Norwegian University
of Life Sciences

Master's Thesis 2022   60 ECTS
Faculty of Biosciences

Detection and genotyping of Atlantic salmon structural variants with genome graphs

Anna Sofie Kjelstrup
M.Sc. Bioinformatics and Applied Statistics

Acknowledgements

First of all I want to thank my supervisors, Sigbjørn Lien and Kristina Stenløkk for their guidance throughout the work in this thesis. Their support has been invaluable. I would also like to thank Øyvind Sætren Gulbrandsen for assistance in establishing the PRDM9 dataset, and taking the time to discuss and answer any questions. The computing power provided by CIGENE has been critical to the success of this thesis, as well as the open source genome graph community providing tools, documentation and guidance. Finally, I would like to thank my family and Simon who has supported me through this masters degree.

Anna S. Kjelstrup
May 16th 2022

Abstract

Structural variants (SVs) are defined as genomic rearrangements of 50 base pairs (bp) or larger. Although they are less frequent in the genome, they can account for ten folds more variable base pairs than the widely studied singe nucleotide polymorphisms (SNPs). SVs have been hard to detect by short-read sequencing, especially in repeat rich regions. The recent addition of a new reference genome (GCA_905237065.2) and long-read sequencing data for eleven Atlantic salmon individuals has allowed for a more extensive characterization of SVs, revealing a significantly higher count than previously reported. By constructing a genome graph with new high-quality assemblies based on long-reads, we aim to genotype salmon SVs in short-read data, not detectable by traditional methods.

We demonstrate how genome graphs, generated with the bioinformatic pipeline PGGB, can be used to detect and accurately represent SVs in Atlantic salmon genomes. We also present two pipelines for graph-based genotyping using short-reads and discuss alternative metrics for genome graph quality improvement. Eventually, this work will contribute to building a whole genome graph for Atlantic salmon, enabling population scale SV-calling based on already available short-read data.

Sammendrag

Strukturelle varianter (SVer) er definert som genomisk endring på 50 basepar eller mer. Selv om de er i mindretall i genomet, står SVer for mange ganger antallet variable basepar enn de mye studerte enkeltnukleotidpolymorfismer (SNPs). Strukturelle varianter har tidligere vært utfordrende å oppdage ved bruk av eldre teknologi som short-read sekvensering, spesielt i regioner med høyt innhold av repetativt DNA. Et nytt refereanse genom for atlanterhavslaks (GCA_905237065.2), samnt long-read sekvenseringsdata for elleve individer, har åpnet opp for utvidet karakterisering/deteksjon av strukturelle varianter. Dette har avdekket høyere forekomster enn hva som tidligere har blitt rapportert. Ved å konstruere en genomgraf fra nye assemblies av høy kvalitet, basert på long-read sekvenseringsdata, åpner vi for mulighetene til å genotype flere strukturelle varianter med short-read data fra Atlantisk laks.

Vi demonstrerer hvordan det bioinformatiske verktøyet PGGB kan produsere genomgrafer som kan brukes til å detektere og representere strukturelle varianter i atlanterhavslaks. Videre presenterer vi to datastrømmer for grafbasert genotyping ved bruk av short-read data, og diskuterer ulike målbare kvaliteter som kan brukes til å forbedre grafen. Hensikten med dette arbeidet er å bidra til utviklingen av en helgenom graf for atlanterhavslaks som vil muliggjøre SV-calling på populasjonsnivå ved bruk av allerede eksisterende short-read data.

Table of Contents

List of Figures

List of Tables

List of Acronyms

bp	Base pair
FN	False negative
FP	False positive
GFA	Graphical fragment assembly
LR	Long-read
ONT	Oxford Nanopore Technologies
PGGB	Pangenome graph builder
SNP	Single nucleotide polymorphism
SV	Structural variant
TP	True positive
VCF	Variant calling format

1. Introduction

1.1 Structural variants

Genome variation between individuals in a species range in size, from a single base change to large rearrangements that are possible to see in a microscope (Conrad and Hurles, 2007). Single nucleotide polymorphisms (SNPs) can be efficiently typed by high-throughput sequencing and SNP-arrays. SNPs have been extensively used in genome wide association studies to understand the basis of phenotypes of aqua-and agriculture importance, such as tomato flavour (Zhang et al., 2015) and age at maturity of Atlantic salmon (Sinclair-Waters et al., 2020). Linking genomic positions with phenotypes, can identify candidate pathways and lead to a greater understanding of the underlying biological mechanisms of a trait (Alkan et al., 2011). Genotype of a sample is inferred through mapping reads to a reference genome, or with SNP-arrays, and the results are presented in the variant calling format (VCF). The introduction of SNP-arrays decreased costs and enabled genome wide association studies with a large number of samples, but is limited to genotype known variants in the genome.

Structural variants (SVs) is another class of genome variation. They are diverse in type and size, ranging from 50 to thousands of base pairs and include different types of sequence rearrangements (Figure 1.1) like insertions, deletions, inversions, translocations and duplications (Mahmoud et al., 2019). Although SVs are fewer in numbers, they affect more bases than SNPs. About 30 000 SVs are expected in any human genome, but this type of variation is still not well understood due to limitations in sequencing technology (Ho et al., 2020).

Structural variants can disrupt functional elements of the genome, ultimately affecting the phenotype. Rearrangements in regulatory elements and copy number variants can have an impact on gene expression and dosage, while deletions can cause gene truncation or fusion (Mahmoud et al., 2019). SVs have been studied in a number of teleost species of interest to the aquaculture industry. Hundreds of high confidence SVs found in domesticated rainbow trout were identified as exon loss variants or gene fusion variants (Liu et al., 2021), and in lake whitefish, SVs are suggested to contribute to speciation

(Mérot et al., 2022). SVs have also been identified in plants, where they have been associated with traits like tolerance, fruit yield and quality (Mahmoud et al., 2019).

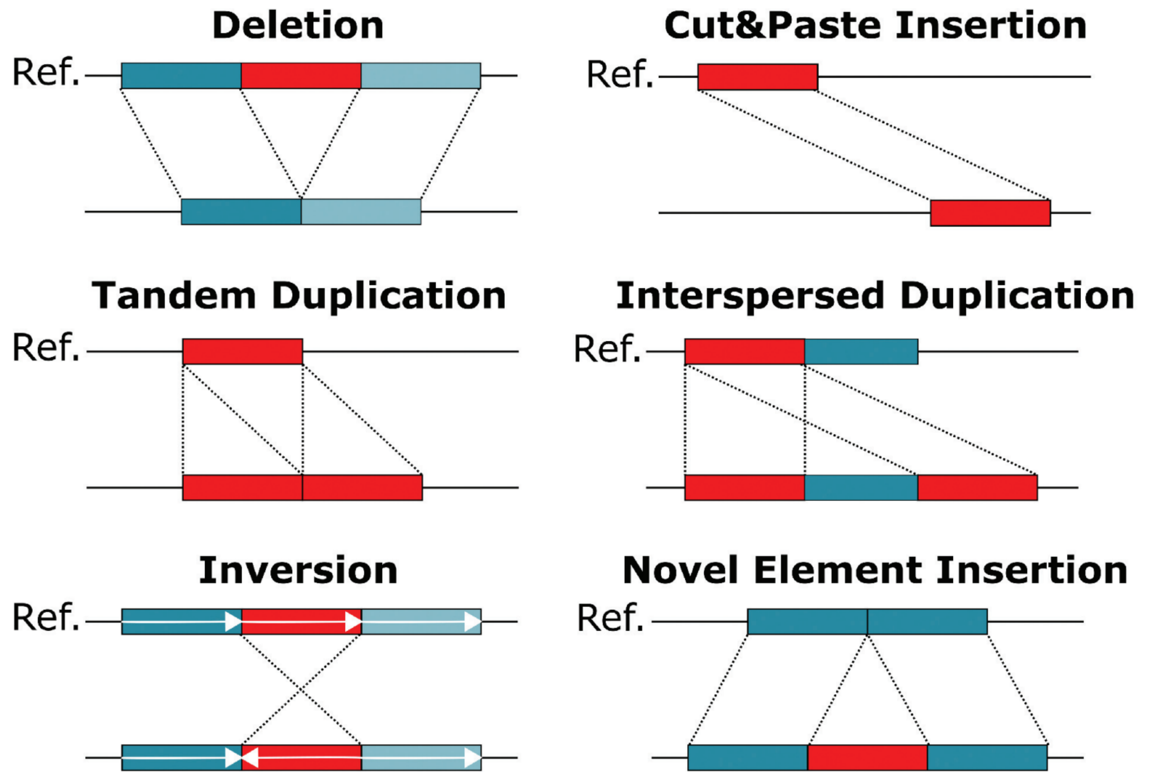


Figure 1.1: Overview of different classes of SVs Different types of SVs include deletion, tandem duplication, inversion, cut and paste insertion (translocation), interspersed duplication and novel element insertion. Here shown in separate genomes, compared to a reference. The rearranged sequence is colored red. Figure from Heller and Vingron, 2019

Repetitive DNA is an important source of SVs and make up large parts of eukaryote genomes (de Koning et al., 2011). There are two main categories of repetitive DNA, transposable elements (TEs) and tandem repeats (TRs). TEs are mobile elements of DNA, often classified by the mechanism of transposition, which can be either cut and paste, known as transposons or copy-paste, known as retrotransposons (Hartley and O'Neill, 2019; Bourque et al., 2018). TRs are defined as adjacently repeated stretches of DNA, where the length of the repeated unit (array size) and sequence composition vary greatly (Lu et al., 2021; Sulovari Arvis et al., 2019). Their prevalence often differ between chromosome regions (Hartley and O'Neill, 2019). TRs are believed to contribute to structural variants through polymerase slipping (Raz et al., 2019), tandem duplications (Farnoud et al., 2019) and template switching (Course et al., 2020). Reconstructions of full genome assemblies based on sequencing data are often flawed in TR-regions due to limitations of short-read sequencing data to read through TR-arrays (Tørresen et al., 2019).

The read length of high-throughput short-read sequencing technologies has also been a limiting factor for reliable SV-detection and genotyping. When the sequencing reads are shorter than the variant length, mapping to a linear reference genome will be a challenge. Genotypes are inferred through patterns found in mapped reads, like split read alignments or discordant read pairs (Wang et al., 2022). Poor mapping will make it difficult to distinguish the different types of SVs (Mahmoud et al., 2019). Thus far, no bioinformatic tool has been able to detect all SV-types and sizes reliably (Mahmoud et al., 2019). Mid to large size insertions are particularly challenging to identify. The fraction of all SVs detected, known as recall, is expected to be between 10% and 70%, while the false positive rate is reported to be as high as 89 % (Mahmoud et al., 2019).

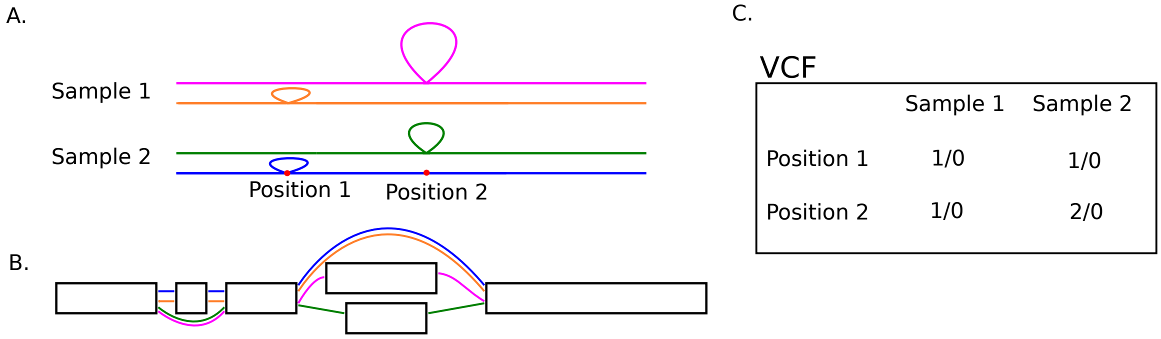
The development of long-read sequencing technologies has exceedingly improved SV-calling. The prominent technologies from PacBio and Oxford Nanopore Technologies (ONT), generate kilobases long reads. Long-read data makes genome assembly and read mapping easier, which in turn has improved SV-detection (Sedlazeck et al., 2018). Mapping patterns are easier to distinguish as the reads cover full variants and flanking sequence. The count of SVs detected by the use of long-read data has been reported to be in the range of 2 to 8.33 times the count found with short-read data, depending on the organism (Mahmoud et al., 2019).

