inferred from PCA across variant types was also observed in previous studies of SVs in soybean (Lemay et al., 2022), in cacao (*Theobroma cacao*; Hämälä et al., 2021), in grapevine (*Vitis vinifera* ssp. *Sativa*; Zhou et al., 2019), lake whitefish (Mérot et al., 2023) and *Corvus* genus species (Weissensteiner et al., 2020). Patterns of fluctuations in per-window variant density and F_{ST} along the genome were also strongly conserved e.g., regions of high SV density were usually dense in SNPs and short indels as well. We reported a quick linkage disequilibrium decay in all pairs of variants, with minimal linkage between SNPs and SVs for distances greater than 250 bp (Suppl. Fig. 9). This suggests that the observed correspondence between all three types of variants might not be attributable to strong physical linkage between them, but rather that SVs, SNPs and short indels may be subject to similar evolutionary processes in the Romaine and Puyjalon system, despite them being very different in size. Per-variant and per-window F_{ST} was usually slightly lower for SVs than for SNPs and small indels, which was also reported in the lake whitefish study (Mérot et al., 2023). We suspect that this slight discrepancy is attributable to the fact that the F_{ST} calculation relies on fewer markers for SVs, thus introducing more noise in F_{ST} estimates than with SNPs and small indels.

By contrast, in American lobster, a non-related and less structured marine species (Kenchington et al., 2009; Benestan et al., 2015), copy number variants harbored stronger interpopulation differentiation than SNPs and a more defined population structure, correlated with environmental variables (Dorant et al., 2020). Similarly, deletions showed stronger spatial population structure and were under stronger selection than duplications in human populations (Sudmant et al., 2015b). In European starling (*Sturnus vulgaris*), SVs and SNPs revealed different patterns of population structure, interpopulation genetic diversity and divergence across the genome (Stuart et al., 2023). Consequently, we cannot assume that SVs, SNPs and small indels are interchangeable and equally informative in all systems and species as we observed in the Romaine and Puyjalon system. We therefore argue that we ought to characterize SVs in population

genomics studies to the same extent as SNPs, as they may display different signatures and provide relevant insights on evolutionary and adaptive processes shaping population structure.

4.4 Genetic divergence between the Romaine and Puyjalon populations is likely driven by divergent selection

Given their spatial proximity and habitat overlap, gene flow is expected to occur between the Romaine and Puyjalon populations. The moderate average genome-wide F_{ST} values estimated from the different classes of variants, as well as a previous estimation based on microsatellites (Albert & Bernatchez, 2006), are consistent with a rate of two to seven migrants per generation, based on Wright's approximation (1978). However, we reported numerous outliers of differentiation and peaks of very strong F_{ST} dispersed throughout the genome that are seemingly resisting such gene flow. The persistence of pronounced divergence in these regions could be explained by a few alternative and not mutually exclusive mechanisms. Genetic drift could lead to random differences in variant allelic frequency in both populations. Alternatively, given that recombination rates are known to differ within species and even within populations (Kong et al., 2010; Ritz, Noor & Singh, 2017), some localized regions of low recombination could have emerged independently in both populations, capturing different alleles and being subject to increased genetic drift as a result of apparent reduced effective population size (N_e). Such "differentiation islands" can result from the interplay of variation in recombination rate due to genetic architecture (e.g., the presence of SVs or large rearrangements) and natural selection (Wolf & Ellegren, 2017). Finally, variants with a functional impact could be subject to divergent selection between habitats leading to differences in allelic frequencies.

Our findings tend to support the hypothesis of local adaptation. First, we reported a repeated enrichment for GO terms related to nervous system function for genes nearby outlier and RDA candidate variants, regardless of variant type. We initially expected enrichment for functions related to other observed phenotypic trait variation in the Romaine and Puyjalon system, such as growth, sexual maturation and reproduction. On the contrary, we observed enrichment mainly related to nervous functions. We hypothesize that enrichment for nervous functions could be linked to variation in these traits through their link with age at smoltification that differ between these two populations. Indeed, changes in photoperiod, which are recognized as the main factor triggering smoltification (Hoar, 1988), are perceived

and processed through the light-brain-pituitary axis, inducing the hormonal cascade responsible for physiological, morphological and behavioral changes underlying smolt-to-parr transition (Stefansson et al., 2008). Smoltification itself causes reorganization of nervous connections, both at the structural and the chemical level (Ebbesson et al., 2003). Although empirical evidence is required to support this hypothesis, polymorphism around genes involved in nervous system function, development and plasticity could alter the expression or function of these genes and lead to physiological differences underlying variation in age at smoltification and other relevant life history traits in Romaine and Puyjalon salmon. This indirect, but plausible link between observed phenotypic variation and genetic polymorphism does not support the hypothesis of persistent genetic differentiation due to genetic drift alone. Moreover, peaks of differentiation are unlikely a result of low recombination alone because such peaks and outliers of differentiation are dispersed throughout chromosomes and across the whole genome, and not clustered into contiguous and localized regions. Along with preliminary knowledge of the Romaine and Puyjalon system, our results suggest that the persistence of localized regions of strong differentiation could at least partly be attributable to local adaptation in response to divergent selection. Indeed, both rivers differ in habitat quality, substrate and hydrological profiles (Schieffer,1975; Fontaine et al., 2000; GENIVAR, 2002; Belles-Isles et al., 2004; WSP Global, 2019), which may impose different constraints on salmon and thus favor alternate life strategies in both populations.

Besides showing variation in age at smoltification, Romaine and Puyjalon salmon also differ in regards to age at sexual maturity in a controlled hatchery environment (Therrien et al., 2017; Langlois-Parisé et al., 2018; T. Dion et al., 2020; T. Dion, Langlois-Parisé & Proulx, 2020). Interestingly, we found no variant of interest overlapping with major-effect loci previously associated with life history variation in age at maturity in wild and domesticated European salmon populations, such as *vgll3* (Ayllon et al., 2015; Barson et al., 2015; Czorlich et al., 2018) and *six6* (Sinclair-Waters et al., 2020; Waters et al., 2021). While one study in North America highlighted a correlation between *vgll3* polymorphism, sea age and sex in the Trinité river population (Kusche et al., 2017), located in the same geographic region as the Romaine and Puyjalon rivers, other studies did not reveal a significant association between polymorphism in these major-effect loci and age at maturity in other North American populations (Boulding et al., 2019; Mohamed et al., 2019). The genetic architecture of such complex life history traits is possibly variable across

populations, especially between highly divergent populations from different continents. In addition, age at maturity was found to have a mixed genetic architecture in both North American-derived farmed salmon (Eisbrenner et al., 2014; Mohamed et al., 2019) and European-origin salmon (Sinclair-Water et al., 2020), involving both major-effect loci and multiple small-effect loci. Since we identified numerous candidate small-effect variants through RDA, we propose that phenotype variation in age at maturity in Romaine and Puyjalon salmon might have a polygenic basis as well.

Further work is required to understand the genetic architecture of major life history trait variation in Atlantic salmon populations as well as other salmonid species. Such work would considerably benefit from an improved knowledge of the full spectrum of genetic variation segregating in populations, especially SVs. The pipelines developed and optimized for this study may therefore contribute this knowledge by facilitating population-scale characterization of SV, as well as serve as a basis for further refinement of variant calling and genotyping procedures in the near future. With ongoing and rapid developments in computational genomic approaches, such as pangenome-based tools or machine-learning-based variant detection and validation, SV analysis is bound to take a significant leap towards robust and reliable characterization in the upcoming years, which will foster their inclusion in evolutionary genomics.

Literature cited

Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., . . . NHGRI Centers for Common Disease Genomics. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature, 583*(7814), 83-89. https://doi.org/10.1038/s41586-020-2371-0

Ahsan, M. U., Liu, Q., Fang, L., & Wang, K. (2021). NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biology, 22*(1), 261. https://doi.org/10.1186/s13059-021-02472-2

Albert, V., & Bernatchez, L. (2006). *Complexe de la Romaine – Caractérisation génétique des populations de saumon atlantique. Rapport sectoriel présenté à GENIVAR Société en commandite et à Hydro-Québec Équipement.* Université Laval.

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In B. J. Turner (Ed.), *Evolutionary genetics of fishes* (pp. 1-53). Boston, MA: Springer US.

Ameur, A. (2019). Goodbye reference, hello genome graphs. *Nature Biotechnology, 37*(8), 866-868. https://doi.org/10.1038/s41587-019-0199-7

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., . . . National Eye Institute N. I. H. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. https://doi.org/10.1038/nature15393

Ayllon, F., Kjaerner-Semb, E., Furmanek, T., Wennevik, V., Solberg, M. F., Dahle, G., . . . Wargelius, A. (2015). The *vgll3* locus controls age at maturity in wild and domesticated Atlantic salmon (*Salmo salar* L.) males. *Plos Genetics, 11*(11), 15. https://doi.org/10.1371/journal.pgen.1005628

Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution, 23*(1), 38-44. https://doi.org/10.1016/j.tree.2007.09.008

Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., . . . Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature, 528*(7582), 405-408. https://doi.org/10.1038/nature16062

Belles-Isles, M., Plourde, Y., Pelletier, P., Théberge, C., & Thibodeau, P. (2004). *Complexe Romaine – Aménagement intégral de la rivière Romaine : Étude préliminaire sur les débits réservés et la faune ichtyenne*. Environnement et Services techniques d'Hydro-Québec. GENIVAR.

Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., & Layer, R. M. (2021). Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biology, 22*(1), 161. https://doi.org/10.1186/s13059-021-02380-5

Belyeu, J. R., Nicholas, T. J., Pedersen, B. S., Sasani, T. A., Havrilla, J. M., Kravitz, S. N., . . . Layer, R. M. (2018). SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *GigaScience, 7*(7). https://doi.org/10.1093/gigascience/giy064

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology, 24*(13), 3299-3315. https://doi.org/10.1111/mec.13245

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berg, P. R., Star, B., Pampoulie, C., Bradbury, I. R., Bentzen, P., Hutchings, J. A., . . . Jakobsen, K. S. (2017). Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. *Heredity, 119*(6), 418-428. https://doi.org/10.1038/hdy.2017.54

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., . . . Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications, 11*(1), 5176. https://doi.org/10.1038/s41467-020-18972-x

Boulding, E. G., Ang, K. P., Elliott, J. A. K., Powell, F., & Schaeffer, L. R. (2019). Differences in genetic architecture between continents at a major locus previously associated with sea age at sexual maturity in European Atlantic salmon. *Aquaculture, 500*, 670-678. https://doi.org/10.1016/j.aquaculture.2018.09.025

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project

Phase I data. *G3 Genes|Genomes|Genetics, 5*(5), 931-941. https://doi.org/10.1534/g3.114.015784

Brenna-Hansen, S., Li, J., Kent, M. P., Boulding, E. G., Dominik, S., Davidson, W. S., & Lien, S. (2012). Chromosomal differences between European and North American Atlantic salmon discovered by linkage mapping and supported by fluorescence in situ hybridization analysis. *BMC Genomics, 13*(1), 432. https://doi.org/10.1186/1471-2164-13-432

Broad Institute. (2019). Picard Toolkit. Retrieved from http://broadinstitute.github.io/picard/

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications, 10*(1), 3240. https://doi.org/10.1038/s41467-019-11146-4

Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology, 28*(6), 1210-1223. https://doi.org/10.1111/mec.15051

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

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., . . . Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications, 10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., . . . Eberle, M. A. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology, 20*(1), 291. https://doi.org/10.1186/s13059-019-1909-7

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

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., . . . Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics, 32*(8), 1220-1222. https://doi.org/10.1093/bioinformatics/btv710

Circulomics. (2021). Nanobind HMW DNA Extraction – Nucleated Blood Protocol (EXT-NBH-001). Retrieved from https://15a13b02-7dac-4315-baa5-b3ced1ea969d.filesusr.com/ugd/5518db_d6e1e77bc90148f387559fd51e2d85a0.pdf?index=true

Crête-Lafrenière, A., Weir, L. K., & Bernatchez, L. (2012). Framing the Salmonidae family phylogenetic portrait: A more complete picture from increased taxon sampling. *Plos One, 7*(10), 19. https://doi.org/10.1371/journal.pone.0046662

Crown, K. N., Miller, D. E., Sekelsky, J., & Hawley, R. S. (2018). Local inversion heterozygosity alters recombination throughout the genome. *Current Biology, 28*(18), 2984-2990.e2983. https://doi.org/10.1016/j.cub.2018.07.004

Czorlich, Y., Aykanat, T., Erkinaro, J., Orell, P., & Primmer, C. R. (2018). Rapid sex-specific evolution of age at maturity is shaped by genetic architecture in Atlantic salmon. *Nature Ecology & Evolution, 2*(11), 1800-1807. https://doi.org/10.1038/s41559-018-0681-5

Dallaire, X., Bouchard, R., Hénault, P., Ulmo-Diaz, G., Normandeau, E., Mérot, C., . . . Moore, J.-S. (2023). Widespread deviant patterns of heterozygosity in whole-genome sequencing due to autopolyploidy, repeated elements, and duplication. *bioRxiv*, 2023.2007.2027.550877. https://doi.org/10.1101/2023.07.27.550877

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England), 27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). https://doi.org/10.1093/gigascience/giab008

Davidson, W. S., Koop, B. F., Jones, S. J., Iturra, P., Vidal, R., Maass, A., . . . Omholt, S. W. (2010). Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biol, 11*(9), 403. https://doi.org/10.1186/gb-2010-11-9-403

de Boer, J. G., Yazawa, R., Davidson, W. S., & Koop, B. F. (2007). Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. *BMC Genomics, 8*(1), 422. https://doi.org/10.1186/1471-2164-8-422

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34*(15), 2666-2669. https://doi.org/10.1093/bioinformatics/bty149

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics, 22*(9), 572-587. https://doi.org/10.1038/s41576-021-00367-3

Debes, P. V., Piavchenko, N., Ruokolainen, A., Ovaskainen, O., Moustakas-Verho, J. E., Parre, N., . . . Primmer, C. R. (2021). Polygenic and major-locus contributions to sexual maturation timing in Atlantic salmon. *Molecular Ecology, 30*(18), 4505-4519. https://doi.org/10.1111/mec.16062

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *Plos One, 16*(10), e0257521. https://doi.org/10.1371/journal.pone.0257521

Dionne, M., Caron, F., Dodson, J. J., & Bernatchez, L. (2008). Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology, 17*(10), 2382-2396. https://doi.org/10.1111/j.1365-294X.2008.03771.x

Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., . . . Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype-temperature association in a marine species. *Molecular Ecology*, 18. https://doi.org/10.1111/mec.15565