As long-read sequencing has made reference quality assemblies more obtainable, we expect to see an increase in collections of genome assemblies, referred to as pangenomes (Eizenga et al., 2020). This creates a need for methods to compare variation between a large number of assemblies.

1.2 Genome graphs

Genome graphs are data structures well suited for representing pangenomes, and they are suggested as a variant aware alternative to linear reference genomes (Eizenga et al., 2020). Graph-based SV genotyping makes it possible to utilize existing short-read data to infer variants undetectable using a linear reference (Li et al., 2020; Hickey et al., 2020). As a result, it will be feasible to carry out population scale SV-studies without resequencing using costly long-read technology. Another major motivation for graph-based genotyping is the decreased reference bias, as proven in Garrison et al., 2018. Reads containing alternative alleles are less likely to be mapped to a linear reference than a read with a reference allele, which may create a bias towards calling reference allele (Brandt et al., 2015; Sirén Jouni et al., 2022; Martiniano et al., 2020). The human pangenome consortium estimates that more than 70% of SVs have been undetected in SV-studies due reference bias and short-read limitation. They aim to improve detection of structural variants through the construction of a human genome reference graph, based on alignment of long-read based assemblies (Wang et al., 2022).

There are two main approaches to constructing a genome graph. The most common method is to base the graph on a linear reference genome, often referred to as the backbone of the graph. Subsequently, known variants are added from a VCF file, resulting in a directed acyclic graph (DAG). GraphTyper2 (Eggertsson et al., 2019) and Paragraph (Chen et al., 2019) are examples of graph based short-read genotypers using the VCF-approach. The graph handling toolkit vg (Hickey et al., 2020) will also build graphs with VCF files and linear reference sequence. The second approach, as applied by the pangenome graph builder (PGGB)(Garrison et al., 2021) and minigraph (Li et al., 2020) is alignment based graphs, where *de novo* assemblies are aligned, and identical sequence is collapsed into nodes (Figure 1.2). Variants between the sequences in the graph, often referred to as bubbles due to the shape they form, are detected during graph construction and can be extracted in the variant call format (VCF)(Figure 1.2 C). This approach is made possible because genome assemblies have become obtainable with the rise of long-read sequencing technology. Multiple sequence alignment is challenging and computationally expensive, but Minigraph and PGGB have solved this by making their own alignment algorithms that are specialized for graph construction.



	Sample 1	Sample 2
Position 1	1/0	1/0
Position 2	1/0	2/0

Figure 1.2: Genome graph representation of SVs A. The four colored lines represent aligned haplotype-resolved assemblies from two samples. **B.** Simplified visualisation of genome graph for the four sequences. Identical sequences are collapsed into nodes (the squares), and each sample is represented as paths linking the nodes (colored lines). **C.** Corresponding variant calls in a simplified version of a VCF, with the blue sequence as reference. For every position of variation (bubble) in the graph, the genotype inferred from the graph is recorded for each sample. There is one call for each haplotype, separated by a slash. The genotypes are encoded, 0 for reference call, other numbers are referring to alternative allele number. The figure is made from figures in Garrison and Guarracino, 2022 and Ebler et al., 2022

.

VCF-based graphs have proved to reliably genotype high confidence SVs, but have limited SV-representation. Eggertsson et al., 2019 reported to have improved genotyping sensitivity with GraphTyper2 compared to linear reference SV genotyping. However, GraphTyper2's data structure was identified as unable to represent a full pangenome, because it could not represent complex structures like nested variants. In addition, Paragraph and GraphTyper2 are dependent on an accurate breakpoint sequence for the set of known SVs to be genotyped accuratly (Chen et al., 2019; Eggertsson et al., 2019). Recall has as been shown to decrease with shifts in the breakpoint. Another VCF-approach from the vg toolkit (Garrison et al., 2018), has a more extensive set of graph handling tools and a different data structure, making it possible to include complex structures (Hickey et al., 2020). Vg is not as affected by breakpoint inaccuracy, but was outperformed by alignment based graphs using short-read mapping identity as a metric (Hickey et al., 2020).

Alignment based graph construction tools are still under active development. Minigraph is the most established of these tools, and has been used to create bovine and whitefish genome graphs. Crysnanto Danang et al., 2021 discovered novel functional sequence by creating a bovine pangenome graph and identified a large percentage of the variants in multiallelic bubbles. Mérot et al., 2022 identified insertions and deletions linked to TEs as a key component of divergence between two whitefish species. Both studies utilized short-reads to do graph-based genotyping. Minigraph is fast, but does not perform base-level alignment and will not include variarion <50 bp. Without base level alignment, minigraph will have trouble with aligning sequences in TR-regions (Li et al., 2020). PGGB will do base level alignment and include smaller variants in addition to SVs. As a recently developed tool, there are yet to be any published results, but the human Pangenome consortium are using PGGB to construct a human pangenome, focusing on inclusion of diversity, made for research and medical application (Wang et al., 2022).

Multiple specialized graph-tools for downstream analysis of genome graphs have been developed in recent years. Through a collaborative effort, a standardized format for assembly graphs has been developed (GFA group, 2022). A variation of the graphical fragment assembly (GFA) format, as first suggested by Li et al., 2020 is now accepted or under implementation for most graph tools. Odgi is the most comprehensive toolkit for analyzing giga base scale genome graphs efficiently (Guarracino et al., 2022), including tools for detecting complex regions, extracting regions of interest, exploratory analysis, manipulation, validation, and visualization. Historically, graph-based genotyping has been time consuming due to slow read mapping. Newly developed giraffe from the vg toolkit has made significant headway on this problem, and has reported read mapping runtimes close to that of a linear reference (Garrison et al., 2018; Sirén Jouni et al., 2022). Another genotyper of note is the PanGenie tool, which claim to improve genotyping by k-mer based methods (Ebler et al., 2020).

1.3 Analyzing structural variants in the Atlantic salmon genome

Atlantic salmon (*Salmo salar*) is an anadromous fish species of high economical, cultural and ecological importance. Aquaculture species like the Atlantic salmon are early in domestication and genetic improvement toward increased quality of production has motivated a great number of genome wide association studies (Houston et al., 2020). Atlantic salmon also presents an opportunity to study vertebra genome evolution, as it is undergoing rediploidization after a salmonid specific whole genome duplication (WGD) 80-120 million years ago (Lien et al., 2016; Gundappa et al., 2022). Duplicates with 75-100 percent identity exists for at least half of the protein coding genes. There is large genomic divergence between European and North American Atlantic salmon populations, including distinct karyotypes (Brenna-Hansen et al., 2012). The most recent reference genome for Atlantic salmon (GCA_905237065.2) adds up to 2.76 giga base pairs (Stenløkk et al., 2022). The content of repeated DNA based on the previous reference genome was estimated to be 58-60%, which is one of the highest found in any vertebra (Lien et al., 2016).

Population scale detection of SVs based on short-read data was reported by Bertolotti et al., 2020 based on the previous reference sequence ICSASG_V2 assembly (NCBI accession GCA_000233375; Lien et al., 2016). Lumpy, a probabilistic tool was used for SV-detection (Layer et al., 2014) in addition to manual curation to ensure only high confidence SVs were kept. The manual curation filtered out all insertions, as they were not possible to confirm by visualization of mapped reads, which was an important step of the curation process. The final set of high confidence variants were of 15,483 unique SVs called for 492 individuals. More than 90 % of the total number of SVs were deletions, being the easiest SV-type to detect and confirm during curation. As much as 1432-1436 of the deletions were caused by a recently active transposon. False positive rate was estimated to be 0.91. In conclusion, the combination of high repeat content, homologous regions with high sequence similarity and large diversity between phylogeographic groups, makes it challenging to call SVs in Atlantic salmon with short-reads.

Recently, eleven long-read based *de novo* assemblies were generated as the first step towards making an Atlantic salmon pangenome resource. Long-read technology data read through repeat regions, making repeat regions accessible for genome analysis. The data incorporates one individual sampled from aquaculture (Simon) and ten individuals from wild populations in North America and Europe (Figure 1.3). The new Atlantic salmon reference genome sequence Ssal_v3.1 (GCA_v905237065.2) has been generated from an aquaculture strain (AquaGen) named Simon. The estimate of repeat content was increased to 60-70 % in the new reference sequence. TEs made up 40.61 % of the assembly, and 11% of all base pairs in the genome was identified as Tc1-mariner elements. TRs were estimated to 34 % of the genome, and was found to be enriched in telomeres. The other ten individuals represent the four different phylogeographical groups, North American, Baltic, Barents/White Sea and Atlantic (Stenløkk et al., 2022). The motivation for sampling from different phylogeographical groups was to include as much diversity as possible to build a pangenomic resource.

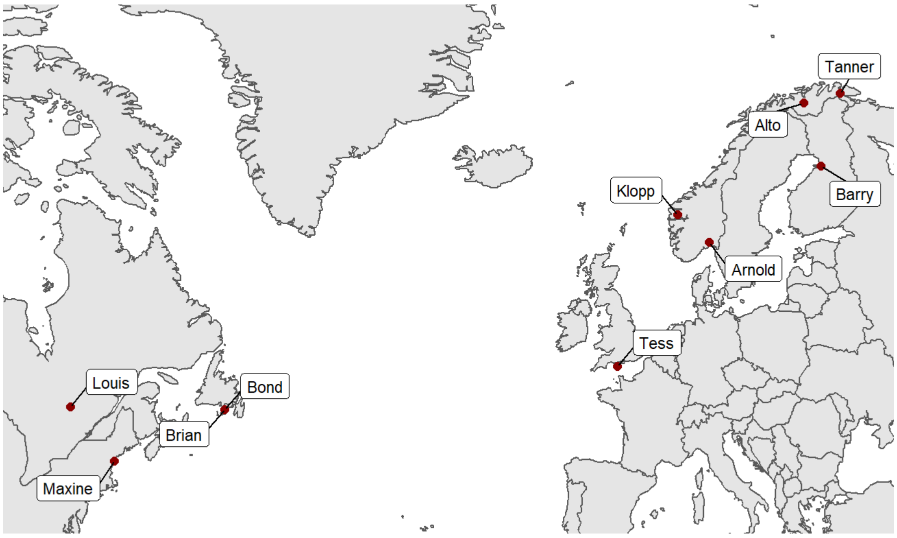


Figure 1.3: Map showing origin of long-read sequenced Atlantic salmon Wild salmon was sampled from four phylogeographic groups; North American (Louis, Bond, Brian and Maxine), Baltic (Barry), Barents/White Sea (Tanner and Alto) and Atlantic (Klopp, Arnold and Tess). The aquaculture individual used to build Ssal_v3.1 (Simon) is not shown in the map, but it origins mainly from the Atlantic phylogeographic group. The map is based on data from Stenløkk et al., 2022

The pangenome has been used to call high confidence SVs with three different tools, sniffles, SVIM and NanoVar (Stenløkk et al., 2022), keeping only SVs detected by multiple tools to ensure a higher precision. SV-calling was done by continent, as there is expected to be high genetic divergence between the European and North American Atlantic salmon. Simon was used as reference for Europe, while Brian was used for the North-American individuals. The resulting set of SVs include more than 700 000 SVs detected in European Atlantic salmon, and more than 300 000 detecten in the North American samples. The SVs consists of 59.89 % deletions, 39.72 % insertions, 0.17 % inversions and 0.22 % duplications. This pattern is shared with TRs, and despite TRs only covering 34 % of the genome, more than 80 % of deletion base pairs overlap TRs. The SV distribution in the genome showed a striking enrichment in telomeric regions (Monsen et al., 2022).