Ebbesson, L. O. E., Ekstro□m, P., Ebbesson, S. O. E., Stefansson, S. O., & Holmqvist, B. (2003). Neural circuits and their structural and chemical reorganization in the light-brain-pituitary axis during parr-smolt transformation in salmon. *Aquaculture, 222*(1), 59-70. https://doi.org/10.1016/S0044-8486(03)00102-9

Eisbrenner, W. D., Botwright, N., Cook, M., Davidson, E. A., Dominik, S., Elliott, N. G., . . . Davidson, W. S. (2014). Evidence for multiple sex-determining loci in Tasmanian Atlantic salmon (*Salmo salar*). *Heredity, 113*(1), 86-92. https://doi.org/10.1038/hdy.2013.55

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics, 7*(2), 85-97. https://doi.org/10.1038/nrg1767

Feuk, L., Marshall, C. R., Wintle, R. F., & Scherer, S. W. (2006). Structural variants: changing the landscape of chromosomes and design of disease studies. *Human Molecular Genetics, 15*(Spec No 1), R57-R66. https://doi.org/10.1093/hmg/ddl057

Fontaine, P.-M., Levesque, F., Proulx, M., & Heppell, M. (2000). *Étude du saumon de la rivière Romaine en 1999 - Rapport présenté à la direction Expertise et support technique de production, Hydro-Québec par le Groupe conseil GENIVAR.*

Forester, B. R., Laporte, M., & Manel, S. (2018). Detecting multilocus adaptation using Redundancy Analysis (RDA). Retrieved from https://popgen.nescent.org/2018-03-27_RDA_GEA.html

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology, 27*(9), 2215-2233. https://doi.org/10.1111/mec.14584

Fraser, D. J., & Bernatchez, L. (2005). Adaptive migratory divergence among sympatric brook charr populations. *Evolution, 59*(3), 611-624. https://doi.org/10.1111/j.0014-3820.2005.tb01020.x

Gamazon, E. R., & Stranger, B. E. (2015). The impact of human copy number variation on gene expression. *Briefings in Functional Genomics, 14*(5), 352-357. https://doi.org/10.1093/bfgp/elv017

GENIVAR. (2002). *Aménagement hydroélectrique de la Romaine-1 – Étude de la population de saumon atlantique de la rivière Romaine en 2001*. GENIVAR.

Gjedrem, T., & Rye, M. (2018). Selection response in fish and shellfish: a review. *Reviews in Aquaculture, 10*(1), 168-179. https://doi.org/10.1111/raq.12154

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences, 118*(35), e2102914118. https://doi.org/10.1073/pnas.2102914118

Harris, R. S. (2007). Improved pairwise alignment of genomic DNA. (Philosophiae Doctor (PhD). The Pennsylvania State University, University Park, United-States.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., . . . Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology, 21*(1), 35. https://doi.org/10.1186/s13059-020-1941-7

Ho, S. V. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics, 21*(3), 171-189. https://doi.org/10.1038/s41576-019-0180-9

Hoar, W. S. (1988). The physiology of smolting salmonids. In W. S. Hoar & D. J. Randall (Eds.), *Fish Physiology* (Vol. 11, pp. 275-343): Academic Press.

Houston, R. D., & Macqueen, D. J. (2019). Atlantic salmon (*Salmo salar* L.) genetics in the 21st century: taking leaps forward in aquaculture and biological understanding. *Animal Genetics, 50*(1), 3-14. https://doi.org/10.1111/age.12748

Hübner, S. (2022). Are we there yet? Driving the road to evolutionary graph-pangenomics. *Current Opinion in Plant Biology, 66*. https://doi.org/10.1016/j.pbi.2022.102195

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., . . . Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research, 27*(5), 677-685. https://doi.org/10.1101/gr.214007.116

Jain, C., Rhie, A., Hansen, N. F., Koren, S., & Phillippy, A. M. (2022). Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature Methods, 19*(6), 705-710. https://doi.org/10.1038/s41592-022-01457-8

Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Weighted minimizer sampling improves long read mapping. *Bioinformatics, 36*(Supplement_1), i111-i118. https://doi.org/10.1093/bioinformatics/btaa435

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . Team Whole Genome Assembly. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature, 484*(7392), 55-61. https://doi.org/10.1038/nature10944

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., . . . ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature, 477*(7363), 203-206. https://doi.org/10.1038/nature10341

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research*. https://doi.org/10.1101/gr.176552.114

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters, 7*(12), 1225-1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x

Kenchington, E. L., Harding, G. C., Jones, M. W., & Prodöhl, P. A. (2009). Pleistocene glaciation events shape genetic structure across the range of the American lobster, *Homarus americanus*. *Molecular Ecology, 18*(8), 1654-1667. https://doi.org/10.1111/j.1365-294X.2009.04118.x

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics, 173*(1), 419-434. https://doi.org/10.1534/genetics.105.047985

Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., & Schatz, M. C. (2023). Jasmine and Iris: population-scale structural variant comparison and analysis. *Nature Methods*. https://doi.org/10.1038/s41592-022-01753-3

Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., . . . Andersen, O. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology, 25*(10), 2130-2143. https://doi.org/10.1111/mec.13592

Klemetsen, A., Amundsen, P.-A., Dempson, J. B., Jonsson, B., Jonsson, N., O'Connell, M. F., & Mortensen, E. (2003). Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): a review of aspects of their life histories. *Ecology of Freshwater Fish, 12*(1), 1-59. https://doi.org/10.1034/j.1600-0633.2003.00010.x

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., . . . Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports, 8*(1), 10872. https://doi.org/10.1038/s41598-018-28948-z

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., . . . Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature, 467*(7319), 1099-1103. https://doi.org/10.1038/nature09525

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics, 15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Kosugi, S., Momozawa, Y., Liu, X. X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology, 20*, 18. https://doi.org/10.1186/s13059-019-1720-5

Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., . . . Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics, 48*(1), 79-83. https://doi.org/10.1038/ng.3443

Kusche, H., Cote, G., Hernandez, C., Normandeau, E., Boivin-Delisle, D., & Bernatchez, L. (2017). Characterization of natural variation in North American Atlantic Salmon populations (Salmonidae: *Salmo salar*) at a locus with a major effect on sea age. *Ecology and Evolution, 7*(15), 5797-5807. https://doi.org/10.1002/ece3.3132

Langlois-Parisé, I., T. Dion, M.-C., & Therrien, J.-C. (2018). *Rapport d'activité 2016 au LARSA, Bilan récapitulatif des opérations, Programme de restauration des populations de saumons de la rivière Romaine*. LARSA. Université Laval. Québec.

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

Lehnert, S. J., Bentzen, P., Kess, T., Lien, S., Horne, J. B., Clément, M., & Bradbury, I. R. (2019). Chromosome polymorphisms track trans-Atlantic divergence and secondary contact in Atlantic salmon. *Molecular Ecology, 28*(8), 2074-2087. https://doi.org/10.1111/mec.15065

Lemay, M.-A., Sibbesen, J. A., Torkamaneh, D., Hamel, J., Levesque, R. C., & Belzile, F. (2022). Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. *BMC Biology, 20*(1), 53. https://doi.org/10.1186/s12915-022-01255-w

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., . . . Kwok, P.-Y. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications, 10*(1), 1025. https://doi.org/10.1038/s41467-019-08992-7

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv, 1303*.

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics, 211*(1), 289-304. https://doi.org/10.1534/genetics.118.301747

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., . . . Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature, 533*(7602), 200-+. https://doi.org/10.1038/nature17164

Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G., & de Ridder, D. (2015). Making the difference: integrating structural variation detection tools. *Briefings in Bioinformatics, 16*(5), 852-864. https://doi.org/10.1093/bib/bbu047

Liu, S., Gao, G., Layer, R. M., Thorgaard, G. H., Wiens, G. D., Leeds, T. D., . . . Palti, Y. (2021). Identification of high-confidence structural variants in domesticated rainbow trout using whole-genome sequencing. *Frontiers in Genetics, 12*. https://doi.org/10.3389/fgene.2021.639355

Liu, Y., Huang, Y., Wang, G., & Wang, Y. (2020). A deep learning approach for filtering structural variants in short read sequencing data. *Briefings in Bioinformatics, 22*(4). https://doi.org/10.1093/bib/bbaa370

Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): Support for the ecological speciation hypothesis. *Evolution, 53*(5), 1491-1505. https://doi.org/10.1111/j.1558-5646.1999.tb05413.x

Ma, J., & Amos, C. I. (2012). Investigation of inversion polymorphisms in the human genome using principal components analysis. *Plos One, 7*(7), e40224. https://doi.org/10.1371/journal.pone.0040224

Mahmoud, M., Gobet, N., Cruz-Davalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology, 20*(1), 14. https://doi.org/10.1186/s13059-019-1828-7

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology, 14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

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*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics, 210*(2), 719-731. https://doi.org/10.1534/genetics.118.301336

Mérot, C., Berdan, E. L., Babin, C., Normandeau, E., Wellenreuther, M., & Bernatchez, L. (2018). Intercontinental karyotype - environment parallelism supports a role for a chromosomal inversion in local adaptation in a seaweed fly. *Proceedings of the Royal Society B: Biological Sciences, 285*(1881), 10. https://doi.org/10.1098/rspb.2018.0519

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution, 35*(7), 561-572. https://doi.org/10.1016/j.tree.2020.03.002

Mérot, C., Stenløkk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., . . . Bernatchez, L. (2023). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Molecular Ecology, 32*(6), 1458-1477. https://doi.org/10.1111/mec.16468

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., . . . 1000 Genomes Project. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature, 470*(7332), 59-65. https://doi.org/10.1038/nature09708

Mohamed, A. R., Verbyla, K. L., Al-Mamun, H. A., McWilliam, S., Evans, B., King, H., . . . Kijas, J. W. (2019). Polygenic and sex specific architecture for two maturation traits in farmed Atlantic salmon. *BMC Genomics, 20*(1), 139. https://doi.org/10.1186/s12864-019-5525-4

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*(6), 443-451. https://doi.org/10.1038/nrg2986

Norwegian University of Life Sciences. (2022). Ssal_Brian_v1.0 (GCA_923944775.1). Retrieved from https://www.ncbi.nlm.nih.gov/assembly/GCA_923944775.1/?shouldredirect=false

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., . . . Weedon, J. (2022). vegan: Community Ecology Package (Version 2.6-4). Retrieved from https://github.com/vegandevs/vegan

Outten, J., & Warren, A. (2021). Methods and developments in graphical pangenomics. *Journal of the Indian Institute of Science, 101*(3), 485-498. https://doi.org/10.1007/s41745-021-00255-z

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research, 27*(5), 665-676. https://doi.org/10.1101/gr.214155.116

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., . . . Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology & Evolution, 3*(12), 1731-1742. https://doi.org/10.1038/s41559-019-1044-6

Pearse, D. E., Miller, M. R., Abadía-Cardoso, A., & Garza, J. C. (2014). Rapid parallel evolution of standing variation in a single, complex, genomic region is associated with life history in steelhead/rainbow trout. *Proceedings of the Royal Society B: Biological Sciences, 281*(1783), 20140012. https://doi.org/10.1098/rspb.2014.0012

Pedersen, B. S., Layer, R., & Quinlan, A. R. (2020). smoove: structural-variant calling and genotyping with existing tools. Retrieved from https://github.com/brentp/smoove

Popic, V., Rohlicek, C., Cunial, F., Hajirasouliha, I., Meleshko, D., Garimella, K., & Maheshwari, A. (2023). Cue: a deep-learning framework for structural variant discovery and genotyping. *Nature Methods, 20*(4), 559-568. https://doi.org/10.1038/s41592-023-01799-x

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

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology, 19*(1), 90. https://doi.org/10.1186/s13059-018-1462-9

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics, 28*(18), i333-i339. https://doi.org/10.1093/bioinformatics/bts378

Reisle, C., Mungall, K. L., Choo, C., Paulino, D., Bleile, D. W., Muhammadzadeh, A., . . . Jones, S. J. M. (2018). MAVIS: merging, annotation, validation, and illustration of structural variants. *Bioinformatics, 35*(3), 515-517. https://doi.org/10.1093/bioinformatics/bty621

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology, 24*(17), 4348-4370. https://doi.org/10.1111/mec.13322

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution, 16*(7), 351-358. https://doi.org/10.1016/s0169-5347(01)02187-5

Ritz, K. R., Noor, M. A. F., & Singh, N. D. (2017). Variation in recombination rate: Adaptive or not? *Trends in Genetics, 33*(5), 364-374. https://doi.org/10.1016/j.tig.2017.03.003

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R., & Weigel, D. (2019). An ultra high-density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. *Genetics, 213*(3), 771-787. https://doi.org/10.1534/genetics.119.302406

Schieffer, K. (1975). *Atlantic salmon management study of the Romaine river*. Beak Consultants Ltd. Sorel, Québec.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018a). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods, 15*(6), 461-468. https://doi.org/10.1038/s41592-018-0001-7

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018b). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics, 19*(6), 329-346. https://doi.org/10.1038/s41576-018-0003-4

Sinclair-Waters, M., Ødegård, J., Korsvoll, S. A., Moen, T., Lien, S., Primmer, C. R., & Barson, N. J. (2020). Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon. *Genetics Selection Evolution, 52*(1), 9. https://doi.org/10.1186/s12711-020-0529-8

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., . . . Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science, 374*(6574), abg8871. https://doi.org/10.1126/science.abg8871

Smit, A.F.A, Hubley, R., & Green, P. (2013). RepeatMasker Open-4.0

Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., . . . Sedlazeck, F. J. (2022). Comprehensive structural variant detection: From mosaic to population-level. *bioRxiv*, 2022.2004.2004.487055. https://doi.org/10.1101/2022.04.04.487055

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Research, 39*(10), e68-e68. https://doi.org /10.1093/nar/gkr123

Spielmann, M., Lupiáñez, D. G., & Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews Genetics, 19*(7), 453-467. https://doi.org/10.1038/s41576-018-0007-0

Spies, N., Zook, J. M., Salit, M., & Sidow, A. (2015). svviz: a read viewer for validating structural variants. *Bioinformatics, 31*(24), 3994-3996. https://doi.org/10.1093/bioinformatics/btv478

Stefansson, S. O., Björnsson, B. T., Ebbesson, L. O., & McCormick, S. D. (2008). Smoltification. In R. N. Finn (Ed.), *Fish Larval Physiology* (pp. 639-681). https://doi.org/10.1201/9780429061608

Stenløkk, K. S. R. (2023). *Genomic structural variations as drivers of adaptation in salmonid fishes.* (Philosophiae Doctor (PhD)). Norwegian University of Life Sciences, Ås, Norway. (2023:26)

Stuart, K. C., Edwards, R. J., Sherwin, W. B., & Rollins, L. A. (2023). Contrasting patterns of single nucleotide polymorphisms and structural variation across multiple invasions. *Molecular Biology and Evolution, 40*(3). https://doi.org/10.1093/molbev/msad046

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . 1000 Genomes Project Consortium. (2015a). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. https://doi.org/10.1038/nature15394

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., . . . Eichler, E. E. (2015b). Global diversity, population stratification, and selection of human copy-number variation. *Science, 349*(6253), aab3761. https://doi.org/10.1126/science.aab3761

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of Gene Ontology terms. *Plos One, 6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

T. Dion, M.-C., Chayer, M., Proulx, E., Langlois-Parisé, I., & Therrien, J.-C. (2020). *Rapport d'activité 2017 au LARSA, Bilan récapitulatif des opérations, Programme de restauration des populations de saumons de la rivière Romaine*. LARSA. Université Laval. Québec.