1.3.1 Benchmarking SV detection and genotyping in Atlantic salmon genome

Generating a relevant dataset with known variants is useful for evaluating the performance of variant calling tools. Ideally, one would want to use a benchmarking dataset like the one created for human (Zook et al., 2020). Unfortunately, no such dataset is available for Atlantic salmon. Acquiring a benchmark is costly and will not be possible for most non-model organisms. An alternative approach will be to simulate data or select small regions suited for validating the results manually. It is important to keep in mind that simulated data will be a simplification of real data.

The following genomic regions in Atlantic salmon are identified as representative of the species with regard to the characteristics described earlier, which makes them ideal for testing new data analysis pipelines. Chromosome 22 contains telomeric enrichment of TRs, as well as regions with less TRs. This makes the chromosome a well suited dataset when investigating the impact of TRs on SV-detection. The zinc finger region of the PRDM9 gene was selected as a biological interesting, small and structurally complex region suitable for testing. PRDM9 is a DNA binding protein, defining meiotic recombination sites, and has been linked to speciation (Grey et al., 2018). In the last exon of the Atlantic salmon PRDM9 copy on chromosome 5, there is a variable array of zinc fingers (Figure 1.4). A single repeat may vary slightly in composition, and as the zinc finger is the DNA-binding part of the protein, polymorphims will affect the position and affinity of the binding sites, thus the recombination frequency can be changed (Grey et al., 2018).

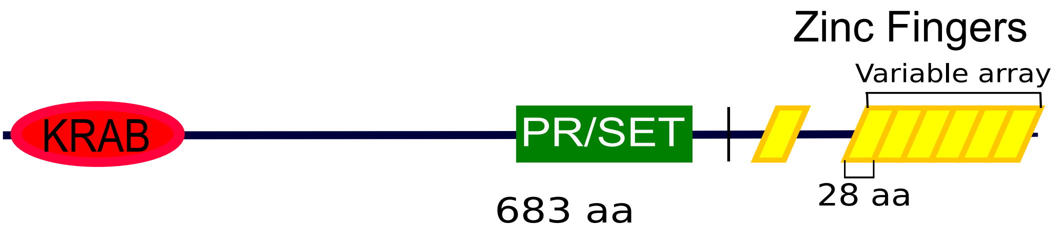


Figure 1.4: Structure of the functional PRDM9 gene found in chromosome 5 in the Atlantic salmon genome The full gene spans 683 amino acids and is well conserved between individuals with the exception of the variable zinc finger array. This figure specifically shows variants found in Simon (GCA_905237065.2) allele 1 with six repeats. The figure is adapted from Guldbrandsen, manuscript in preparation

1.4 Aims of the study

In this study, we will explore genome graph construction with PGGB, and compare two graph based genotyping pipelines. One pipeline includes short-read graph mapper giraffe (Sirén Jouni et al., 2022), while the other is a k-mer based genotyper called PanGenie (Ebler et al., 2020) (Figure 1.5 A). We will evaluate how repeats influence PGGBs performance, and we will create a simulated sequence with known SVs in order to validate the variants detected.

Next, we want to extend the pipeline to using real data, constructing a graph on assemblies and using real reads by focusing on the zinc finger array of a functional PRDM9 gene on Atlantic salmon chromosome 5 (Figure 1.5 B). Graph-based SV genotyping has already been proven to be better than traditional linear based approaches, (Hickey et al., 2020; Sirén Jouni et al., 2022; Ebler et al., 2022) but we want investigate if it works for the complex genome of Atlantic salmon.

The aim is to evaluate the feasibility of making a Atlantic salmon whole genome graph, with the interest of using it for population SV studies. For the pipelines to be of use, we have to establish an approach which satisfies the following criteria:

- The graph must be able to detect and represent SVs.
- The graph-based genotying pipeline must reliably call SVs with short-reads.

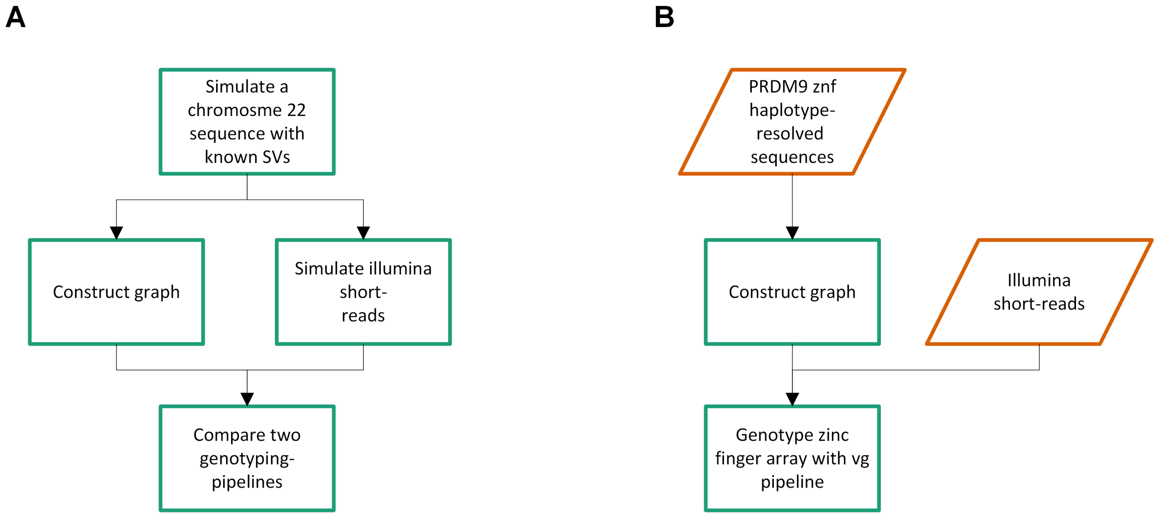

A
Simulate a chromosme 22 sequence with known SVs
Construct graph
Simulate illumina short-reads
Compare two genotyping-pipelines
B
PRDM9 znf haplotype-resolved sequences
Construct graph
Illumina short-reads
Genotype zinc finger array with vg pipeline

Figure 1.5: Overview of datasets and bioinformatic pipelines used to detect SVs in the Atlantic salmon genome Green boxes represent a process, orange represent data. **A:** shows an overview of the pipeline for evaluating PGGB and the genotypers with simulated data. **B:** PGGB and vg toolkit is will be used genotype PRDM9 zinc finger on chromosome 5 with real sequencing data and assemblies

2. Methods

2.1 Simulating Atlantic salmon chromosome 22 data

From the approximately 65 Mbp long chr. 22, two 10 Mbp regions were selected based on the repeat density of the reference genome. One region with a low count of TRs, and another with a high count of TRs. SVs from the long-read detected catalog (Stenløkk et al., 2022) were used to simulate a sequence from chr. 22. To decrease the count of SVs and ensure true variants, we opted to only keep SVs detected with at least 3 tools and found in 2 individuals. Three of the SV were removed because they overlapped with other variants. The remaining SVs were inserted into the reference sequence of chr. 22 using VISOR (Bolognini et al., 2020) in order to make a new sequence with SVs inserted into the two regions.

In order to asses the two different graph-based genotyping pipelines, we simulated reads from the new sequence (Figure 2.3). ART (Huang et al., 2012) was used to simulate short paired-end reads based on the original chr. 22 sequence and the simulated sequence. When genotyping this set of reads, we expect heterozygous callings for all variants. Parameters were chosen to resemble 150 bp Illumina reads with approximately 400 bp in fragment size. Simulation was carried out for multiple levels of read depth.

2.2 PRDM9 dataset

The PRDM9 zinc finger region of the assemblies had previously been manually phased, resulting in a total of 12 haplotype-resolved sequences. The array showed between five and eight zinc fingers in these sequences (table 2.1). In order to align the sequences better to the full chromosome 22, each haplotype sequence was extended with 5 kbp of reference sequence on each end.

Table 2.1: Number of repeats in the zinc finger array for each haplotype sequence Four individuals were successfully phased, Simon, Klopp, Arnold and Maxine. The remaining individuals are represented by one haplotype. Based on data from Gulbrandsen, in preparation

Name	Haplotype	Number of repats in znf-array
Simon	1	6
Simon	2	5
Klopp	1	6
Klopp	2	8
Arnold	1	6
Arnold	2	7
Alto	1	6
Tanner	1	8
Maxine	1	8
Maxine	2	6
Brian	1	6
Bond	1	6

2.3 Graph construction

2.3.1 PGGB - the pangenome graph builder

The pangenome graph builder (PGGB) is a three-step pipeline for making alignment-based genome graphs. A pangenome refers to a collection of all genomic sequences found within a species, population, clade or metagenome (Eizenga et al., 2020). A pangenome graph will be a graphical model to represent them. Pangenome graphs and genome graphs are terms used interchangeably. In this thesis, we are using the term genome graphs for simplicity. PGGB will include all kinds of variants into the graph, including variants < 50 base pairs long. However, this study will only focus on the structural variants.

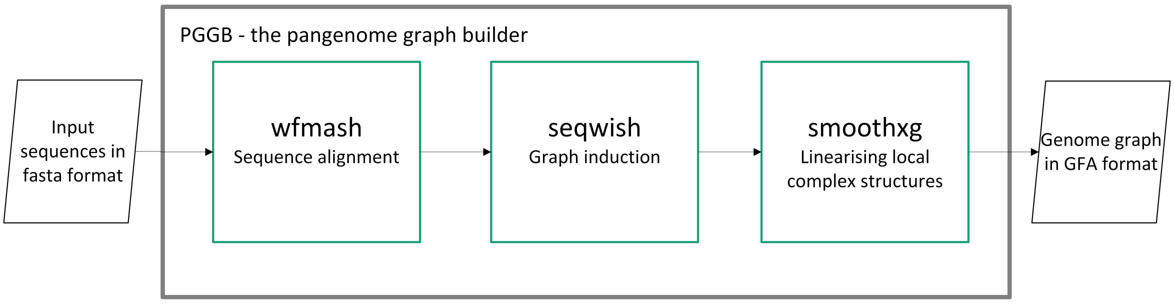

PGGB - the pangenome graph builder
Input sequences in fasta format
wfmash
Sequence alignment
seqwish
Graph induction
smoothxg
Linearising local complex structures
Genome graph in GFA format

Figure 2.1: An overview of the pangenome graph builder pipeline (PGGB) Three tools are developed specifically for the purpose of constructing genome graphs. Input to the pipeline is a fasta file with assemblies, and the output is a genome graph. Wfmash will align the input sequences, seqwish induces the graph before smoothxg ensures local linearity in the final step.

The first step of the pipeline is a pairwise sequence alignment of the input sequences with wfmash (Figure 2.1). The queries are divided into non-overlapping segments, which are then mapped to the other sequence in the pair using a version of mashmap. Only mappings with identity over a certain threshold will be kept, thus the segment size parameter works as a minimum alignment filter. These approximate mappings will be used as a target for alignment with wflign. Wfmash is quick and conserves synteny while also being able to ensure base level alignment (Garrison, 2022b).

Graph induction using seqwish is the second step of the PGGB pipeline. An alignment graph is built using the output from wfmash, before collapsing the nodes into a variantion graph (Garrison and Guarracino, 2022) (Figure 1.2 A and B).

The last step of the pipeline is smoothxg. This tool will perform partial order alignment (PoA) on blocks in the graph, ensuring local linearity. It is the most computational expensive step in the pipeline, but very important as it decreases the complexity of the graph, making it possible to use for down stream analysis such as graph-based read mapping (Garrison, 2022b).

2.3.2 Graph evaluation

The graph quality was assessed by the count of SVs represented in the graph. This metric was used for parameter tuning and to evaluate PGGB's ability to detect SVs from *de novo* assemblies. Variants found in the graph were extracted into a VCF with vg deconstruct and compared to the truth set of SVs. True positives (TP) were defined as the variants in the graph with start and end position ±60 bp from any of the original variants inserted into the simulated sequence. False positive (FP) was a variant detected by PGGB but not found in the SV-catalog. False negative was the SVs in the simulated squence not detected in the graph. A script was written with functions for comparing positions and finding recall, precision and F1 score. A metric taking both recall and precision into consideration as defined in equation 2.1-2.3. https://github.com/ankjelst/SalmonGraph/blob/main/scripts/rscripts/metrics.R.