T. Dion, M.-C., Langlois-Parisé, I., & Proulx, E. (2020). *Rapport d'activité 2018 au LARSA, Bilan récapitulatif des opérations, Programme de restauration des populations de saumons de la rivière Romaine*. Université Laval. Québec.

Taylor, E. B. (1991). A review of local adaptation in Salmonidae, with particular reference to Pacific and Atlantic salmon. *Aquaculture, 98*(1), 185-207. https://doi.org/10.1016/0044-8486(91)90383-I

Tham, C. Y., Tirado-Magallanes, R., Goh, Y., Fullwood, M. J., Koh, B. T. H., Wang, W., . . . Benoukraf, T. (2020). NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biology, 21*(1), 56. https://doi.org/10.1186/s13059-020-01968-7

Therrien, J.-C., T. Dion, M.-C., Ouellet-Cauchon, G., & Langlois-Parisé, I. (2017). *Rapport d'activité 2014-2015 au LARSA, Bilan récapitulatif des opérations, Programme de restauration des populations de saumons de la rivière Romaine*. LARSA. Université Laval. Québec.

Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity, 113*(1), 1-8. https://doi.org/10.1038/hdy.2014.20

Tuttle, E. M. (2003). Alternative reproductive strategies in the white-throated sparrow: behavioral and genetic evidence. *Behavioral Ecology, 14*(3), 425-432. https://doi.org/10.1093/beheco/14.3.425

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., . . . Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature, 534*(7605), 102-105. https://doi.org/10.1038/nature17951

Vincent, B., Dionne, M., Kent, M. P., Lien, S., & Bernatchez, L. (2013). Landscape genomics in Atlantic salmon (*Salmo salar*) searching for gene-environment interactions driving local adaptation. *Evolution, 67*(12), 3469-3487. https://doi.org/10.1111/evo.12139

Waters, C. D., Clemento, A., Aykanat, T., Garza, J. C., Naish, K. A., Narum, S., & Primmer, C. R. (2021). Heterogeneous genetic basis of age at maturity in salmonid fishes. *Molecular Ecology, 30*(6), 1435-1456. https://doi.org/10.1111/mec.15822

Watson, K. B., Lehnert, S. J., Bentzen, P., Kess, T., Einfeldt, A., Duffy, S., . . . Bradbury, I. R. (2022). Environmentally associated chromosomal structural variation influences fine-scale population structure of Atlantic Salmon (*Salmo salar*). *Molecular Ecology, 31*(4), 1057-1075. https://doi.org/10.1111/mec.16307

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*(6), 1358-1370. https://doi.org/10.2307/2408641

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., . . . Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications, 11*(1), 3403. https://doi.org/10.1038/s41467-020-17195-4

Wellband, K., Mérot, C., Linnansaari, T., Elliott, J. A. K., Curry, R. A., & Bernatchez, L. (2019). Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Molecular Ecology, 28*(6), 1439-1459. https://doi.org/10.1111/mec.14965

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution, 33*(6), 427-440. https://doi.org/10.1016/j.tree.2018.04.002

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology, 28*(6), 1203-1209. https://doi.org/10.1111/mec.15066

Wold, J. R., Guhlin, J. G., Dearden, P. K., Santure, A. W., & Steeves, T. E. (2023). The promise and challenges of characterizing genome-wide structural variants: A case study in a critically endangered parrot. *Molecular Ecology Resources*, Advance online publication. https://doi.org/10.1111/1755-0998.13783

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics, 18*(2), 87-100. https://doi.org/10.1038/nrg.2016.133

Wright, S. (1984). *Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations*: University of Chicago Press.

WSP Global. (2019). *Complexe de la Romaine - Suivi environnemental 2017 en phase exploitation : Suivi de la population de saumon atlantique*. Retrieved from https://www.ree.environnement.gouv.qc.ca/dossiers/3211-12-086/3211-12-086-16.pdf

Yan, S. M., Sherman, R. M., Taylor, D. J., Nair, D. R., Bortvin, A. N., Schatz, M. C., & McCoy, R. C. (2021). Local adaptation and archaic introgression shape global diversity at human structural variant loci. *eLife, 10*, e67615. https://doi.org/10.7554/eLife.67615

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., . . . Gaut, B. S. (2019). The population genetics of structural variants in grapevine domestication. *Nature Plants, 5*(9), 965-979. https://doi.org/10.1038/s41477-019-0507-8

Tables

Table 1. Number of SVs reported by each short-read-based caller, and number of SV calls merged across these callers (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Short-read caller	SV type				Per-caller total	Merged across callers
	DELs	DUPs	INSs	INVs		
DELLY	104,738	650	15,086	1,390	121,864	
Manta	114,857	3,879	28,546	3,821	151,103	238,492
smoove	27,627	292	0	245	28,164	

Table 2. Number of SVs in the filtered short-read SV set (SR SVs), obtained by merging calls across the three short-reads-based callers, then filtering for a minimum of two supporting tools and a minimum length of 50 bp (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Short-read caller pair	SV type				Per-pair total
	DELs	DUPs	INSs	INVs	
DELLY + Manta	7,807	108	849	894	9,658
DELLY + smoove	2,216	122	0	21	2,359
Manta + smoove	1,434	26	0	69	1,529
DELLY + Manta + smoove	23,304	62	0	129	23,495
Per-type total	34,761	318	849	1,113	37,041

Table 3. Number of SVs reported by each long-read-based caller, and number of SV calls merged across these callers (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Long-read caller	SV type				Per-caller total	Merged across callers
	DELs	DUPs	INSs	INVs		
Sniffles	260,921	635	211,588	1,233	474,377	
SVIM	1,738,486	41,164	1,754,905	4,914	3,539,469	3,832,032
NanoVar	178,957	144,922	123,349	7,469	454,697	

Table 4. Number of SVs in the filtered long-read SV set (LR SVs), obtained by merging calls across the three long-read-based callers, then filtering for a minimum of two supporting tools and a minimum length of 50 bp (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Long-read caller pair	SV type				Per-pair total
	DELs	DUPs	INSs	INVs	
Sniffles + NanoVar	1,831	47	2,628	208	4,714
Sniffles + SVIM	107,490	38	83,797	0	191,325
SVIM + NanoVar	17,287	7,169	13,550	0	38,006
Sniffles + SVIM + NanoVar	72,871	451	38,328	0	111,650
Per-type total	199,479	7,705	138,303	208	345,695

Table 5. Merged SV count by type and sequencing platform (SR: short-read; LR: long-read; SR + LR: short- and long-read). These represent putative SVs in the genomes of Romaine and Puyjalon salmon, prior to genotyping and filtering (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Platform	SV type				Per-platform total
	DELs	DUPs	INSs	INVs	
SR	13,972	305	101	1,034	15,412
LR	178,690	7,692	137,555	129	324,066
SR + LR	20,789	13	748	79	21,629
Per-type total	213,451	8,010	138,404	1,242	361,107

Table 6. Number of filtered SVs, SNPs and small indels used for population genomics analyses along with summary statistics for each variant set. These variants met the population-level filters, i.e., had a minor allele frequency between 0.05 and 0.95 and were genotyped in at least 50% of the 60 samples. The proportion of base pairs covered by variants of a given type was estimated for the whole genome and by 100-kb windows.