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.3}$$

2.3.3 Parameter tuning

There are three parameters required for running PGGB version 0.2.0, -s, -n, and -p. The number of mappings reported for each segment (-n) was set to the number of sequences in the input fasta as recommended by the developers. The percent identity, (-p) was decided by the mash distances between sequences within the chromosome, selecting the largest distance in this equation $100 - (maxdivergence \times 100)$. Percent identity (-p) should be set to this value or lower. The last required parameter, segment size (-s) will require some tuning. The segment size is the minimum size of the sequences for the approximate mapping of wfmash, and works as a minimum alignment filter. If you set this too low, you will risk keeping small matches caused by repeats from two different regions. If you set it to high, your sequences will not be aligned. PGGB was run with different -s values. F1 score, clock time and memory consumption was recorded for each run.

Two optional parameters were adjusted. Minimum match length (-k) was set to 311, based on recommendations for human data. This value was constant for all runs in this study. The -k parameter will filter out small matches in the graph induction step of seqwish. The second optional parameter adjustment was block size (-G), which is relevant in smoothxg. This parameter was selected for tuning as it is stated to have a large impact on memory and time usage of the pipeline (Garrison et al., 2022).

Two values are passed to -G as default, resulting in two rounds of graph refinement. -s was set to 100 000 like in the human genome for -G tuning. All the other parameters were left to default. Be aware of differences in defaults values and required input between versions of PGGB.

2.4 Genotyping

For down-stream analysis, we want to use the graph to genotype SVs for individuals where only short-read data is available. Two pipelines were tested for graph-based SV-genotyping, representing two different approaches 2.3. The vg pipeline represent a read mapping approach, including Giraffe (Sirén Jouni et al., 2022). PanGenie is a k-mer based genotyper, an approach known to be quicker then a mapping approach, but has been unreliable in repeats and duplicate regions. PanGenie address this vulnerability by inferring genotypes based on neighbouring variants when it is not possible to tell genotype based on unique k-mers (Ebler et al., 2022). This is why the tool requires haplotype-resolved sequences. In addition, it is only applicable to diploid organisms.

The two pipelines presented some requirements which had to be considered when constructing the input graph. To meet PanGenies requirement of at least two samples with haplotype information, the input fasta was made up of three copies of the original sequence and one of the simulated sequence (Figure 2.2), representing two samples by using the PanSN-spec convention (Garrison, 2022a). PGGB parameters were set to -s 50 000 and -G 7919,8069.

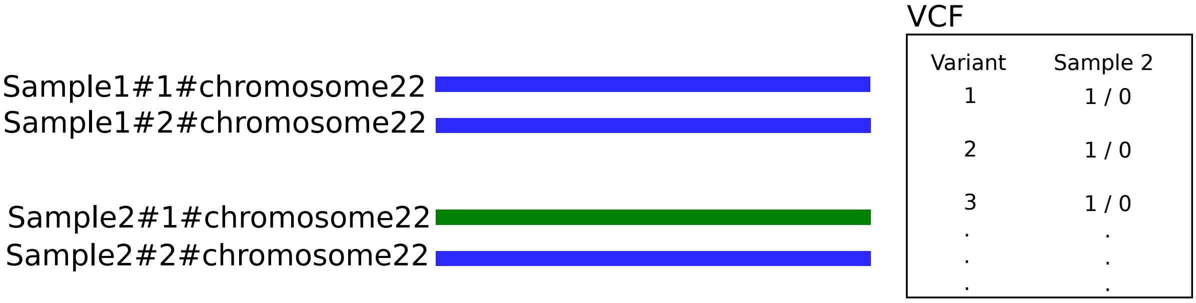


Figure 2.2: Input for construction of chr. 22 graph The green line represents the simulated sequence, and the blue lines represent the original reference chr. 22. The simplified VCF visualisation shows expected genotype for sample 2, which will be 1/0 for every single variant.

Giraffe has not been tested on PGGB-graphs by the developers, and it did require some work-arounds to run successfully (Novak, 2022). The giraffe preprocessing had to be done manually in 4 steps, as the graph lost all path lines when running the autoindex tool. For the manual preprocessing to work, we had to chop our graph where nodes exceeded 1024 bp. This step was preferential to do manually as opposed to letting the preprossessing tools chop the graph, as it allowed us to retain control over the coordinates. After mapping reads to the graph, the vg call function was used to call genotypes. This step required a preprocesing of vg pack. The same set of reads were used as input into PanGenie, as well as the VCF from vg deconstruct as PanGenie requires specification of the sites to call. The full pipeline from simulation to genotyping is presented in Figure 2.3.

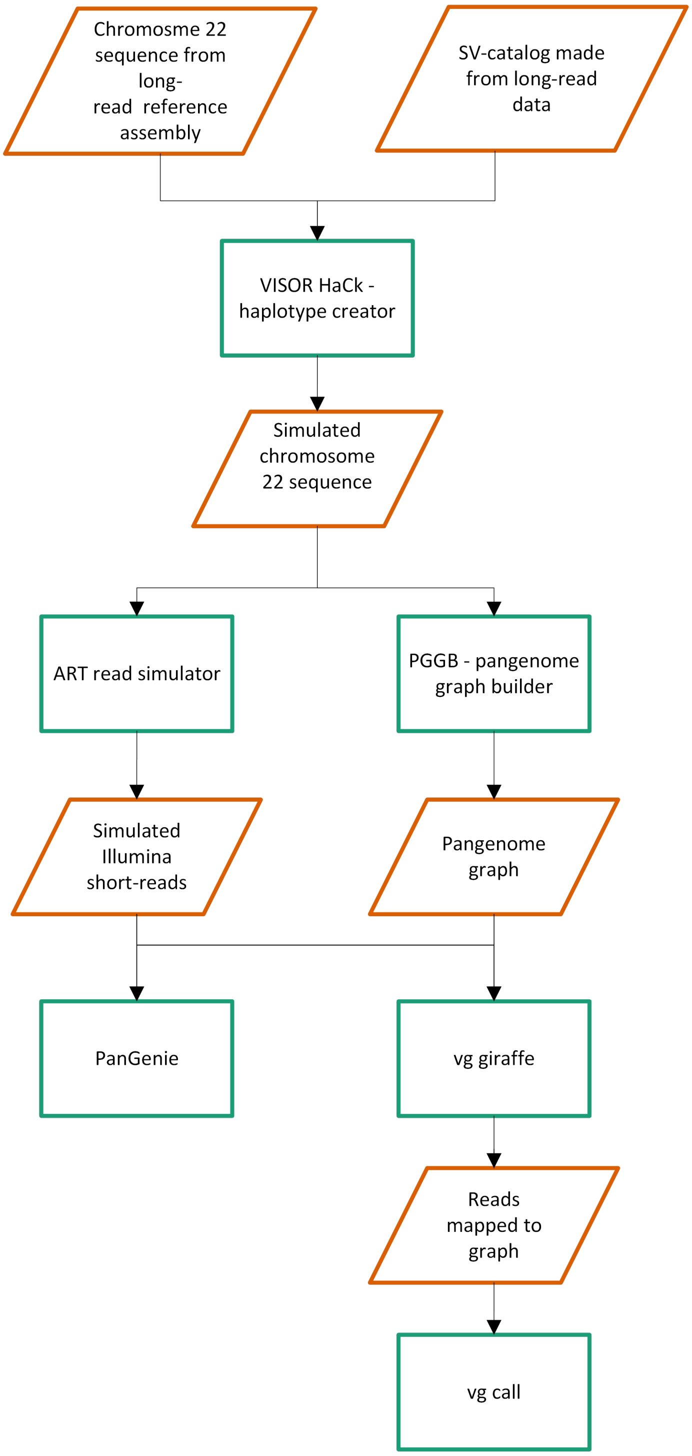

Chromosme 22 sequence from long-read reference assembly
SV-catalog made from long-read data
VISOR HaCk - haplotype creator
Simulated chromosome 22 sequence
ART read simulator
PGGB - pangenome graph builder
Simulated Illumina short-reads
Pangenome graph
PanGenie
vg giraffe
Reads mapped to graph
vg call

Figure 2.3: Overview of the full pipeline for the chr. 22 based graph construction and down-stream analysis An overview of all the tools (in green), input and output data (in orange) from each step in the pipeline for construction and down stream analysis of the simulated data based on chromosomes 22. Simulation of a sequence with known variation, based on chromosome 22 was carried out by inserting SVs from a SV-catalog with VISOR. The simulated sequence and the original reference was the input to PGGB (Figure 2.2), which generated a genome graph. Reads were simulated with ART and used to compare two different approaches to graph-based genotyping, PanGenie and vg. Preprocessing steps for the genotypers are not included.

2.5 PRDM9 zinc finger graph

After testing the pipeline on simulated data, PGGB and the vg pipeline was rerun with real sequencing data on the PRDM9 zinc finger (Figure 2.4). A genome graph was constructed with the whole ONT long-read based reference genome and the phased PRDM9 zinc finger sequences from Arnold, Klopp, Maxine and Simon, as well as the individuals that were not successfully phased (Table 2.1). By mapping short reads from the same individuals from which the graph was built, we can evaluate how well the graph-based genotype pipeline works in a complex region like the PRDM9 znf-array.

This input to PGGB required parameter tuning because some of the sequences are very short. The zinc finger sequences are between 671 and 923 bp long. We extended each sequence with 10 000 bp to improve alignment, but we still have to use a smaller segment length (-s) parameter in order to prevent filtering out all approximate mappings in the alignment step of the pipeline. lllumina short-reads were mapped to the graph and genotyping was performed with the vg pipeline. The graph was visualized using odgi (Guarracino et al., 2022), and the genotypes were manually compared to expected callings.

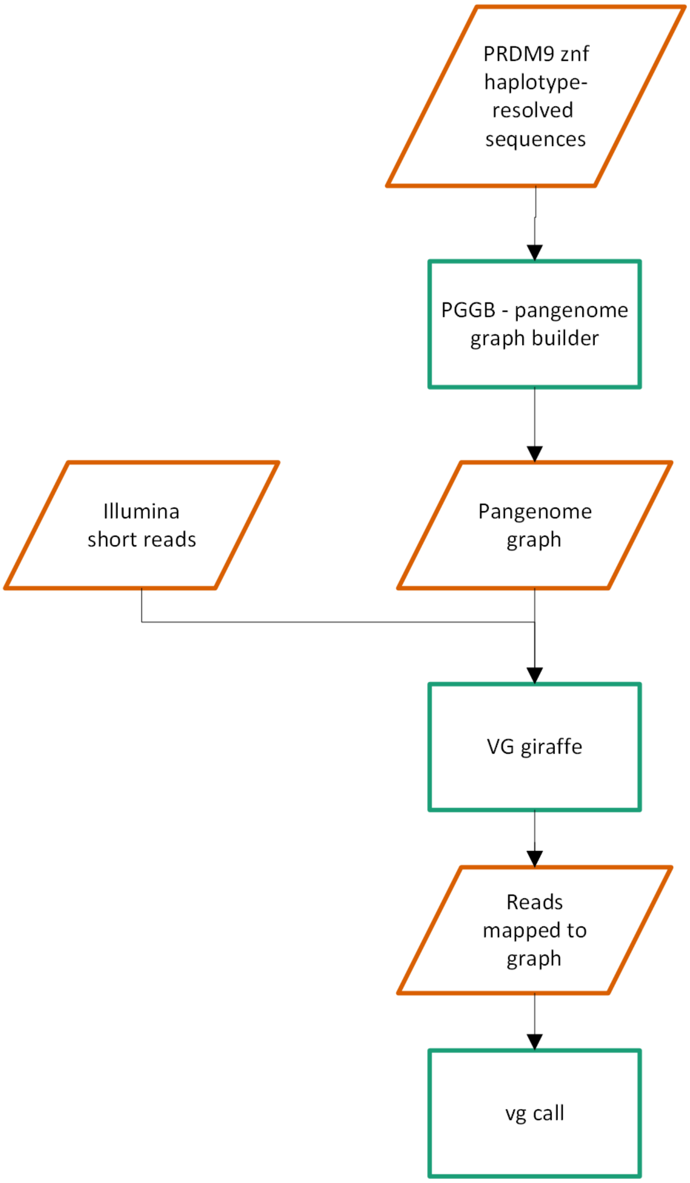


Figure 2.4: Overview of pipeline for graph construction and graph-based genotyping PRDM9 zinc finger array on Atlantic salmon chr. 5 The green boxes are processes/tools, while the orange boxes are data. A graph was constructed with haplotype-resolved sequences (Figure2.1) and the full reference genome (GCA_905237065.2). Giraffe was used to map Illumina short-reads to the graph, before inferring genotypes with vg call. Short-read based genotyping was carried out for the four individuals where we have acquired two haplotype sequences (Arnold, Maxine, Klopp, Simon, Figure 2.1)

3. Results and discussion

3.1 Simulation on Atlantic salmon chromosome 22

A sequence was simulated based on Atlantic salmon reference chromosome 22. The counts of SVs in the simulated dataset are shown in Table 3.1. Deletions are the most frequent SV-type, and we can observe an increase of SVs in the high repeat region as reported in earlier work, making our SVs representative of their respective regions. The two regions were selected based on repeat content visualised in Figure 3.1 B. We restricted SVs to two regions in order to simplify the test runs, which kept the count of SVs down and allowed us to focus on two regions with different repeat content, and identifying key parameters and repeat impact, before expanding to a more complex input.

Table 3.1: Number of SVs inserted in the two regions of the simulated sequence The regions of interest are selected based on the tandem repeat distribution shown together with the start position of all SVs in the simulated sequence as shown in Figure 3.1.

SV type	Low repeat region	High repeat region	Total
Insertion	73	140	213
Deletion	135	259	394
Total	208	399	604

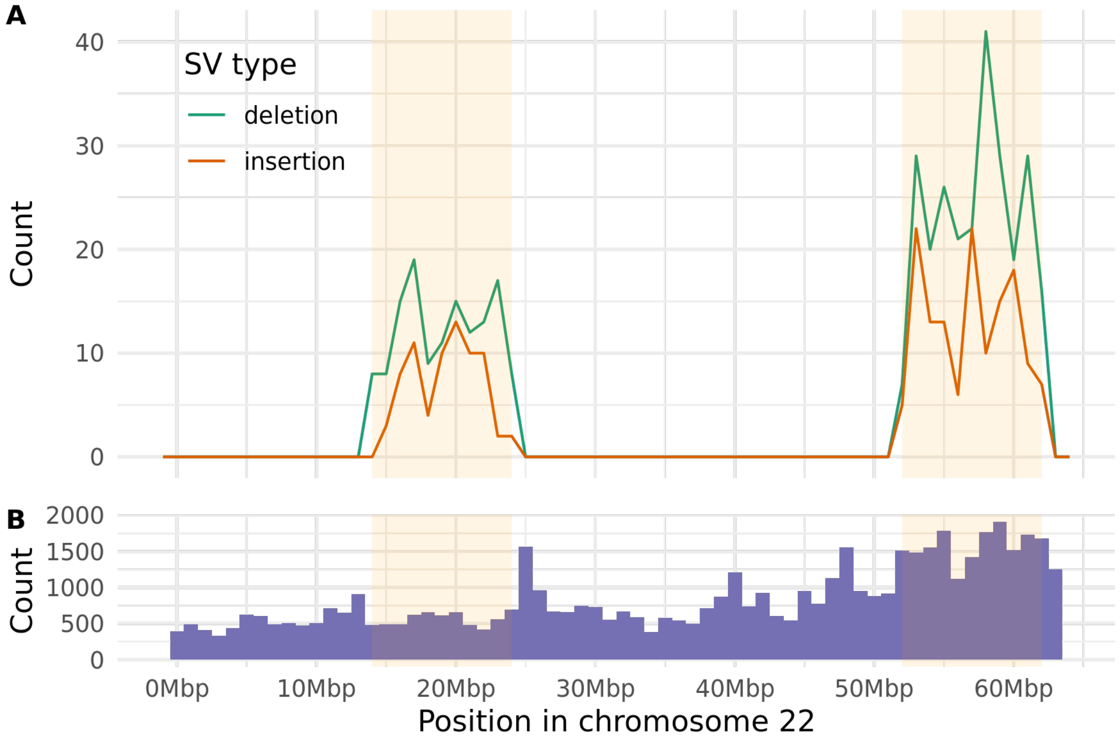


Figure 3.1: SV distribution in simulated sequence and TR density on chromosome 22 The yellow areas show the regions selected for SV simulation. **A.** The figure presents the distribution of SVs across chromosomes 22 in the simulated sequence. **B.** The figure presents the TR count in 1 Mbp bins. The data is from Monsen et al., 2022 and made with the tool *Tandem repeats finder* (Benson, 1999). SV count peaks in the same positions as TR count, which is in accordance with reports of repeats being a source of SVs.

SV-length distribution of the selected SVs are different between the two regions (Figure 3.2). We can observe an increase of shorter SVs in the high repeat region, which is likely due to repeat number differences. We can also observe a peak of approximately 1500 bp, which possibly represents the recently active transposable element reported by Bertolotti et al., 2020.

In order to simulate a realistic sequence, the simulation tool must take positioning into consideration. As previously mentioned, SVs are not evenly dispersed in a chromosome, but are expected to be enriched in repeat sequences. Visor was selected for simulation as it will insert SVs into positions defined by the user. This allowed us to base the simulated data on real SV sequence and positioning.

Even with a representative sequence and a realistic sequence length distribution within the selected regions of interest, there are important simplifications implicated by this dataset. There are no overlapping variants and no SNPs in the simulated sequence. A real sequence will include more complex structures of variation. In addition, there are only variants in two regions and we only simulated one sequence.

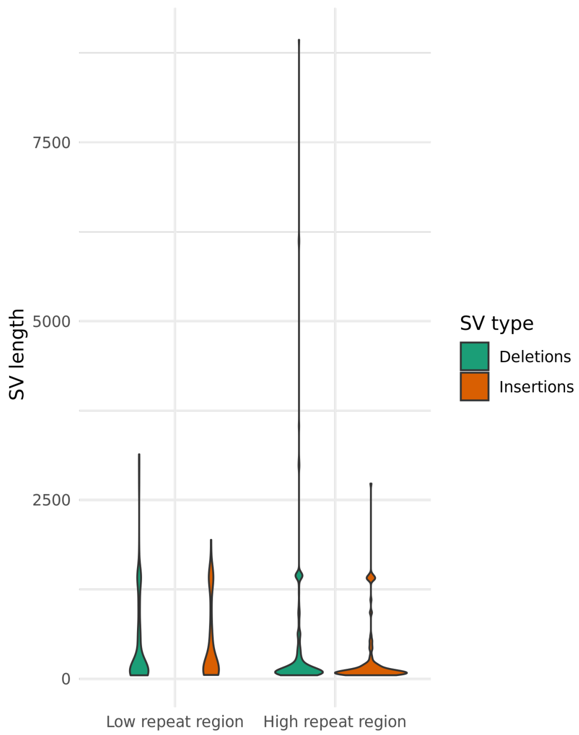


Figure 3.2: SV-length distribution in the simulated dataset There are differences in distribution between the two selected regions. In the high repeat region, the SVs were shorter in length. In addition, the longest SVs were much longer than in the low repeat region. We can observe a peak at 1500 bp which is assumed to be a recently active TE.

3.2 Optimization of parameters in PGGB

Understanding the key parameters, as well as being able to optimize them according to your data and use, is essential to make a genome graph suited to your application with PGGB. Although only three parameter inputs are required, the total number of parameters is much larger. The number of parameters makes it possible to build genome graphs from data with different levels of identity and for different applications. This comes at the cost of having to spend time on parameter tuning, and until recently, documentation on PGGB has been sparse. This has been improved, making the tool more user friendly, with recommended settings for multiple organisms and identification of the key parameters (Garrison, 2022b).

Two parameters in PGGB, -s and -G were selected for parameter optimisation as they were identified as having the largest effect on runtime and graph complexity. The minimum segment size (-s) will impact the initial step and alignment of the sequences, while the second parameter optimized (-G), will influence the "smoothening" of the graph, ensuring local linearity which is important for the graph to be applicable for down stream analysis (Garrison, 2022b). Figure 3.3 presents high F1 scores for all parameters values. Lowest F1 score is found for -s set to 5000, but the score is higher

(>0.98) for all other values of -s. The results suggests that the two parameters have little impact on PGGBs ability to detect SVs with the simulated data as input.

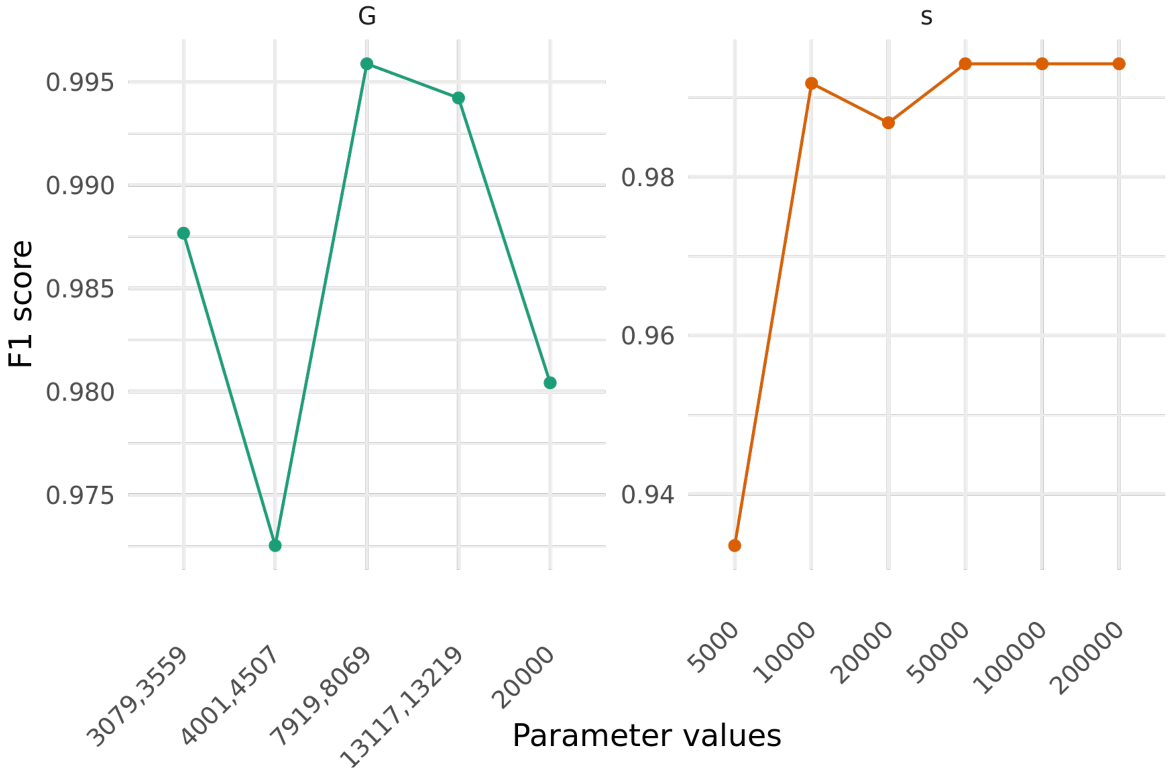


Figure 3.3: Parameter tuning PGGB Green points show F1 score for different values of parameter -G when -s is set to 100 000. Orange point show F1 score for different values of parameter -s when -G is set to 7919,8069. F1 is based on number of SVs found in the resulting graph as explained in methods section 2.3.2.

Time and memory consumption, like the F1 score, were not impacted by different values of -s (Figure 3.4). For -G on the other hand, there is a small increase in time usage with higher block size in two passes to -G, and a large shift in memory and time usage when passing only one value of 20 000. The single pass results in using less time, but required a lot more memory without improving F1 score. While there is no evidence of increased SV detection performance when increasing -G, it seems to be advantageous to include two rounds of refinement with lower block sizes compared to one round with a large block size, as memory consumption is much lower with two passes to -G (see Figure 3.4).