Variant type	Number of variants	Genome base pairs covered			F_{ST}	
		Number	Proportion	Maximum per-window proportion	Genome-wide weighted average	Per-variant maximum
SVs	115,907	41,858,002	0.0168	0.941	0.044	0.702
SNPs	8,777,832	8,777,832	0.0035	0.029	0.065	0.981
Indels	1,089,321	10,316,768	0.0041	0.102	0.079	0.949

Table 7. Outlier and candidate variant sets for each type of polymorphism. The intersection outlier set corresponds to variants that were among the upper 3% quantile of per site F_{ST} and that had a Fisher test *q*-value lower than the 0.01 threshold, whereas RDA candidate variants were identified through RDA (redundancy analysis) using a threshold of three standard deviations.

Variant type	Outliers			RDA candidates
	F_{ST}	Fisher	Intersection	
SVs	3,484	5,298	2,079	1,280
SNPs	263,316	591,621	71,953	114,637
Indels	32,544	103,829	6,223	13,871

Table 8. Number of genes overlapping intersection outlier variants and RDA candidate variants for each type of polymorphism, along with the number of significant enriched GO terms for each gene set. Variants and genes were considered as overlapping if they were located within 10 kb of each other, either from their start or their end positions.

Variant type	Outliers of differentiation			RDA candidates		
	Variants	Nearby genes	Significant enriched GO terms	Variants	Nearby genes	Significant enriched GO terms
SVs	2,079	1,407	108	1,280	940	27
SNPs	71,953	11,578	223	114,637	15,711	528
Indels	6,223	3,285	212	13,871	5,872	292

Figure legends

Figure 1. Overview of **A)** polymorphism detection pipelines used for population-scale characterization of structural variants (SVs), single nucleotide polymorphisms (SNPs) and small indels within the genomes of Romaine (RO) and Puyjalon (PU) salmon (SR: short reads; LR: long reads), **B)** location of the Romaine and Puyjalon rivers (red dot) in Québec, Canada, and **C)** comparative genomics analyses performed on catalogued variants (F_{ST} : fixation index; GO: Gene Ontology; PCA: principal component analysis; RDA: redundancy analysis).

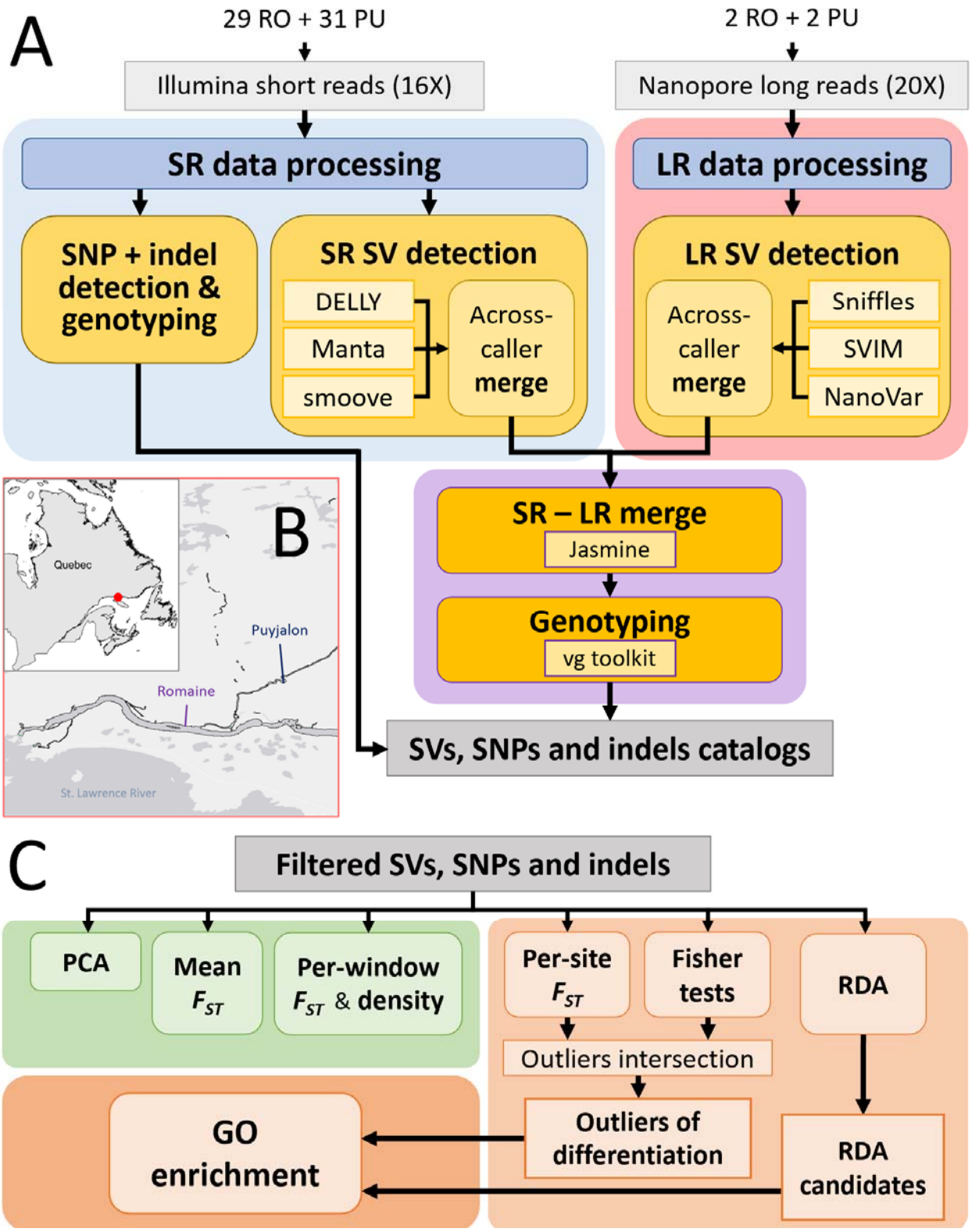
Figure 2. Merged SV count by SV type, length and sequencing platform (SR: short-read; LR: long-read; SR + LR: short- and long-read). These represent putative SVs in the genomes of Romaine and Puyjalon salmon, prior to genotyping (DELs: deletions; DUPs: duplications; INSs: insertions; INVs: inversions).

Figure 3. Proportion of base pairs covered by filtered SVs (top), SNPs (middle) and short indels (bottom) per 100-kb windows along the genome. Each vertical panel represents one chromosome.