The small differences in SV detection performance as well as time and memory usage is likely due to our very simple input of two sequences. Using a larger number of input sequences will increase the number of pairwise mappings in wfmash, and -s will have a greater impact on run time and memory consumption. There will also be more complex multiallelic regions with an increase of sequences, which we expect will be more challenging for smoothxg, making -G have a greater impact on SV detection as well as time and memory consumption. Another artefact of the input data are the large stretches of identical sequence in the two input sequences (Figure 3.1). If variation was distributed across the chromosome, the first step of alignment would become a bigger computational challenge, thus changing -s would have more impact. This test gives us a starting point and a framework for exploring the effect of PGGB parameters on larger and more complex input data.

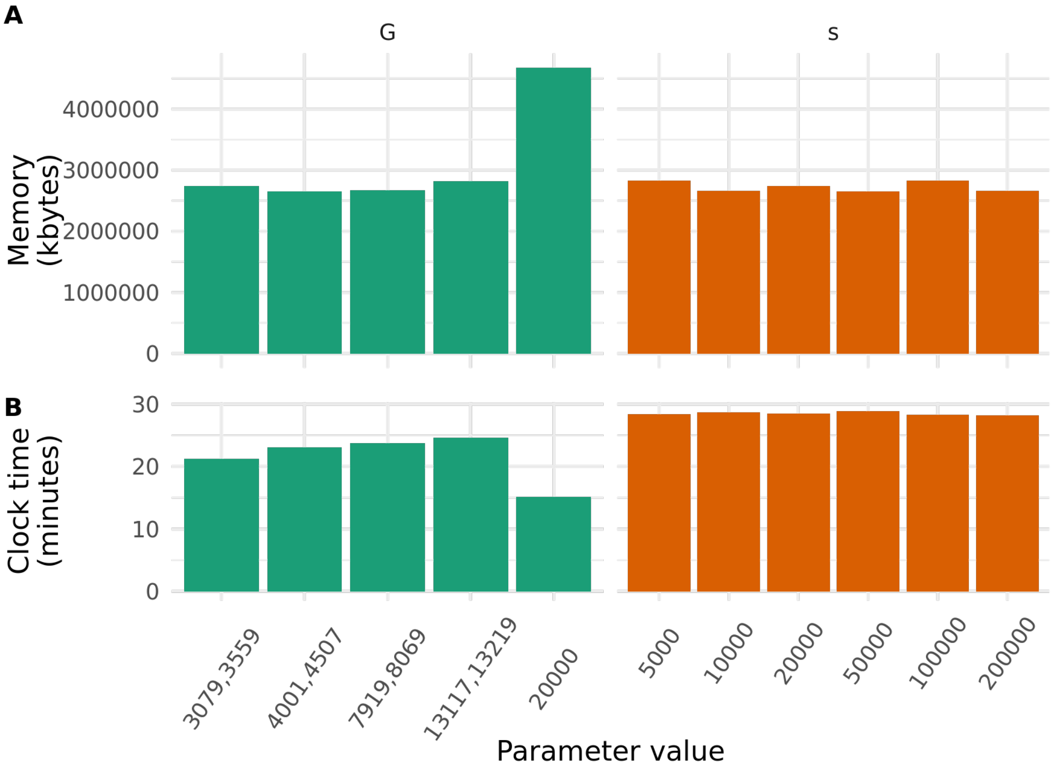


Figure 3.4: Runtime and memory usage for PGGB Running PGGB with the chr. 22 sequences (Figure 2.2) with different values of -G and -s to optimize parameters. **A.** Memory is stable except for one run where -G is set to 2000. **B.** Clock time increases with larger values for -G, and decreases when run with only one round of smoothening (one imput value).

3.3 PGGB graphs for SV detection

A chromosome 22 graph was made for comparison of the two graph-based genotyping pipelines. Table 3.2 shows high recall and precision for all regions and types of SVs. PGGB is able to detect almost every single SVs and calls very few false positive variants. All missed and falsely called variants are found in the high repeat region. A closer look reveals only one FP and one FN SV, both of approximately the same length of 100 bp, and with start positions 100 bp apart. It is likely the same variant, but repeated sequence makes identification of the start position ambiguous.

Table 3.2: SVs detected with PGGB Precision, recall and F1 score by region and SV type. Metrics are based on number of SVs found in the resulting graph as explained in methods section 2.3.2

Region	SV type	Precision	Recall	F1
Low repeat	Insertion	1.000	1.000	1.000
Low repeat	Deletion	1.000	1.000	1.000
High repeat	Insertion	1.000	1.000	1.000
High repeat	Deletion	0.996	0.996	0.996

The simulated input data may present artificially good results because of the simple nature of input. Non-simulated sequences contain complex structures and nested variants, which will be harder to detect. This is a two-sequence genome graph. When increasing the number of input sequences, the graph is going to be more complex.

The metric used to evaluate PGGB SV-detection was to compare positions between variants found in the graph and the SVs in the SV-catalog. A variant was classified as true or false based on the start and end positions, allowing a slack of 60 bp. This was an arbitrary threshold with unclear importance. Positions were compared with functions written in R, but there are exiting tools for comparing positions in VCF files like *bedtools intersect* (Quinlan and Hall, 2010). This tool may be preferable to use in order to save time, as it is well tested and already exist.

3.4 PanGenie and vg comparison on chromosome 22

The two graph based genotype pipelines, PanGenie and vg, were run with the same short-read and graph inputs in order to compare their results (Figure 3.5). The vg pipeline shows the highest precision for all depths of reads and types of SVs (Figure 3.5). PanGenie performs well with higher depth set of reads, but reveals a lower precision for calling insertions.

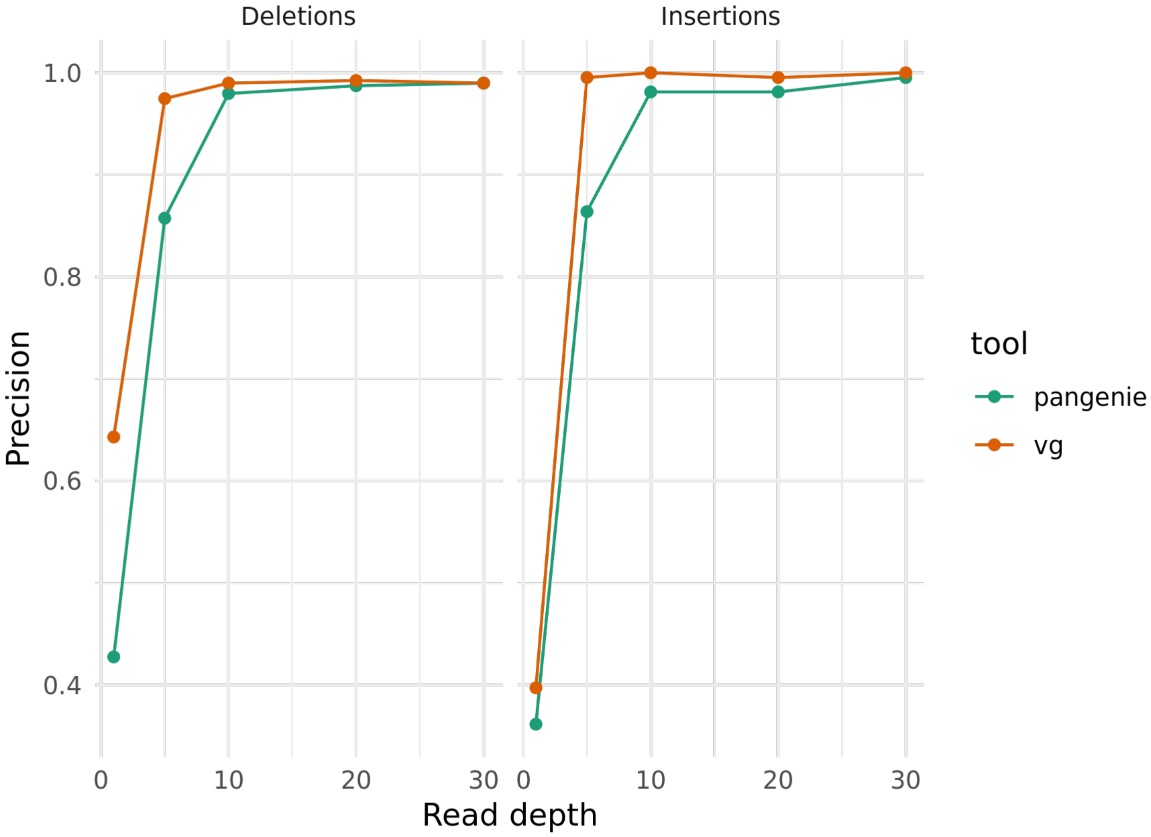


Figure 3.5: Precision comparison of two graph-based genotyping pipelines on simulated data The vg pipeline shows the highest precision for all depths of reads and types of SVs.

While PanGenie requires more memory, it is much quicker than the vg pipeline (Figure 3.6). In addition, the memory requirements are stable, not showing an increase with read depth. PanGenie finds all possible k-mers in the graph for each run, which is probably the cause of the high, but stable memory consumption. These time and memory measures must be considered as estimates, as it not consider which node of the computer cluster the command is run at as the different nodes have different CPUs. This is likely why we see larger time usage for read depth of 10 compared to depth of 20. Traffic on the cluster may also impact the runtime.

Where the vg-pipeline have multiple steps for prepossessing input files for vg giraffe and vg call, PanGenie is run by a single command. The requirements for the input VCF are strict, but other than that, PanGenie is easy to run. PanGenie does require haplotype resolved assemblies, which is unavailable for Atlantic salmon at this point.

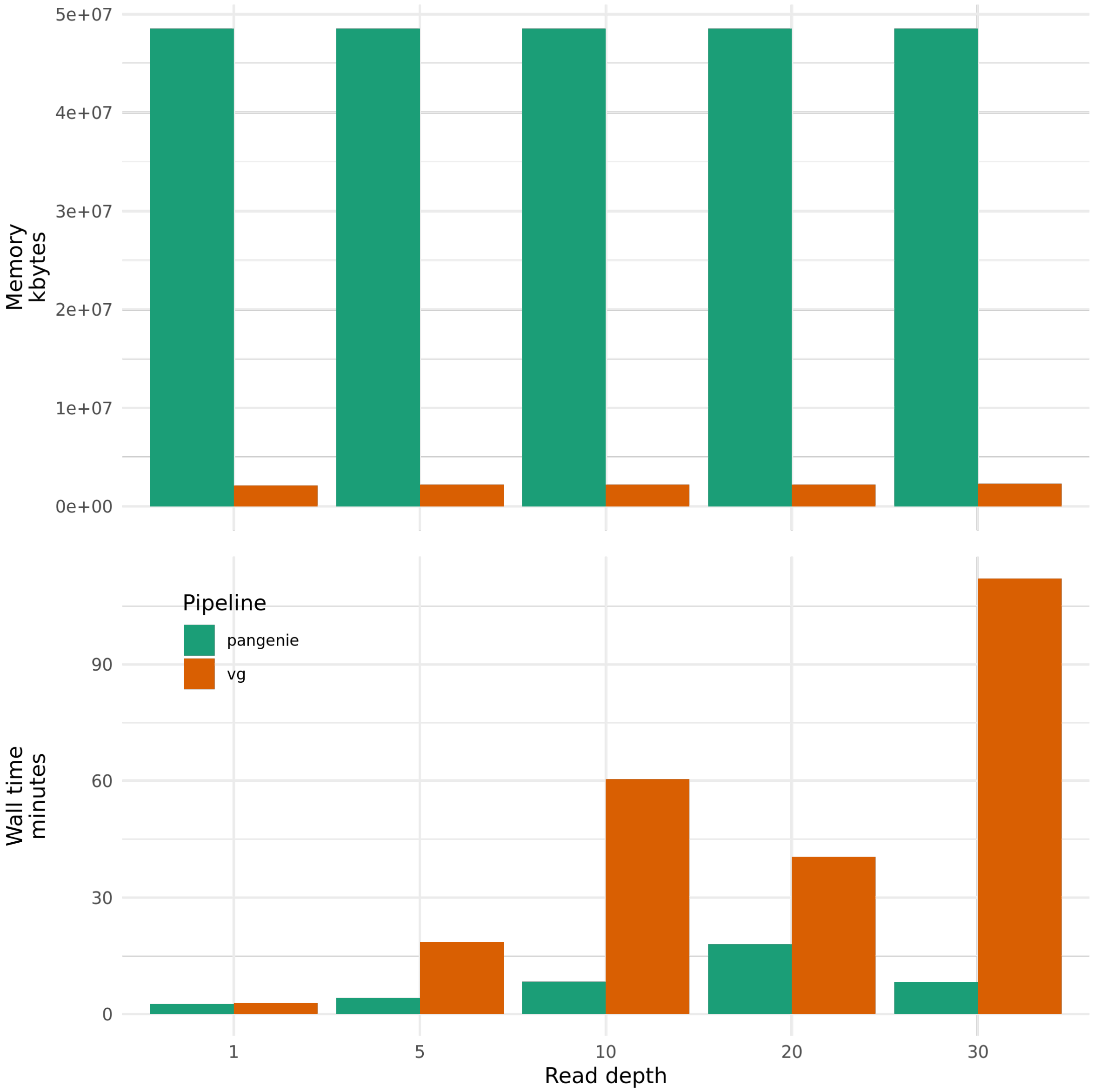


Figure 3.6: Runtime and memory usage for graph-based genotyping pipelines PanGenie shows higher memory usage, but is much faster than the vg pipeline. In general, runtime increases with depth for both tools.

3.5 PRDM9 zinc finger graph

The zinc finger graph shows the length differences of the sequences in one position (Figure 3.7). Alternative alleles corresponds to the length and sequence of one, two or three zinc finger repeats when calling genotypes in the graphs using the shortest sequence, Simon haplotype 2 (Simon#2) as a reference to call variation. One exception is the reference assembly Ssal_v3.1. which is not haplotype-resolved, meaning that the sequence is a mix of the two haplotypes of one individual. This has resulted in a collapsed zinc finger array in the reference sequence based on Simon, which is also included into the graph. A way to avoid including the collapsed variant, could be to make a smaller graph with only the haplotype-resolved sequences. However, it would not be possible to map real reads to this graph, as reads span the full genome and thus needs a full genome to be mapped to.

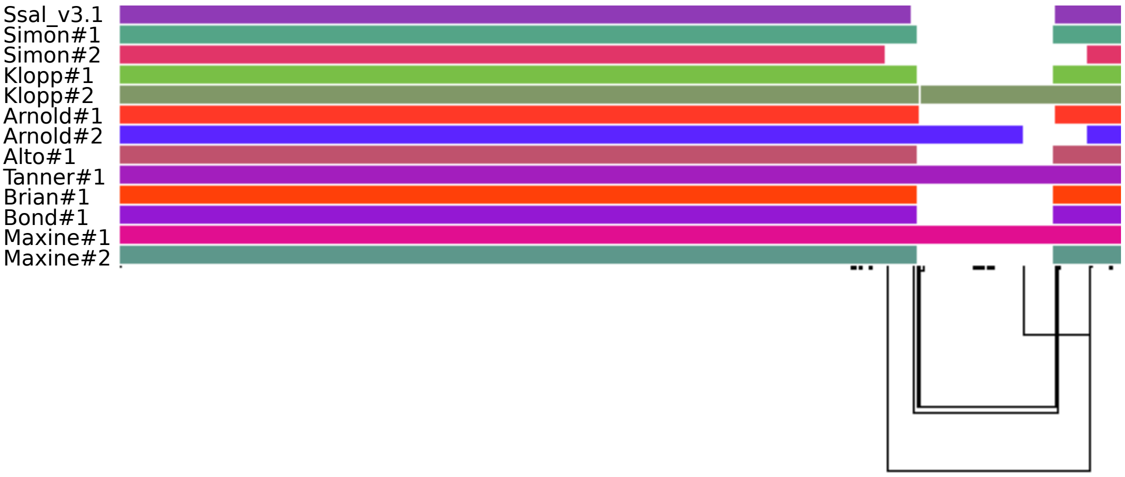


Figure 3.7: Visualisation of the variable zinc finger array with ODGI Colored lines represent genomic sequence, while the black lines show graph paths. The first sequence is the reference genome Ssal_v3.1 (GCA_905237065.2) which includes a collapsed zinc finger array. The numbers of repeats is between five and eight.

SV representation with PGGB and graph-based genotyping showed promising results on simulated data, but it remains to be seen how the complexity of real data and a larger number of input sequences will affect the pipeline. The PRDM9 zinc finger array is a very challenging region to map reads to, as it consists of a variable number of almost identical repeats. In addition, we built the graph from 13 sequences which created bubbles with up to 12 alternative alleles, making this a complex graph which could make down-stream analysis challenging.

Ideally, the graph should be constructed on chromosome scale input sequences. This was not possible since no phased Atlantic salmon assemblies of this size currently exist. Phased sequences are needed to have an accurate presentation of the PRDM9 region. Aligning shorter sequences with flanks of reference sequence was the approach we utilized to get the different length sequences to align.

3.6 Graph-based genotyping of zinc finger repeat number

We chose the vg pipeline for graph-based genotyping of the PRDM9 zinc finger region because it performed better than PanGenie on the simulated dataset. In addition, PanGenie requires a haplotype resolved input sequences, which we do not have for the full genome. Even if we were able to run PanGenie, the tool would not be able to rely on unique k-mers in this repeat region, and the short haplotype resolved sequences would not include any nearby variants to infer genotypes from.

Calling repeat numbers in the zinc finger array of PRDM9 with short-reads turned out to be challenging. For Klopp and Maxine, one repeat less then expected was called, while for Simon, one repeat more than expected was called (Figure 3.8). The results indicate a bias towards calling 7 repeats, and suggests that we are unable to reliably genotype the zinc finger repeats with this graph.

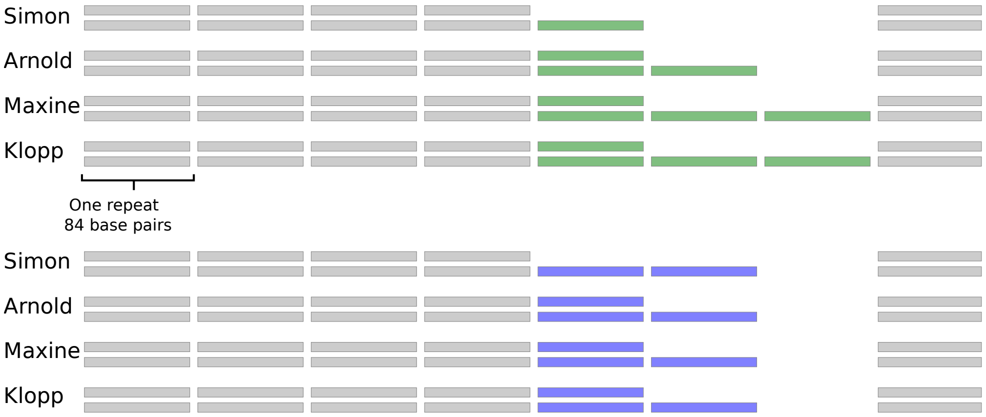


Figure 3.8: Repeat numbers called in zinc finger array with graph based genotyping Number of repeats in the PRDM9 zinc finger graph for a given individual called from the assemblies (top) and with short reads (bottom). The colored bars indicates a bubble in the graph where variation is called. Green bars show repeats represented in the graph, which is in accordance to previously reported repeat numbers 2.1. The blue bars represent the number of repeats called with short reads and the vg pipeline. The correct number of repeats was called for Arnold, but not for any of the other individuals. SNP level differences between the repeats are not taken into consideration.

Although this specific graph represent the repeat differences well, there could be qualities of the graph which is negating accurate graph-based genotyping. Genome graph quality will be limited by the quality of the assemblies, in this sense, errors from sequencing or assembly will impact graph-based genotyping. In addition, the quality of the PRDM9-sequences is dependent on phasing. The nature of the PRDM9 zinc finger array has made phasing challenging. This has resulted in missing haplotypes for several of the long-read sequenced individuals. Errors in the graph can potentially cause difficulties in mapping reads to the graph for an already challenging region. The adjacent, near identical zinc finger motifs, makes it hard to tell reads from the different repeats apart. Improving quality of the sequences used in construction of the graph can possibly make read mapping to the graph better and genotyping more successful.

Assembly-based graphs can include very complex structures. Complex structures often origin from errors in assemblies or regions that are hard to align (Guarracino et al., 2022) like we see in the PRDM9 region. It is possible to remove complex structures, which is expected to decrease the computational burden. Simplifying the graph may also make read mapping less vulnerable to artifacts introduced by complex regions. There will be fewer possible positions to map to, leading to less ambiguous mappings. The odgi toolkit has suggested pipelines to identify and simplify these regions.

Read mapping can be used as a metric of graph quality in order to optimize the graph for genotyping. This metric can be a measure of quality independent on the data set, and will allow for evaluation of graphs based on sequences without known SVs. It is possible to extract mapping statistics such as counts of un-mapped reads and perfectly mapped reads from giraffe.

3.7 Conclusion and further work

This study presents a viable method for representing the Atlantic salmon pangenome through a PGGB graph. With simulated data and the highly variable PRDM9 zinc finger repeat array, PGGB proved to successfully detect SV variation in genome assemblies by graph construction.

Two different approaches for graph-based genotyping with short-reads showed promising results on simulated data. However, PGGB allows for complex datastructures which are vulnerable to errors from sequencing, assembly and phasing. This became particularly apparent for the PRDM9 zinc finger region, where we were unable to genotype the correct number of repeats through short-reads with the vg pipeline. The contrasting results to the initial run with simulated data demonstrates the importance of solid test data in order to produce reliable results.

Ultimately, we would like to create a Atlantic salmon pangenome graph which will allow for population scale SV-studies at low cost. A full genome graph would be constructed one chromosome at a time, merging the graphs together before downstream analysis. This will be a be a large and complex data structure, which will likely require simplifications to be useful for down stream analysis. A combination of metrics like SV-count, read-mapping statistics or graph complexity will make sure to construct a graph which represent variation, but also allows for analysis such as genotyping.

References

Alkan, C., Coe, B. P., and Eichler, E. E. (May 2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics* 12 (5): 363–376. DOI: 10.1038/nrg2958.

Benson, G. (Jan. 1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27 (2): 573–580. DOI: 10.1093/nar/27.2.573.

Bertolotti, A. C., Layer, R. M., Gundappa, M. K., Gallagher, M. D., Pehlivanoglu, E., Nome, T., Robledo, D., Kent, M. P., Røsæg, L. L., Holen, M. M., Mulugeta, T. D., Ashton, T. J., Hindar, K., Sægrov, H., Florø-Larsen, B., Erkinaro, J., Primmer, C. R., Bernatchez, L., Martin, S. A. M., Johnston, I. A., Sandve, S. R., Lien, S., and Macqueen, D. J. (Oct. 2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications* 11 (1): 5176. DOI: 10.1038/s41467-020-18972-x.

Bolognini, D., Sanders, A., Korbel, J. O., Magi, A., Benes, V., and Rausch, T. (Feb. 2020). VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics* 36 (4): 1267–1269. DOI: 10.1093/bioinformatics/btz719.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., and Feschotte, C. (Nov. 2018). Ten things you should know about transposable elements. *Genome Biology* 19 (1): 199. DOI: 10.1186/s13059-018-1577-z.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (May 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. DOI: 10.1534/g3.114.015784.

Brenna-Hansen, S., Li, J., Kent, M. P., Boulding, E. G., Dominik, S., Davidson, W. S., and Lien, S. (Aug. 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. DOI: 10.1186/1471-2164-13-432.

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., and Eberle, M. A. (Dec. 2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology* 20 (1): 291. DOI: 10.1186/s13059-019-1909-7.

Conrad, D. F. and Hurles, M. E. (July 2007). The population genetics of structural variation. *Nature Genetics* 39 (7): S30–S36. DOI: 10.1038/ng2042.

Course, M. M., Gudsnuk, K., Smukowski, S. N., Winston, K., Desai, N., Ross, J. P., Sulovari, A., Bourassa, C. V., Spiegelman, D., Couthouis, J., Yu, C.-E., Tsuang, D. W., Jayadev, S., Kay, M. A., Gitler, A. D., Dupre, N., Eichler, E. E., Dion, P. A., Rouleau, G. A., and Valdmanis, P. N. (Sept. 2020). Evolution of a Human-Specific Tandem Repeat Associated with ALS. eng. *American journal of human genetics* 107 (3). Edition: 2020/08/03 Publisher: Elsevier: 445–460. DOI: 10.1016/j.ajhg.2020.07.004.