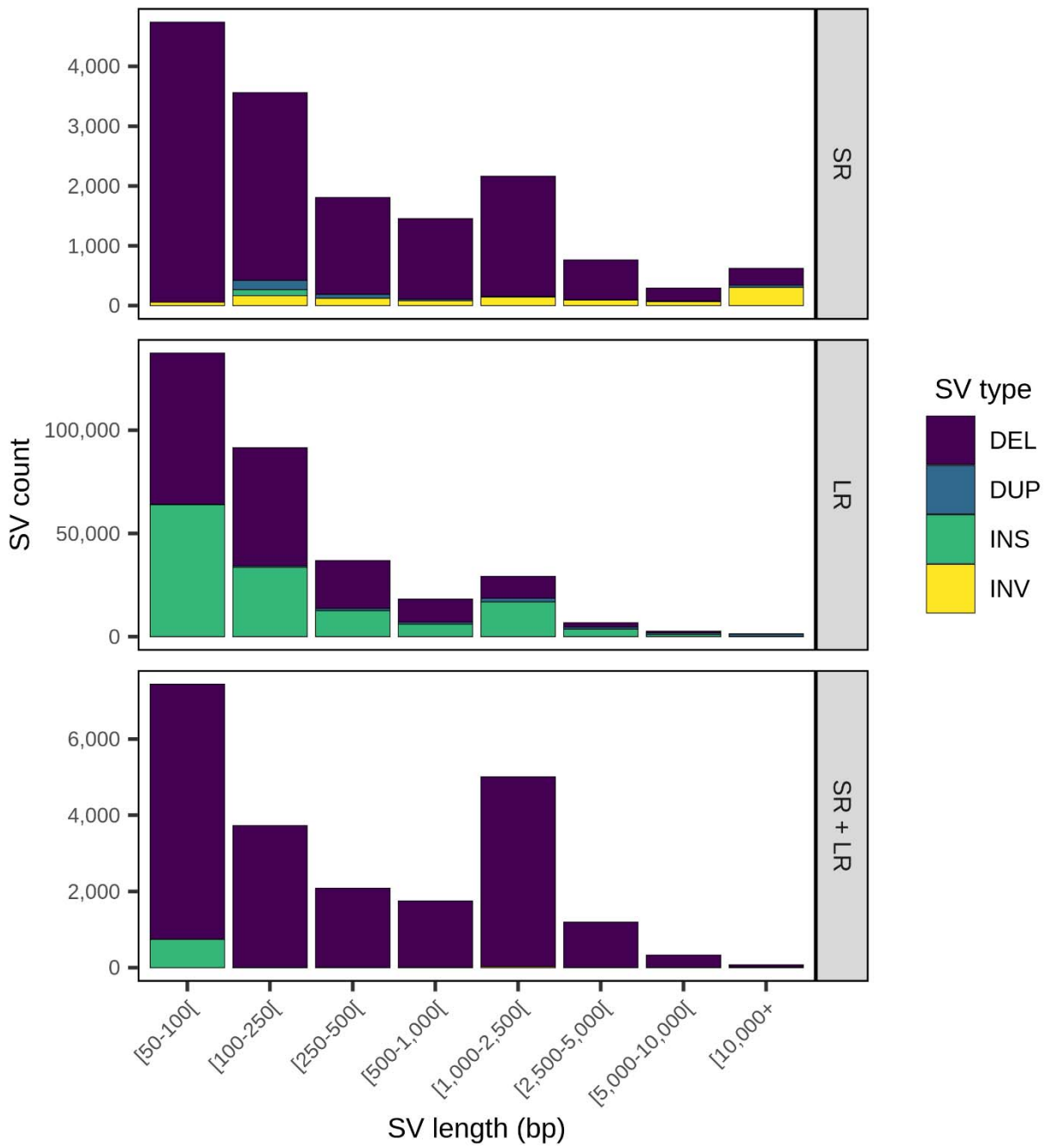
Figure 4. Population structure of Romaine (RO) and Puyjalon (PU) salmon based on principal component analysis from filtered and genotyped **A)** SVs, **B)** SNPs and **C)** short indels. Each point represents one of the 60 salmon sampled for the study.

Figure 5. F_{ST} along the genome between Romaine and Puyjalon salmon, estimated by 100-kb windows (with 10-kb steps) from SVs (top), SNPs (middle) and short indels (bottom). Dotted white lines correspond to the weighted mean genome wide F_{ST} for each variant type. Each vertical panel represents one chromosome.

1223 **Figures**

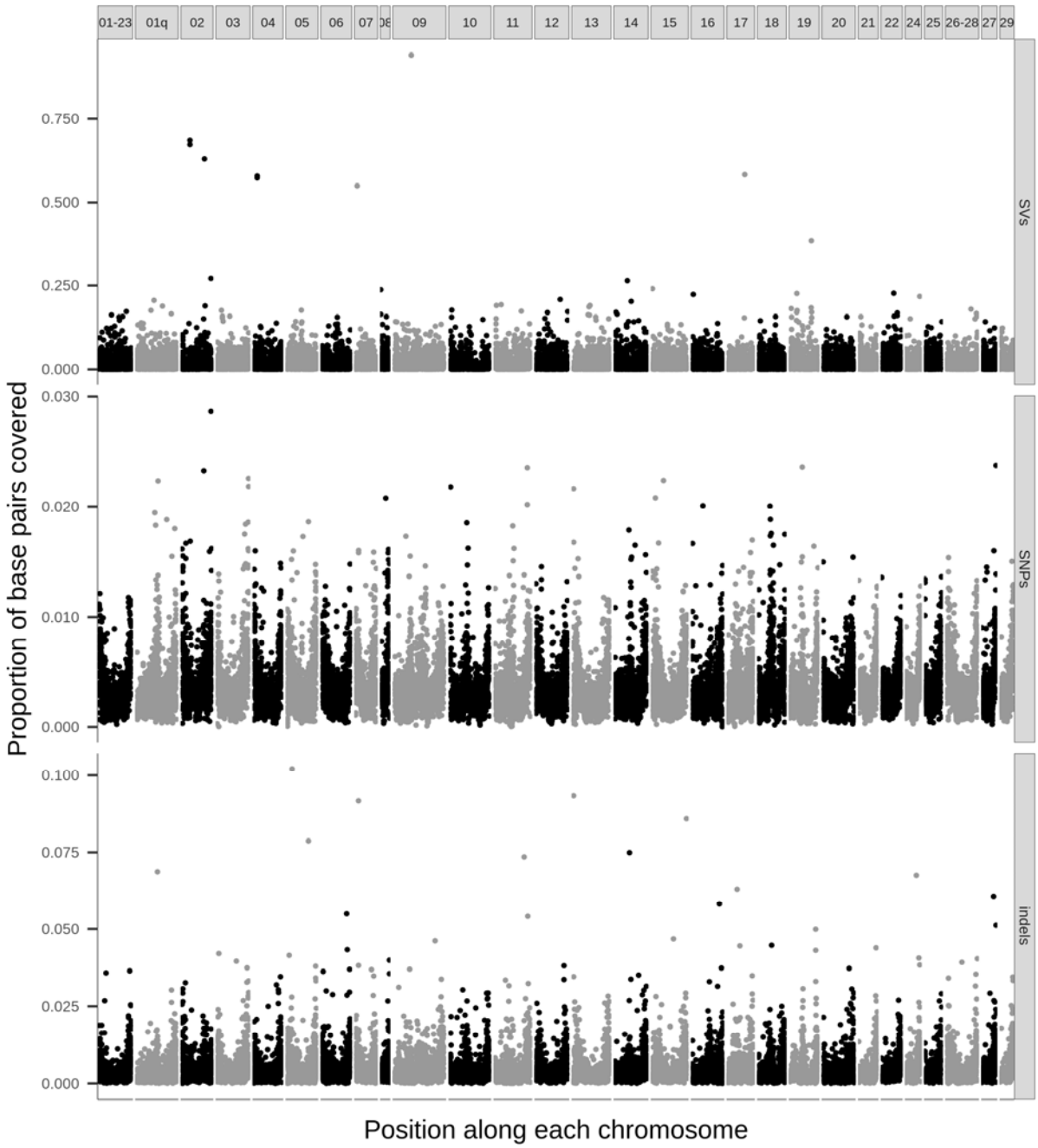


1224

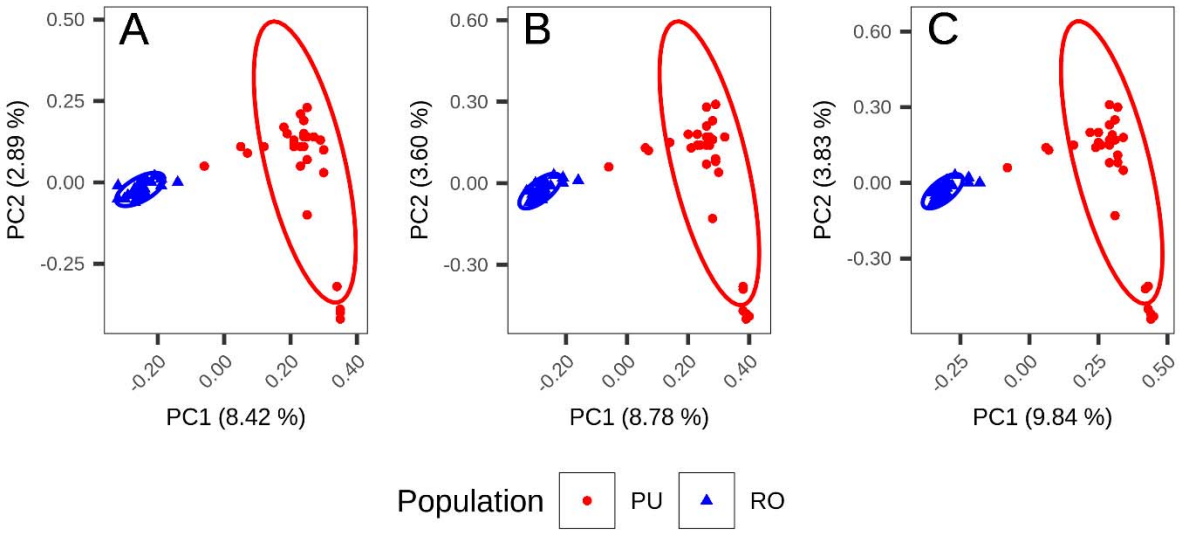

SR
LR
SR + LR
SV count
4,000
3,000
2,000
1,000
0
100,000
50,000
0
6,000
4,000
2,000
0
[50-100[
[100-250[
[250-500[
[500-1,000[
[1,000-2,500[
[2,500-5,000[
[5,000-10,000[
[10,000+
SV length (bp)
SV type
DEL
DUP
INS
INV

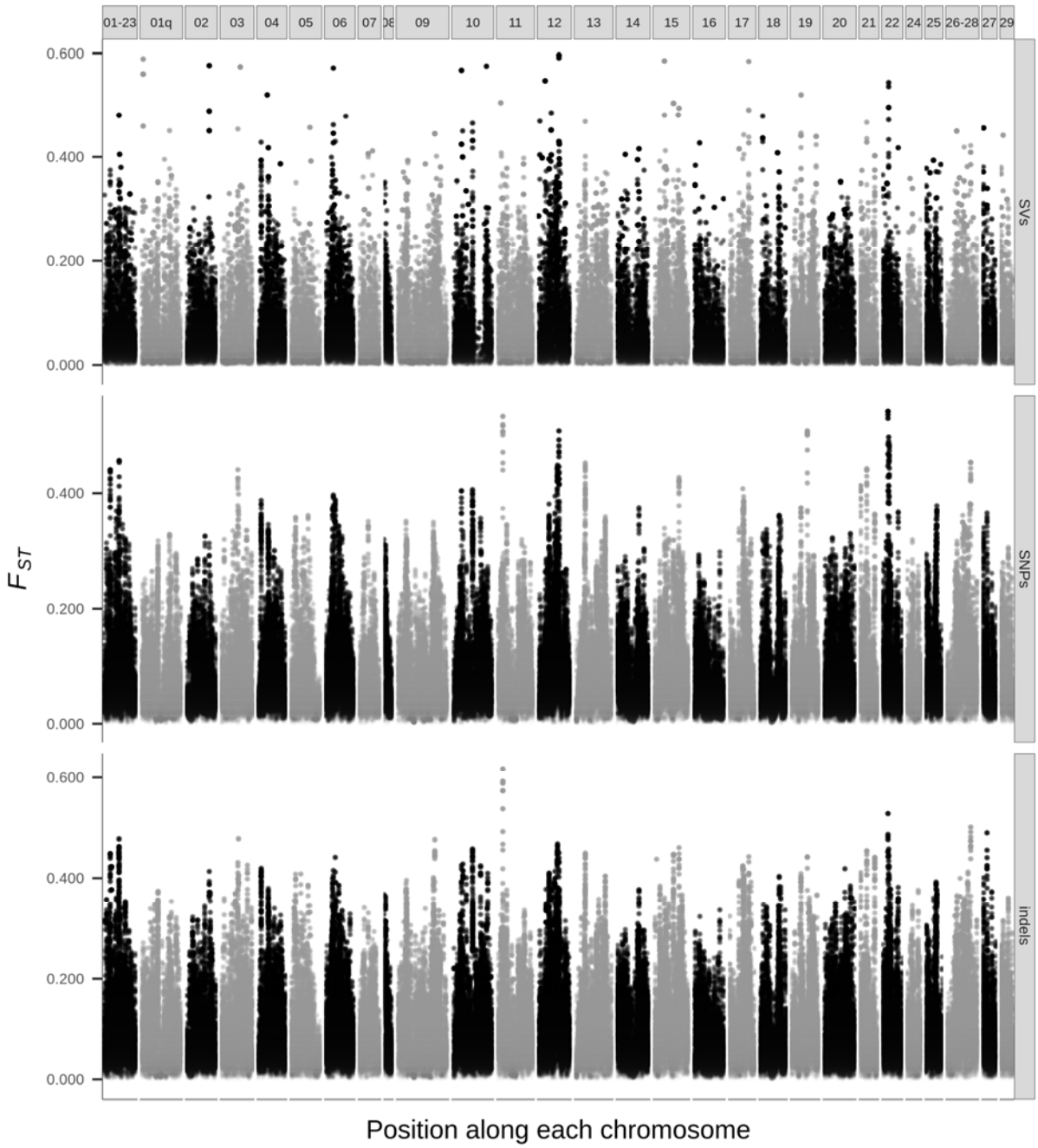
1225

1226


01-23
01q
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
24
25
26-28
27
29
SVs
SNPs
indels
0.750
0.500
0.250
0.000
0.030
0.020
0.010
0.000
0.100
0.075
0.050
0.025
0.000
Proportion of base pairs covered
Position along each chromosome

1227


A
B
C
PC2 (2.89 %)
PC1 (8.42 %)
PC2 (3.60 %)
PC1 (8.78 %)
PC2 (3.83 %)
PC1 (9.84 %)
Population
PU
RO


01-23
01q
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
24
25
26-28
27
29
0.600
0.400
0.200
0.000
SVs
SNPs
indels
F_{ST}
Position along each chromosome

1230

1231