Crysnanto Danang, Leonard Alexander S., Fang Zih-Hua, and Pausch Hubert (May 2021). Novel functional sequences uncovered through a bovine multiassembly graph. *Proceedings of the National Academy of Sciences* 118 (20). Publisher: Proceedings of the National Academy of Sciences: e2101056118. DOI: 10.1073/pnas.2101056118.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (Dec. 2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics* 7 (12). Publisher: Public Library of Science: e1002384. DOI: 10.1371/journal.pgen.1002384.

Ebler, J., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Korbel, J., Eichler, E. E., Zody, M. C., Dilthey, A. T., and Marschall, T. (Jan. 2020). Pangenome-based genome inference. *bioRxiv*: 2020.11.11.378133. DOI: 10.1101/2020.11.11.378133.

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., and Marschall, T. (Apr. 2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics* 54 (4): 518–525. DOI: 10.1038/s41588-022-01043-w.

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., Gudbjartsson, D. F., Stefansson, K., Halldorsson, B. V., and Melsted, P. (Nov. 2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications* 10 (1): 5402. DOI: 10.1038/s41467-019-13341-9.

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J. D., Rounthwaite, R., Ebler, J., Rautiainen, M., Garg, S., Paten, B., Marschall, T., Sirén, J., and Garrison, E. (Aug. 2020). Pangenome Graphs. *Annual Review of Genomics and Human Genetics* 21 (1). Publisher: Annual Reviews: 139–162. DOI: 10.1146/annurev-genom-120219-080406.

Farnoud, F., Schwartz, M., and Bruck, J. (Feb. 2019). Estimation of duplication history under a stochastic model for tandem repeats. *BMC Bioinformatics* 20 (1): 64. DOI: 10.1186/s12859-019-2603-1.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., and Durbin, R. (Oct. 2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* 36 (9): 875–879. DOI: 10.1038/nbt.4227.

Garrison, E., Guarracino, A., and Heumos, S. (2021). *The pangenome graph builder*. URL: https://github.com/pangenome/pggb/tree/v0.2.0.

Garrison, E., Heumos, S., Guarracino, A., and Gao, Y. (May 2022). *Homogenizing and ordering the graph with smoothxg*. URL: https://github.com/pangenome/pggb/tree/v0.2.0#homogenizing-and-ordering-the-graph-with-smoothxg.

Garrison, E. (May 2022a). *PanSN-spec: Pangenome Sequence Naming*. URL: https://github.com/pangenome/PanSN-spec.

Garrison, E. (May 2022b). *PGGB documentation*. URL: https://pggb.readthedocs.io/en/latest/.

Garrison, E. and Guarracino, A. (Jan. 2022). Unbiased pangenome graphs. *bioRxiv*: 2022.02.14.480413. DOI: 10.1101/2022.02.14.480413.

GFA group (May 2022). *GFA: Graphical Fragment Assembly (GFA) Format Specification*. URL: https://github.com/GFA-spec/GFA-spec.

Grey, C., Baudat, F., and de Massy, B. (Aug. 2018). PRDM9, a driver of the genetic map. *PLOS Genetics* 14 (8). Publisher: Public Library of Science: e1007479. DOI: 10.1371/journal.pgen.1007479.

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., and Garrison, E. (Jan. 2022). ODGI: understanding pangenome graphs. *bioRxiv*: 2021.11.10.467921. DOI: 10.1101/2021.11.10.467921.

Gundappa, M. K., To, T.-H., Grønvold, L., Martin, S. A. M., Lien, S., Geist, J., Hazlerigg, D., Sandve, S. R., and Macqueen, D. J. (Jan. 2022). Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Molecular Biology and Evolution* 39 (1): msab310. DOI: 10.1093/molbev/msab310.

Hartley, G. and O'Neill, R. J. (2019). Centromere Repeats: Hidden Gems of the Genome. *Genes* 10 (3). DOI: 10.3390/genes10030223.

Heller, D. and Vingron, M. (Sept. 2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35 (17): 2907–2915. DOI: 10.1093/bioinformatics/btz041.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., and Paten, B. (Feb. 2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology* 21 (1): 35. DOI: 10.1186/s13059-020-1941-7.

Ho, S. S., Urban, A. E., and Mills, R. E. (Mar. 2020). Structural variation in the sequencing era. *Nature Reviews Genetics* 21 (3): 171–189. DOI: 10.1038/s41576-019-0180-9.

Houston, R. D., Bean, T. P., Macqueen, D. J., Gundappa, M. K., Jin, Y. H., Jenkins, T. L., Selly, S. L. C., Martin, S. A. M., Stevens, J. R., Santos, E. M., Davie, A., and Robledo, D. (July 2020). Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics* 21 (7): 389–409. DOI: 10.1038/s41576-020-0227-y.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (Feb. 2012). ART: a next-generation sequencing read simulator. eng. *Bioinformatics (Oxford, England)* 28 (4). Edition: 2011/12/23 Publisher: Oxford University Press: 593–594. DOI: 10.1093/bioinformatics/btr708.

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

Li, H., Feng, X., and Chu, C. (Oct. 2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology* 21 (1): 265. DOI: 10.1186/s13059-020-02168-z.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Inge Våge, D., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J. M., Jonassen, I., Maass, A., Omholt, S. W., and Davidson, W. S. (May 2016). The Atlantic salmon genome provides insights into rediploidization. *Nature* 533 (7602): 200–205. DOI: 10.1038/nature17164.

Liu, S., Gao, G., Layer, R. M., Thorgaard, G. H., Wiens, G. D., Leeds, T. D., Martin, K. E., and Palti, Y. (2021). Identification of High-Confidence Structural Variants in Domesticated Rainbow Trout Using Whole-Genome Sequencing. *Frontiers in Genetics* 12. URL: https://www.frontiersin.org/article/10.3389/fgene.2021.639355.

Lu, T.-Y., Munson, K. M., Lewis, A. P., Zhu, Q., Tallon, L. J., Devine, S. E., Lee, C., Eichler, E. E., Chaisson, M. J. P., and The Human Genome Structural Variation Consortium (July 2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nature Communications* 12 (1): 4250. DOI: 10.1038/s41467-021-24378-0.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biol* 20 (1). Type: Journal Article: 246–246. DOI: 10.1186/s13059-019-1828-7.

Martiniano, R., Garrison, E., Jones, E. R., Manica, A., and Durbin, R. (Sept. 2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology* 21 (1): 250. DOI: 10.1186/s13059-020-02160-7.

Mérot, C., Stenløkk, K. S. R., Venney, C., Laporte, M., Moser, M., Normandeau, E., Árnyasi, M., Kent, M., Rougeux, C., Flynn, J. M., Lien, S., and Bernatchez, L. (Apr. 2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (Coregonus sp.) with long and short reads. *Molecular Ecology* n/a (n/a). Publisher: John Wiley & Sons, Ltd. DOI: 10.1111/mec.16468.

Monsen, Ø., Stenløkk, K. S., Sandve, S. R., and Sigbjørn, L. (2022). Structural Variation in Atlantic salmon Strongly Correlated with Telomere Accociated Tandem Repeats. Manuscript in prep.

Novak, A. M. (Apr. 2022). URL: https://github.com/vgteam/vg/issues/3614#issuecomment-1086329910.

Quinlan, A. R. and Hall, I. M. (Mar. 2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6): 841–842. DOI: 10.1093/bioinformatics/btq033.

Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., Onn, A., Chapal-Ilani, N., Tao, L., Marx, T., Feige, U., and Shapiro, E. (Mar. 2019). Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Research* 47 (5): 2436–2445. DOI: 10.1093/nar/gky1318.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M. C. (June 2018). Accurate detection of complex structural variations using single-molecule sequencing. eng. *Nature methods* 15 (6). Edition: 2018/04/30: 461–468. DOI: 10.1038/s41592-018-0001-7.

Sinclair-Waters, M., Ødegård, J., Korsvoll, S. A., Moen, T., Lien, S., Primmer, C. R., and Barson, N. J. (Feb. 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. DOI: 10.1186/s12711-020-0529-8.

Sirén Jouni, Monlong Jean, Chang Xian, Novak Adam M., Eizenga Jordan M., Markello Charles, Sibbesen Jonas A., Hickey Glenn, Chang Pi-Chuan, Carroll Andrew, Gupta Namrata, Gabriel Stacey, Blackwell Thomas W., Ratan Aakrosh, Taylor Kent D., Rich Stephen S., Rotter Jerome I., Haussler David, Garrison Erik, and Paten Benedict (2022). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374 (6574) (). Publisher: American Association for the Advancement of Science: abg8871. DOI: 10.1126/science.abg8871.

Stenløkk, K., Moser, M., Monsen, Ø., Manousi, D., Nome, T., Árnyasi, M., Kent, M., Sandve, S., and Lien, S. (2022). The Atlantic salmon pan-genome provides insight into the structural variation landscape across phylogeographic groups. Manuscript in prep.

Sulovari Arvis et al. (Nov. 2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the National Academy of Sciences* 116 (46). Publisher: Proceedings of the National Academy of Sciences: 23243–23253. DOI: 10.1073/pnas.1912175116.

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., and Linke, D. (Dec. 2019). Tandem repeats lead to sequence assembly errors and impose multilevel challenges for genome and protein databases. *Nucleic Acids Research* 47 (21): 10994–11006. DOI: 10.1093/nar/gkz841.

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., Koenig, B. A., Li, D., Marschall, T., McMichael, J. F., Novak, A. M., Purushotham, D., Schneider, V. A., Schultz, B. I., Smith, M. W., Sofia, H. J., Weissman, T., Flicek, P., Li, H., Miga, K. H., Paten, B., Jarvis, E. D., Hall, I. M., Eichler, E. E., Haussler, D., and the Human Pangenome Reference Consortium (Apr. 2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604 (7906): 437–446. DOI: 10.1038/s41586-022-04601-8.

Zhang, J., Zhao, J., Xu, Y., Liang, J., Chang, P., Yan, F., Li, M., Liang, Y., and Zou, Z. (2015). Genome-Wide Association Mapping for Tomato Volatiles Positively Contributing to Tomato Flavor. *Frontiers in Plant Science* 6. URL: https://www.frontiersin.org/article/10.3389/fpls.2015.01042.

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., Fiddes, I. T., Barrio, A. M., Wala, J., Carroll, A., Ghaffari, N., Rodriguez, O. L., Bashir, A., Jackman, S., Farrell, J. J., Wenger, A. M., Alkan, C., Soylev, A., Schatz, M. C., Garg, S., Church, G., Marschall, T., Chen, K., Fan, X., English, A. C., Rosenfeld, J. A., Zhou, W., Mills, R. E., Sage, J. M., Davis, J. R., Kaiser, M. D., Oliver, J. S., Catalano, A. P., Chaisson, M. J. P., Spies, N., Sedlazeck, F. J., and Salit, M. (Nov. 2020). A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology* 38 (11): 1347–1355. DOI: 10.1038/s41587-020-0538-8.

Appendix A. Software and code

All scripts used in this thesis are available in the following github repository: https://github.com/ankjelst/SalmonGraph

The scripts were run at the Orion High Performance Computing at the University of Norwegian University of Life Sciences at arbitrary nodes with a variable number of cores defined in the slurm scripts.

Software has been run in containers with singularity version 3.8.6-1.el7, with the exception of PanGenie which was installed as described in the readme (https://github.com/eblerjana/pangenie/tree/d03d66d7da2e158a67a9b7e02c604e7fd09a8d57) as there were no existing image available.

The PGGB docker image version 0.2.0 was downloaded from github https://github.com/pangenome/pggb. Odgi was also run in this container.

The vg tools, including giraffe was run with version 1.38.0 Canossa https://github.com/vgteam/vg.

VISOR image from docker https://hub.docker.com/r/davidebolo1993/visor

ART image from galaxy *art:2016.06.05–he1d7d6f_6*

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway