Norwegian University of Life Sciences

Master's Thesis 2023 60 ECTS

Department of Animal and Aquacultural sciences (IHA)
Faculty of Biosciences (BIOVIT)

Characterization of Computational Pipelines for Structural Variant Detection Using Short-Read Sequencing Data in Arctic Charr.

Syed Muneeb Ur Rehman
M.Sc Aquaculture

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my main supervisor, Matthew Peter Kent, for his invaluable guidance, support, and encouragement throughout my master's research and writing of this thesis. This work would not have been possible without his vast knowledge and insightful feedback. I would also like to extend my sincere thanks to my co-supervisors, Kristina Stenløkk and Célian Diblasi, for taking the time to review my work and provide thoughtful suggestions and questions that strengthened my research and writing.

Furthermore, I am grateful to NMBU and the Orion cluster computing at the Centre for Integrative Genetics (CIGENE) for providing the facilities, resources, and assistance that enabled me to pursue my computational research effectively.

Finally, I must thank my family and friends for their love, patience, and belief in me. I am especially appreciative of my parents, and my significant other, for supporting me every step of the way on this long but rewarding journey.

Syed Muneeb Ur Rehman

December 15, 2023

Abstract

Structural variants (SVs) are an important emerging class of genomic variation with pivotal implications for evolution, adaptation, and phenotypic diversity. As a cold-water salmonid fish displaying extensive niche variation and life history plasticity, the Arctic charr *(Salvelinus alpinus*) serves as an ideal model to elucidate the genomic underpinnings of adaptability. This study performs an integrated analysis to comprehensively characterize the SV landscape across 30 genomes of farmed Arctic charr strains. Using a multi-algorithm approach employing Delly, Manta and Smoove for variant detection, overall 47,966 high-confidence were identified, including deletions, duplications, inversions and translocations. The results show variable numbers of SV's between individuals, ranging from 71,866 to 128,116 per fish, and reveal that chromosome 36 is enriched for SVs, containing up to 23% of all structural variations. Additional analyses with sequencing coverage data further support the inferences that patterns in chromosome architecture lead to increased structural variation susceptibility. This project substantiates the ability to reliably capture SVs from short-read resequencing but also highlights limitations when using short-read data. By enumerating SVs differentiated among domesticated strains, this study potentiates future research into SVs allele distributions, segregation, and trait associations in selective breeding programs. Overall, the analytical framework and genomic resources developed considerably advance characterization of structural variation spectra in this salmonid species.

Table of Content

List of Figures

List of tables

Abbreviations

SV	Structural Variants
bp	Base pair
Kb	Kilobase
Mb	Megabyte
SNP	Single Nucleotide Polymorphism
BND	Breakends
PacBio	Pacific Biosciences
HiFi	High Fidelity
VCF	Variant Call Format
BAM	Binary Alignment Map
SAM	Sequence Alignment Map
Gb	Gigabase
VCF	Variant Calling Format

Chapter 1 Introduction

Global aquaculture production has expanded steadily over the past decades, providing a crucial source of affordable protein to a growing human population (FAO 2022). Selective breeding programs have played a key role in the continued growth and advancement of aquaculture production. Traditionally, aquaculture breeding relied on phenotypic selection and family-based models (Zhang et al., 2022)(FAO, 1995). However, the integration of genomic information and technologies is transforming breeding approaches, The study conducted by Houston et al., (2020) thoroughly investigates the progress and effectiveness of genomics in aquaculture breeding, highlighting its transformational impact. Molecular marker data enables more accurate and accelerated genetic improvement through genomic selection and marker-assisted selection (Zenger et al., 2019). While single nucleotide polymorphisms (SNPs) have been the main genetic marker previously, but recent studies increasingly highlight the importance of structural variants as sources of genetic variation that impact quantitative traits and adaptive potential. For example, a comprehensive analysis of Atlantic salmon genomes by (Bertolotti et al., 2020) identified 15,483 high-confidence SVs using whole-genome sequencing of 492 individuals. This research accurately recovered population genetic patterns and provided insights into the role of SVs in genome evolution and genetic basis of domestication-related traits. Such findings are especially relevant for aquaculture like Arctic char, as they offer new perspectives on the genetic factors influencing key trait variations and domestication processes. Similarly, in livestock the recent study by Steensma et al., (2023) reveals how SVs also act as important markers of genetic variation affecting quantitative traits and adaptive abilities. These insights could help guide breeding programs to use SVs for boosting beneficial traits in aquaculture species including Arctic char. Characterization of SVs has important applications in animal genetics and breeding, particularly for agricultural species. In cattle, widespread SVs were found to be breed-specific, enabling genomic prediction of complex traits (Koufariotis et al., 2018). In pigs, thousands of SVs correlate with economically-relevant phenotypes such as meat quality, reproduction, and growth (Zong et al., 2023). In chickens, over 49,000 SVs among diverse breeds impact genes controlling metabolic and immune traits (Zhang et al., 2022).

Based on this understanding, genomic selection (GS) is one such approach that uses genome-wide molecular markers to predict breeding values for quantitative traits (Meuwissen et al., 2001). By genotyping SNPs and SVs effects across the genome, GS can predict genomic estimated breeding values (GEBV) early in life with increased accuracy compared to pedigree-based models. This enables higher selection intensity, reduces generation interval, and facilitates selection for traits that are difficult or expensive to measure directly such as disease resistance and product quality traits (S. Liu et al., 2015; Vallejo et al., 2017).

GS is most advanced in salmonid breeding, where it has been adopted by major breeding companies over the past decade. The implementation of GS in Atlantic salmon is associated with rapid genetic gains per generation for key traits such as harvest weight and fillet color, as well as reduced rates of early maturation (Tsai et al., 2015). Beyond salmonids, GS has shown strong promise in other aquaculture species. Recent studies have demonstrated its application in about 20 different species, significantly enhancing accuracy in breeding values, particularly for growth and disease resistance traits (Allal & Nguyen Hong Nguyen, 2022). The AquaIMPACT project further illustrates this point, revealing that accurate genomic selection can be achieved in species like rainbow trout with fewer SNP markers, thereby reducing costs and facilitating wider adoption in aquaculture breeding programs (AquaIMPACT, 2023). This adaptability of GS, especially for traits not directly measurable in

broodstock, such as product quality and disease resistance, underscores its potential across various aquaculture species (AquaIMPACT, 2023). While GS is still being optimized in many aquaculture species, collectively these studies demonstrate its advantages in accelerating genetic improvement (Houston et al., 2020). Key factors influencing GS success include marker density, population size, pedigree structure, and trait architecture. As genotyping costs continue to decrease, the use of GS is expected to expand and benefit additional aquaculture breeding programs.

1.1 Applications of Marker-Assisted Selection

In addition to genomic selection, marker-assisted selection (MAS) makes use of genotype-phenotype associations to select for specific genes or loci affecting quantitative traits. MAS has been applied in aquaculture breeding both independently and in conjunction with GS models (Abdelrahman et al., 2017; Z. J. Liu & Cordes, 2004). Target traits include growth rate, processing yield, flesh quality, appearance traits, disease resistance, temperature tolerance, and age at maturation. Major target genes that have been integrated into MAS include the growth hormone transgene in Atlantic salmon, myostatin variants for enhanced muscle mass, and the RYR3 gene affecting muscle fiber density (Abdelrahman et al., 2017). Another significant example of MAS in aquaculture is the use of a quantitative trait locus (QTL) for resistance to infectious pancreatic necrosis virus (IPNV) in Atlantic salmon. This QTL, linked to the epithelial cadherin (cdh1) gene, accounts for most of the genetic variation in resistance to the virus. Implementing this QTL in MAS has led to a dramatic reduction in the number of IPN outbreaks in salmon farms, indicating the pivotal role of MAS in enhancing disease resistance in aquaculture breeding programs. This approach has been crucial in the aquaculture industry, as IPNV is one of the most prevalent and economically damaging diseases in farmed Atlantic salmon. The discovery and application of the cdh1 gene in MAS illustrate how targeted genetic interventions can substantially improve the sustainability and productivity of aquaculture operations, making disease resistance a top priority for breeding companies By incorporating markers associated with key traits into selection decisions, MAS provides a means to directly target and optimize specific phenotypic outcomes.

The implementation of both GS and MAS is the development of abundant, validated SNP markers across genomes of farmed species. SNPs are single base pair changes representing the most abundant form of genetic variation. SNP discovery in aquaculture species has been enabled by high-throughput sequencing technologies. Medium to high-density SNP arrays have now been developed for most major aquaculture species including Atlantic salmon (Houston et al., 2014), rainbow trout (Palti et al., 2015), catfish (Z. Liu et al., 2016), Pacific white shrimp (Yu et al., 2015),) and scallops (Gutierrez et al., 2017). These SNP resources are being utilized for a multitude of applications beyond genomic selection, including assessing genetic diversity, mapping quantitative trait loci (QTLs), determining parentage, and enabling selective breeding through marker-assisted introgression from wild relatives (Abdelrahman et al., 2017). A primary use of SNP markers is to conduct genome-wide association studies (GWAS) to identify QTLs associated with complex polygenic traits (Yáñez et al., 2023). The discovery and testing of trait-associated SNPs allows for subsequent integration into GS models and MAS programs. In Atlantic salmon, GWAS using a SNP array led to the identification of markers for resistance against the Salmon Rickettsial Syndrome (SRS), one of the most costly diseases in salmon aquaculture (Correa et al., 2015). Genome-wide SNPs have also been used to dissect the genetic architecture of complex traits including growth, fillet color, and texture in rainbow trout (Gonzalez-Pena et al., 2016). The ability to rapidly generate genome-wide SNP datasets has been pivotal in enabling selective breeding efforts that align with aquaculture industry goals.

1.2 Structural Variants: An Emerging Frontier in Aquaculture Genomics

Structural variations (SVs) are defined as large genomic variations encompassing DNA segments typically ranging from 50 base pairs to megabases in size. The primary classes of SVs include deletions, insertions, duplications, inversions, and translocations (Diblasi et al., 2023; Escaramís et al., 2015; Mahmoud et al., 2019; Balachandran & Beck, 2020). Where deletions represent loss of genomic sequence, insertions constitute gain of sequence. Duplications create additional copies of pre-existing sequences. Inversions rearrange the orientation of genomic segments . Translocations occur when sections are relocated to new positions, either within or between chromosomes (Figure 1).

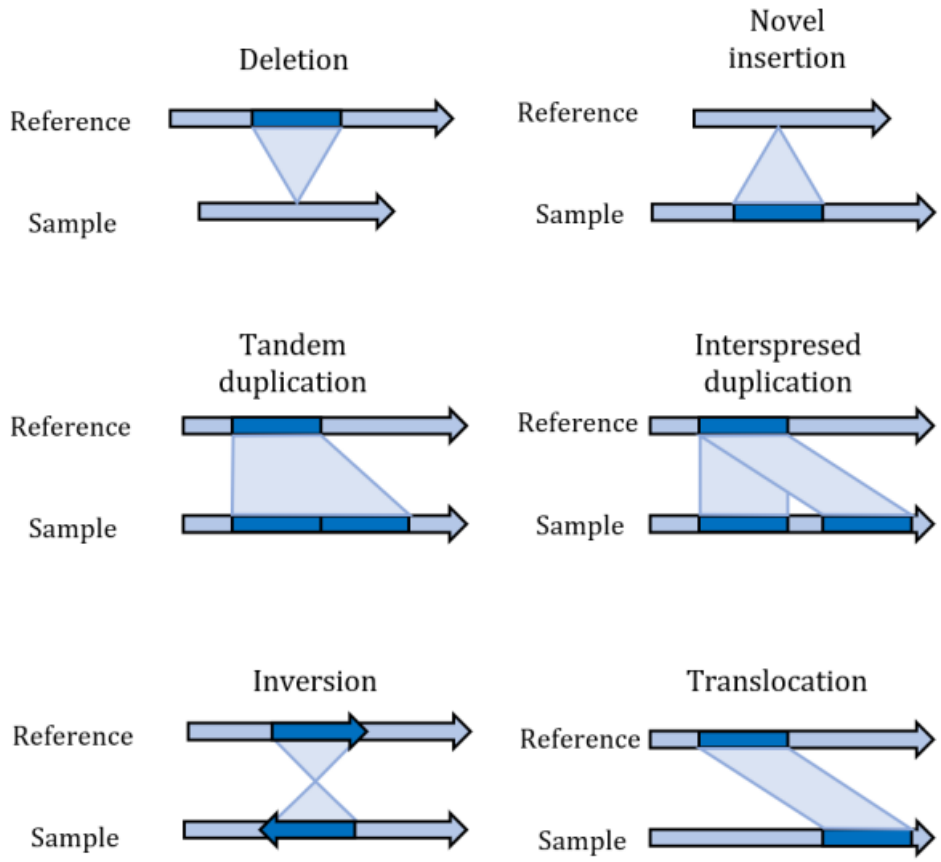


Figure 1. Schematic of common structural variation types. Illustration depicting the four predominant categories of genomic structural variations. Figure from (Stenløkk, 2023).

In addition to these cardinal types, SVs exhibit tremendous variability in size, origin, recurrence, and effects. SVs were initially discovered through karyotype analyses over a half-century ago (Yang, 2020). However, the true extent of SVs was not revealed until the advent of whole-genome array and sequencing platforms in the 2000s (Feuk et al., 2006). In addition to their pivotal roles in genomic evolution and adaptation, SVs have been associated with diverse phenotypic consequences in humans, model organisms, aquacultural and agricultural species. Comprehensive characterization of SVs is therefore crucial for understanding genetic variation and its implications for health, disease, and evolutionary fitness.

While SNPs have dominated molecular breeding applications, there is growing recognition that SVs represent a significant form of underutilized genetic variation for selective breeding programs. Studies across diverse species have shown that SVs account for more divergent genomic content between individuals than SNPs (Redon et al., 2006; Chaisson et al., 2015). This suggests that SVs likely contribute to phenotypic variation. However, the focus has remained on SNP markers as they are easier to genotype. Recent advances in long-read sequencing technologies have enabled considerable progress

in the detection and analysis of SVs across diverse genomes (Jiang et al., 2022; Sakamoto et al., 2021). Additional research is still needed to optimize computational pipelines for identifying SVs from short reads in aquaculture species. Improving methods for SV discovery could lead to better understanding of how SVs influence productive traits. Incorporating SVs into genomic selection models may strengthen predictions of breeding values and rates of genetic gain over the exclusive use of SNPs. Looking forward, a better characterization of structural variants and their associations with complex traits has the potential to further advance selective breeding. Continued research and optimization of genomics tools will ensure that aquaculture productivity can keep pace with global food demands in a responsible and sustainable manner.

1.3 Formation and Prevalence of Structural Variations

SVs arise primarily through errors in DNA replication and repair mechanisms. Non-allelic homologous recombination (NAHR) between repetitive elements or segmental duplications on different chromosomes is a major driver of recurrent SVs (Burssed et al., 2022). NAHR is stimulated by repetitive sequences, but the presence of repetitive elements explains only a fraction of NAHR-derived SVs, indicating involvement of additional genomic architectural features. Non-homologous end joining (NHEJ) also generates rearrangements, particularly non-recurrent SVs, at double-stranded breaks (Chang et al., 2017). Replication-based mechanisms such as fork stalling and template switching (FoSTeS) produce rearrangements including deletions, duplications, and complex SVs. FoSTeS occurs when the replication fork stalls and switches templates, resulting in abnormal joining and rearrangement (Mani & Chinnaiyan, 2010). Finally, retrotransposition via long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and endogenous retroviruses results in de novo insertions throughout the genome (Carvalho & Lupski, 2016).

The prevalence and spectrum of SVs has been characterized through whole-genome sequencing of diverse populations. For instance, a study by Chaisson et al. (2019) found that SVs affect 4.8-9.5% of the human genome sequence, corresponding to 84-164 Mb of cumulative SVs per diploid genome. In contrast, SNPs impact only about 0.1% of the human genome. This disparity emphasizes the substantial scale and sequence diversity shown by SVs. A similar trend is observed in non-human species, with SVs affecting about 7% of the chimpanzee genome (Sudmant et al., 2013) and 4-5% of the cattle genome, where over 5 million SVs were identified (Koufariotis et al., 2018). Notably, in the marine teleost Chrysophrys auratus, a study revealed that SVs outnumber SNPs by a threefold ratio, significantly contributing to genomic variation and potentially affecting ecological and evolutionary processes (Catanach et al., 2019). Collectively, these analyses demonstrate the pervasive distribution and variable scale of SVs across diverse eukaryotic genomes, underscoring their greater cumulative impact on genomic content and architecture compared to smaller variations like SNPs. Collectively, these analyses demonstrate the pervasive distribution and variable scale of SVs across diverse eukaryotic genomes. Compared to smaller variations like SNPs, SVs have a much greater cumulative impact on genomic content and architecture. In addition to effects on phenotypes, SVs play a pivotal role in genomic evolution through their ability to generate genetic diversity (Hollox et al., 2022). SVs are a key source of inter-individual genetic variation within populations, as well as divergence between species (Mérot et al., 2020).

Structural variants (SVs) play a crucial role in shaping genetic diversity and driving evolutionary processes in natural populations (Kirkpatrick & Barton, 2006;Ravinet et al., 2017). However, the significance of SVs in aquaculture setups differs in its implications. Within the scope of this thesis, the focus is on studying structural variants (SVs) present in aquaculture populations, but the methodologies and insights gained can be applicable and valuable to studies on SVs in wild populations The evolutionary dynamics of these populations are predominantly determined by controlled breeding

practices and artificial selection, as outlined by Gjedrem et al. (2012). This shift in focus signifies a transition from the general evolutionary significance of structural variants (SVs) to their particular effects within regulated, production-oriented settings. In contrast to wild populations, in which structural variations (SVs) play a crucial role in driving adaptation and speciation (Lamichhaney et al., 2015;Todesco et al., 2020), aquaculture conditions exhibit a more deliberate and targeted selection process.

1.4 Tools for SV detection and sequencing

Despite the significant evolutionary implications of SVs, several inherent challenges have historically hampered their comprehensive discovery and analysis compared to smaller sequence variants like SNPs (Alkan et al., 2011). The larger size of SVs (>50 bp) means they cannot be directly assayed by conventional approaches optimized for SNPs such as genotyping arrays (Wellenreuther & Bernatchez, 2018). Many SVs occur in repetitive regions of the genome difficult to characterize with short-read sequencing (Chaisson et al., 2019). SV breakpoints are also prone to imprecise mapping, complicating localization (Huddleston & Eichler, 2016). However, recent advances in high-throughput sequencing are enabling more accurate SV detection. Long-read technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) can directly span repetitive segments and with effort allow researchers to phase SVs (Sedlazeck et al., 2018; Leung et al., 2022). Emerging long-range scaffolding approaches produce highly contiguous de novo assemblies ideal for discovering SVs and resolving complex regions (Bachtrog & Charlesworth, 2022). Combining multiple modes of evidence from both long and short reads improves accuracy (Audano et al., 2019; Chakraborty et al., 2016). Hybrid assembly approaches that integrate multiple modes of evidence from both long-read and short-read sequencing technologies have emerged as a strategy for improving SV detection accuracy by leveraging complementary strengths (Fan et al., 2017; Sedlazeck et al., 2018). Long-read sequencing generates reads from tens to hundreds of kilobases in length, capable of directly spanning repetitive segments and capturing large structural variants in a single read (Jain et al., 2018). Steady advances have led to substantial improvements in data accuracy with PacBio's HiFi sequencing achieving error rates as low as 0.1-0.5% , representing over a 10-fold increase in accuracy compared to their previous chemistry. Oxford Nanopore Technologies has also made strides with their R14 flow cells, attaining median Q scores over 30 and estimated error rates below 5% for certain applications (Oehler et al., 2023). This is a notable enhancement from the 10-15% error rates routinely seen in early nanopore data (Ardui et al., 2018). While higher errors were acceptable in the past due to long-read sequencing's ability to characterise complex genomic regions, these new chemistry versions provide the best of both worlds - long reads with precision approaching short-read platforms. Short-read sequencing (e.g. Illumina) produces high accuracy data (error rates <1%) in vast volumes, but is limited in resolving complex regions and large variants beyond the read length (Chaisson et al., 2015). Integrating long and short reads in a hybrid assembly framework combines these complementary strengths to improve SV detection. Long reads can be aligned to an assembly graph or scaffolds constructed from more accurate short reads, encouraging the short reads to correct errors in the long reads (Koren et al., 2012). Discordant paired-end alignments and changes in read depth from short reads provide orthogonal signals to nominate SVs, which are then validated by alignment of long reads spanning the variant breakpoints (Chaisson et al., 2015). Joint analysis algorithms integrating multiple data types have shown dramatically improved sensitivity and precision. For instance, NanoSV achieved >95% precision and sensitivity by combining illumina, 10X Genomics linked-reads and PacBio long reads (Cretu Stancu et al., 2017). By overlaying multiple modes of evidence, hybrid approaches overcome limitations of individual technologies for more comprehensive and accurate SV detection.

The optimal approach for a given study depends on the biological question, samples, and resources available. Generating a high-quality reference assembly provides a crucial foundation for accurate SV calling (Rhie et al., 2021). For population-level analyses, moderate genome coverage (10-30X) of multiple individuals using short or long-reads can identify segregating SVs (Chiang et al. 2017). Comparing long-reads directly to a reference assembly is very effective for genotyping all SV types (Huddleston & Eichler, 2016). Combining orthogonal short and long-read evidence increases validation (Audano et al., 2019). Overall, applying complementary methods tailored to the study system enables rigorous SV analysis.

According to (Illumina), their sequencing platforms short-read sequencing platforms have continued to drive genomics over the past decade. Through iterative advances, Illumina sequencing now routinely generates hundreds of gigabases per run with available read lengths including 50bp, 100bp, 150bp or 300bp and accuracies exceeding Q30 (99.9%). This has enabled population-scale projects like the 100,000 Genomes Project (UK, 2018) and the All of Us Research Program (Denny et al., 2019). However, short-reads perform poorly in repetitive regions and cannot phase variants over long distances (Sedlazeck, Lee, et al., 2018).

1.5 Motivations for Studying Arctic Charr Structural Variation

The recent sequencing of the Arctic charr genome has opened exciting opportunities to explore the structural genomic variation underlying this species' adaptation. While numerous studies have speculated about the presence and adaptive significance of SVs, thorough investigation and validation has been lacking. Comprehensively evaluating SVs differentiated among divergent Arctic charr populations promises fundamental insights into the genomic basis of adaptability and its ecological speciation. From an applied perspective, Arctic charr represent an economically important emerging aquaculture species across Nordic regions, due to its delicate flavour and attractive pink colour (Helgadóttir et al., 2021; Pappas et al., 2023; Pappas & Palaiokostas, 2021). Overall, comprehensive analysis of Arctic charr structural variation has both basic and applied merits warranting in-depth investigation.

This thesis aims to develop a catalogue of structural variants in this species by optimizing and benchmark different computational tools designed for genome-wide SV discovery using short read data as input.

Achieving this objective will significantly advances understanding of the genomic understanding of Arctic charr's adaptability. Furthermore, the methodologies and resources produced will broadly empower structural variant studies in other non-model organisms.

1.6 Study Specie: The Arctic Charr

Aquaculture rearing and selective breeding of Arctic charr provides a model system to study genomic diversity, adaptation, and traits of commercial interest. Characterizing structural variation in farmed strains compared to wild populations can reveal impacts of domestication and artificial selection. Investigation of structural variants in the genomes of aquaculture stocks may uncover rearrangements related to productivity, growth, disease resistance, and other agro-economically important attributes.

Additionally, The Arctic charr (*Salvelinus alpinus*) has become a key model species for studying the genetic basis of adaptation, evolution, and early ecological speciation using genomic approaches. This cold-water fish lives in freshwater lakes, rivers, and coastal marine environments across the Northern hemisphere (Klemetsen et al., 2003). In contrast to most other salmonid species that have anadromous life cycles, Arctic charr generally reside in freshwater throughout their lives, with only limited migration

between interconnected water bodies, particularly in coastal river systems. The different charr morphs within these populations exhibit niche segregation and differences in feeding habits, which can lead to variations in parasite species and abundance (Jonsson & Jonsson, 2001).

Distinct morphs demonstrating adaptations to ecological niches characterized by water depth, available prey, and other limnological factors have evolved (Adams et al., 1998). For instance, dwarf, normal, and piscivorous morphs can be found coexisting and occupying distinct trophic roles in a single lake (Moccetti et al., 2019). Arctic charr thus represents an intermediate stage in the speciation continuum, with intraspecific phenotypic variation approaching interspecific differences normally seen between distinct species (Jonsson & Jonsson, 2001; Anders Klemetsen, 2010).

This signature adaptability of Arctic charr is encouraged by high levels of genetic variation accrued and maintained within genetically distinct populations (Brunner et al., 2001;Kapralova et al., 2011). Microsatellite studies have uncovered some of the highest genetic diversity. Significant allele frequency differences and fine-scale local adaptation are evident even across small spatial scales such as among interconnected lakes (Fraser & Bernatchez, 2005; Kapralova et al., 2011). Overall, the extensive phenotypic plasticity coupled with abundant genetic variation make Arctic charr a compelling model for genomic studies of rapid adaptation. Elucidating the genetic basis underlying this diversity promises to reveal key insights into the genomic architecture of adaptability and the drivers of ecological speciation.

The extensive genetic diversity observed within and among Arctic charr populations provides both opportunities and challenges for developing selective breeding programs. On one hand, the high levels of natural variation offer a rich resource to select for desired traits related to growth, disease resistance, flesh quality, and other aquaculture-relevant characteristics (Kapralova et al., 2011). The fine-scale local adaptation evident across interconnected environments also suggests genomic variants conferring plasticity that could enable adaptation to changing conditions (Fraser & Bernatchez, 2005). Capturing these adaptive alleles through genomics-guided breeding could produce resilient strains tailored to specific aquaculture settings.
However, the same diversity also poses challenges. High genetic differentiation between populations means breeding values for traits estimated in one population may not predict performance well in other populations (Sae-Lim et al., 2016). Connectedness between breeding populations is required for effective genomic selection. The diversity may also reflect high levels of inbreeding within small isolated populations, necessitating careful management of inbreeding depression in a breeding program. Nevertheless, applying genomic tools to understand the architecture and selective pressures shaping diversity will be key to harnessing the variation through selective breeding for continued advancement of Arctic charr aquaculture.

1.7 Emergence of Genomic Resources for Arctic Charr

Early genomic research in Arctic charr focused on targeted investigations of specific molecular markers, genes, and chromosomal regions (e.g. Brunner et al., 2001; Ferguson et al., 1991). Technological advancements in high-throughput sequencing over the past decade have enabled genome-wide analyses to interrogate genetic architecture, adaptation, and evolutionary relationships in this species (Christensen et al., 2021). For instance, SNP arrays and genotyping-by-sequencing approaches have empowered population genomic studies revealing neutral and adaptive genetic structure among Arctic charr populations (Bourret et al., 2013; Kapralova et al., 2011). Additionally, recent years have seen major improvements in Arctic charr genome resources to facilitate more detailed genetic investigations.

1.8 Structural Variant Detection Pipeline used in this research

In this research, structural variants were identified using three complementary detection tools: Delly v1.1.6 (Rausch et al., 2012), Manta v1.6.2 (Chen et al., 2016), and Smoove v0.2.5 (Pedersen, 2020). Employing multiple tools accounts for the complexity of accurately detecting SVs and addresses inherent limitations of individual programs.

The rationale for a multi-tool approach stems from SVs' intricate nature and current computational challenges in reliably identifying them. Each tool utilizes distinct algorithms suited to detecting certain SV types and sizes, leading to variances in sensitivity, specificity, and efficacy. For instance, one tool may excel at finding large SVs while another is optimized for complex genomic rearrangements. By integrating strengths across tools, the research aims to improve overall SV detection reliability and comprehensiveness.

The selection of Delly, Manta, and Smoove was based on their distinct methodologies and proven performance in previous studies. Delly demonstrates robust detection across SV types like deletions and translocations. Manta is optimized for precisely calling SVs and indels, especially in cancer genomics. Smoove efficiently handles large datasets while maintaining high SV calling accuracy.

1.8.1 DELLY v1.1.6

The C++ integrated software programme DELLY combines split-read analysis paired-end mapping, making it an excellent SV detector. In order to successfully detect and genotype SVs, the tool employs the 'delly call' command. A reference fasta file and a BAM file that has been sorted are required for this procedure. This command demonstrates how to use DELLY to detect several types of SVs, such as deletions (DEL), duplications (DUP), inversions (INV), and insertions (INS) and breakends (BND), A breakend is one breakpoint or disrupted endpoint of a structural variant (SV). Most SV detection involves finding paired breakends that reveal the variant type and size. However, limitations can result in single breakends being detected without identifiable pair. These lone breakends indicate a genomic SV but lack information to categorize the specific type. Hence, they classified as BND. You may easily convert the resultant data from its original BCF format to the simpler and more readable VCF format with Bcftools (H. Li, 2011). Detailed information on SVs, including their types, chromosome locations, genomic positions, reference and changed sequences, quality scores, and other pertinent parameters, are included in the final result.

1.8.2 Manta v1.6.2

Manta employs a robust C++ programming to lead the field in precise SV identification from genomic sequences. Its specialized diploid-aware algorithm efficiently analyzes both short-read Illumina and long-read sequencing data to pinpoint SVs with accuracy. By integrating paired-end and split-read evidence, Manta can adeptly detect complex events like Breakends and tandem duplications. Its streamlined "manta" command-line workflow allows straightforward inputs of alignment files in BAM or CRAM formats to initiate intricate SV analysis.

1.8.3 Smoove v0.2.5

Based on go-based programming, this software demonstrates exceptional proficiency in streamlining the complicated procedure of structural variation (SV) detection. Smoove functions by effectively utilizing preexisting tools such as Lumpy and SVTyper, simply integrating them into a single workflow. The collaboration leads to a highly effective approach for accurately detecting deletions, duplications, and other SV categories. The command-line interface of the tool, known as 'smoove', simplifies the execution of SV analysis by simply using input BAM files and a reference genome. The output generated by Smoove, which is formatted as Variant Call Format (VCF), provides a substantial amount

of data. This comprehensive description allows researchers to obtain a concise and accurate understanding of the structural variants, including their positions and distinctive features. In this thesis, we applied a robust computational strategy to detect structural variants (SVs) in Arctic charr (*Salvelinus alpinus*), leveraging the high-throughput SV detection capabilities of Smoove. We executed a multi-step pipeline that included individual SV calling, stringent quality control, and cross-caller variant comparison to distill a consensus SV profile.

1.9 Anticipated Research Outcomes and its significance

A key outcome of this research is to benchmark alternative SV calling programs for their accuracy and efficacy. By assessing different SV calling approaches, this study will help refining protocols for variant detection in genomics research by utilizing short read data as input.

Another major goal is constructing a comprehensive catalog of SVs in Arctic charr. It will enable researchers and breeders to pinpoint genetic markers associated with desirable traits, thereby optimizing breeding strategies for improved productivity and sustainability in aquaculture.

Additionally, the findings could potentially be used in comparative genomics analysis with other salmonids (farmed or wild). Contrasting SVs across related species will provide insights into the genetic architecture of shared and unique traits, advancing knowledge in evolutionary biology and aquaculture.

Chapter 2 Methods and Materials

2.1 Sample Collection

The samples utilized in this study consisted of 30 Arctic charr (*Salvelinus alpinus*) obtained from a commercial aquaculture facility in Norway. The library preparation and DNA sequencing were performed at the Norwegian Sequencing Center. The 30 Arctic charr DNA libraries were sequenced using an Illumina HighSeq-4000 platform to generate 150 bp paired-end reads with ~30X genome coverage depth.

2.2 Coverage Analysis by mosdepth (v 0.3.5)

Genomic coverage for binary alignment/map (BAM) files was computed using Mosdepth (version 0.3.5) (Pedersen & Quinlan, 2018), an ultra-rapid tool designed and optimized specifically for calculating depth distribution across genomic regions. Mosdepth demonstrates superior computational performance over other coverage calculators like SAMtools depth and BEDTools genomecov by utilizing a concatenated hash table and space-efficient binary format (Pedersen & Quinlan, 2018).

2.2.1 Workflow Implementation

An automated Bash script (run_mosdepth.sh) was developed to enable batch processing of Mosdepth across 30 BAM files. The script contained a loop to iterate through each input BAM file stored in the specified directory. For each BAM file, the script executed Mosdepth with the following optimized parameters:

Interval size: 100 bp (non-overlapping 100 bp windows used to calculate coverage metrics, balancing resolution and speed)

Fast mode: Enabled to accelerate computation

Per-base report: Disabled to reduce output volume

The Arctic charr reference genome assembly (Salpinus_reference/Arthur_CHR_polished.fasta) was provided as a benchmark to calculate coverage aligned to genomic coordinates. The scripts and its documentation for run_mosdepth.sh is provided in my GitHub repository (Syed-NMBU, 2023), specifically at this link.
In the output Analysis, mosdepth output files containing coverage distributions, regional metrics, and summary statistics.The coverage data visualization was performed using a custom R script. This script, developed for generating chromosomal coverage plots for each sample, utilized various R packages including ggplot2, dplyr, viridis, ggrepel, and patchwork. The script facilitated the creation of multi-panel figures, such as 'Average Coverage Depth Across Chromosomes' (Figure 3), 'Comprehensive Chromosomal Coverage Depth Analysis for Samples 01 to 10' (Appendix C, Figure C1), and 'Sequential Chromosomal Coverage Depth Profiles for Samples 11 to 20' (Appendix C, Figure C2), which are essential for assessing variability in coverage depth across samples. The script and its documentation are accessible on my GitHub repository at this link.

2.3 Delly: v1.1.6 (Chen et al., 2016)

2.3.1 Computational Framework and Dataset

Analysis was conducted using the Delly SV caller (version 1.1.6) (Rausch et al., 2012), specifically optimized for high-throughput sequencing data. The integrity and uniformity of the reference genome sequence, Salvelinus alpinus reference genome (Arthur_CHR_polished.fasta), were maintained across all samples to ensure consistency in SV detection.
SV calling was executed through a batch script (1_run_delly_analysis.sh), utilizing the singularity container technology to encapsulate the Delly environment, thereby ensuring reproducibility across computational environments. Binary Call Format (BCF) files were generated for each sample, followed by conversion to Variant Call Format (VCF) for downstream analysis using BCFtools. The Delly program was executed with default parameters. The script is accessible for review and use at my GitHub repository (Syed-NMBU, 2023). In the data processing phase, a custom script, 2_delly_vcf_filtering.sh (available at Syed-NMBU, 2023) was utilized. This script was specifically designed to filter Variant Call Format (VCF) files generated by Delly, focusing on calculating and annotating structura variant lengths (SVLEN) where not initially provided, and filtering variants based on size criteria, particularly those exceeding 50 base pairs. This step was crucial for ensuring the accuracy and relevance of the structural variant data used in subsequent analyses.

2.3.2 Data Normalization and Merging

The normalization step corrected for any inconsistencies in reference allele representation across the individual VCF files, facilitating a coherent merged dataset. The BCFtools `'norm'` command was employed for this purpose, ensuring the harmonization of SV calls prior to merging. The merging process was meticulously logged, capturing every step and error to ensure transparency and traceability of the pipeline execution, BCFtools
`'merge'` command was used. The script with detail merging commands available here.
The filtered, normalized SV datasets were subject to detailed statistical analysis using BCFtools (H. Li, 2011) and custom R scripts to quantify the frequency and types of variants across the genome and for visualization with R code (Available at: Syed-NMBU, 2023)
Data Management, all intermediate and final datasets were systematically stored on a project-designated SCRATCH storage space on Orion, adhering to a pre-defined directory structure and file naming convention to facilitate efficient data management and retrieval.

2.3.3 Materials

Genomic Data

- Arctic charr sequencing data: 30 samples (BAM format)

Software and Tools

- Delly SV caller (version 1.1.6) (Rausch et al., 2012)
- BCFtools (version 1.12) (H. Li, 2011)

- Singularity container (delly_1_1_6_ha41ced6_0.sif) (Kurtzer et al., 2017)
  Reference Genome
- Salvelinus alpinus reference genome (Arthur_CHR_polished.fasta)
  Computational Resources
- Orion high-performance computing cluster
  Scripts and Commands
  Detailed scripts used for SV calling, quality filtering, normalization, and merging are available in (https://github.com/Syed-NMBU/StructVar_ComparativePipeline/tree/main).

2.4 Manta v1.6.2 (Chen et al., 2016)

2.4.1 Computational Workflow

Structural variants were identified using Manta (version 1.6.2), a tool designed for the rapid and precise analysis of structural variants in genomic data. Each Arctic char sample was processed separately, yielding three primary VCF outputs: candidateSmallIndels.vcf, candidateSV.vcf, and diploidSV.vcf, each uniquely annotated with genomic coordinates. The computational process was conducted on a high-performance computing cluster (HPC) using the SLURM workload manager. Specific parameters included a single task (--ntasks=1), 15 CPUs per task (--cpus-per-task=15), 60G memory allocation (--mem=60G), and a run time limit of 48 hours (--time=48:00:00). This setup was crucial for managing the computational demands of processing high-throughput sequencing data. Scripts written in Bash were used for the Manta workflow (Available at: Syed-NMBU, 2023), including configuration (configManta.py), execution (runWorkflow.py), and post-processing steps. Significantly, the output VCF files from Manta already contained the SVLEN (structural variant length) column. For data refinement, the `bcftools view` command was applied to exclude structural variants smaller than 50 base pairs, ensuring the retention of variants larger than this threshold. This filtration was vital in isolating and analyzing structural variants of substantial size, for the accuracy and relevance of this genomic study with other tools. The bcftools v1.12 (Danecek et al., 2021) is used for filteration of all files (Appendix A, Script A.1)

2.4.2 Normalization and Merging of VCFs

Post-processing involved normalizing primary VCF files for each sample using BCFtools `norm`. This step was critical in unifying the SV calls, addressing multi-allelic records, and standardizing SV representations across the dataset. After that a custom script (VCF_merge) was utilized in which BCFtools merge command is used to merge the normalized VCF files from all 30 samples, creating a consolidated VCF file containing all unique SVs from 30 samples detected by Manta.

2.4.3 Materials

Genomic Database

A reference genome for Arctic charr (Salpinus_reference) previously constructed at CIGENE was made available to this work and used for read mapping and SV calling.

2.5 Smoove v0.2.5 (Pedersen, 2020)

In this thesis, we applied a robust computational strategy to detect structural variants (SVs) in Arctic charr (*Salvelinus alpinus)*, leveraging the high-throughput SV detection capabilities of Smoove. We executed a multi-step pipeline.

2.5.1 Computational Workflow

The SV calling was initiated with individualized processing of samples. A unified automated script run_smoove.sh was build to process samples 1 through 30. Each sample underwent SV calling where Smoove was invoked within a controlled singularity container environment, ensuring computational reproducibility. The BAM files, served as the input, and the SV calling was conducted in a sample-specific directory to streamline data organization and log collection. The ouput VCF files are filtered to

remove SVs in the range from 1 to 49 bp, by using `bcftools view` command (Appendix A, Script A.1).

2.5.2 Normalization and Merging

BCFTOOLS, a renowned tool for its precision in VCF file manipulation, was employed to normalize and standardize the genomic coordinates and allele representations within the merged VCF file against the reference genome, ensuring uniformity across the dataset. Normalization served as a pivotal step, rectifying allele inconsistencies and facilitating subsequent genomic queries. Following SV detection and normalization, a custom script (vcf_merge) merged the individual .vcf.gz files into a singular, cohesive VCF file for further SV analysis.

2.5.3 Materials

Bioinformatics Tools

- Smoove v0.2.6 (Pedersen, 2020)
- BCFtools (H. Li, 2011) (for VCF normalization and merging)
- **Computational Resources**
  Orion

2.6 Comparing Tools by SURVIVOR v1.0.7 (Jeffares et al., 2017)

We utilized three structural variant (SV) detection tools to comprehensively identify SVs from the Bam files and for comparison, DELLY, Manta and Smoove. Each tool was independently applied to the 30 Bam files, resulting in 90 VCF files containing putative SVs identified per sample. To construct an integrated overview of SVs across samples and tools, we utilized SURVIVOR to harmonize and merge the SV calls

Per-Tool Merging, The SURVIVOR merge command was utilized to consolidate 30 VCF files from each structural variant (SV) detection tool into a single, unified VCF file per tool. The unique SV catalogue results from Delly, Manta and Smoove with each VCF file. The key parameters were a max breakpoint distance of 100 bp (to merge proximal SVs) and a minimum tool support of 1 (to retain all SVs regardless of sample overlap).

Final Merging, The 3 merged tool-specific VCF files were integrated using a final SURVIVOR merge step to create a comprehensive dataset with all identified SVs with one final VCF. The same inclusive parameters were used as before.The final merged VCF file was subjected to comparative analysis to identify common SVs supported by multiple tools.

Chapter 3 Results

3.1 Coverage analysis

Sequencing was performed on 30 arctic charr samples. After aligning reads to the reference genome sample coverage statistics were calculated using the mosdepth package. The analysis generated global distribution files, region-specific distribution files, summary statistics files, and per-window depth files for each sample.

3.1.1 Mean Coverage depth per sample

With an estimated reference genome size of 2.147 gigabases (Gb), the mean sequencing depth across the 30 samples (Sample_01 to Sample_30) ranged from 10X to 14.4X coverage. The overall average sequencing depth across all 30 samples was calculated to be 11.19X with a standard deviation of 0.98X. Most samples displayed mean depths within this typical range. However, two samples, Sample_25 and Sample_26, exhibited markedly higher mean depths of 13.61X and 14.41X respectively. Additionally, a subset of samples showed slightly lower than average mean depths below 10X, specifically Sample_01 (9.97X), Sample_09 (9.96X), Sample_11 (9.97X), and Sample_19 (9.96X) (Figure 2).

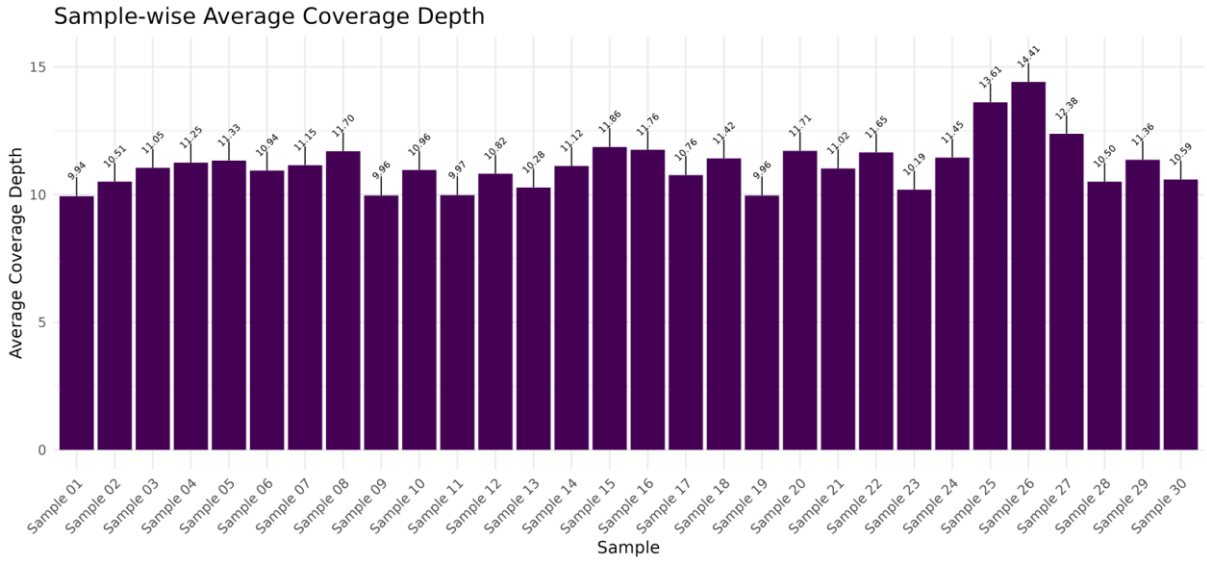


***Figure 2**. Sequencing coverage of 30 Arctic charr Samples*

3.1.2 Per chromosome coverage statistics

Each sample comprising of 39 chromosomes (sal01 to sal39). The chromosome sizes ranged from 25Mb (25,575,790; sal39) to 106Mb (106,341,530; sal10). Across samples, coverage depth per chromosome varied from 7.9X (sample 1, sal 25) to 30.56X (sample 26, sal 17), indicating moderate variability between samples and substantial variability between chromosomes. The mean coverage depth across all chromosomes and samples was 11.38X. Of 39 chromosomes, 17 (sal17) showed high coverage depth in all samples, ranging from 16.41X (sample 23) to 30.56X (sample 26) (Figure 3).

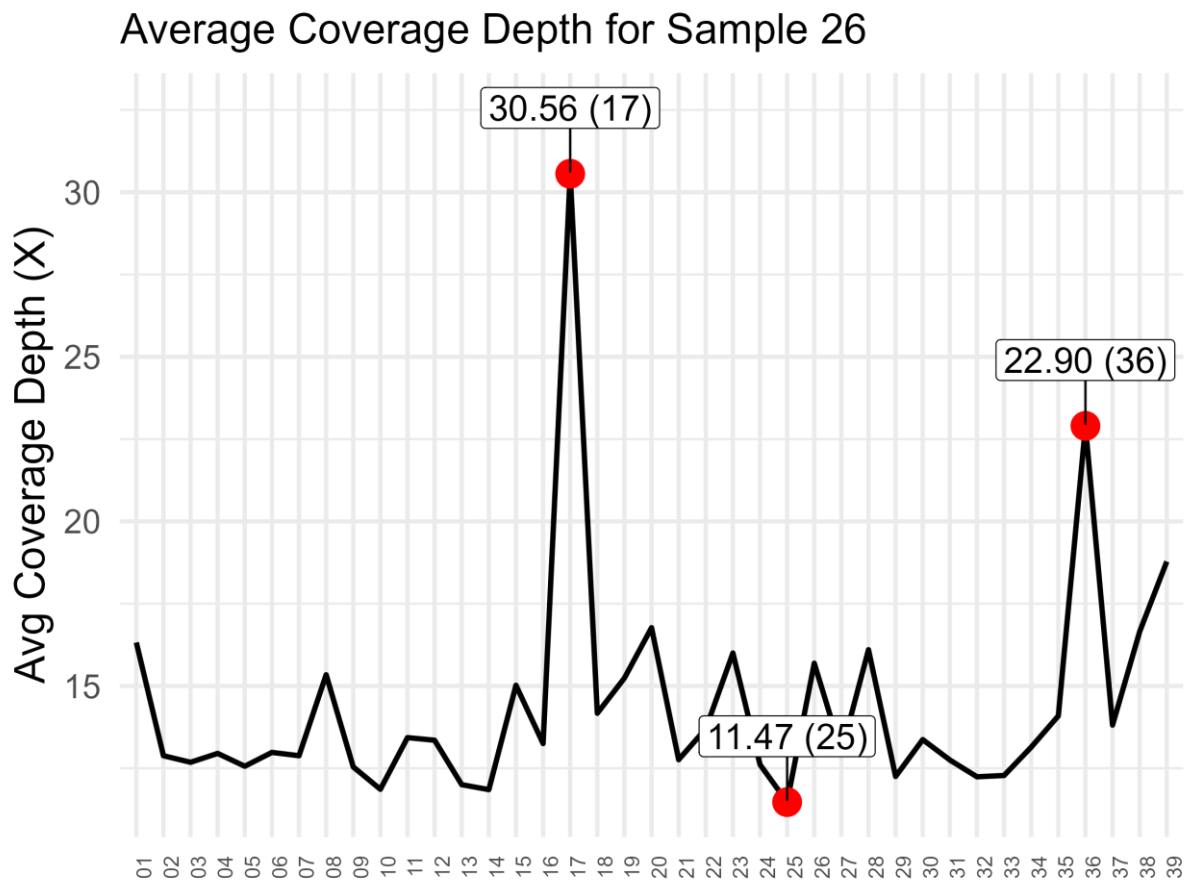


Figure 3. Chromosomal coverage depth analysis for Sample 26 *with maximum, second maximum and minimum mean coverage points representing with red dots.*

Several additional large coverage spikes were noted for sample 14's chromosome 19, reaching 12.53X depth, and sample 25's chromosome sal20 with 16.05X. Figure 3 presents the average coverage depth across chromosomes for sample 26, with sal17 showing high coverage depth across all samples. This trend is consistent across all samples as shown in Figures C1, C2 and C3 in the Appendix C.

3.2 Delly, SV detection

Delly identifies five classes of structural variations (SVs): deletions (DEL), duplications (DUP), inversions (INV), insertions (INS), and breakends (BND) as shown in figure 4, which indicate the presence of a DNA rearrangement with broken ends. After filter-based normalization, specifically filtering out SVs smaller than 50bp in size and normalizing the vcf files retaining variants greater than 50bp, between 70,866 and 128,116 SVs were detected per individual. SVs were detected per individual with an average of 97,341.

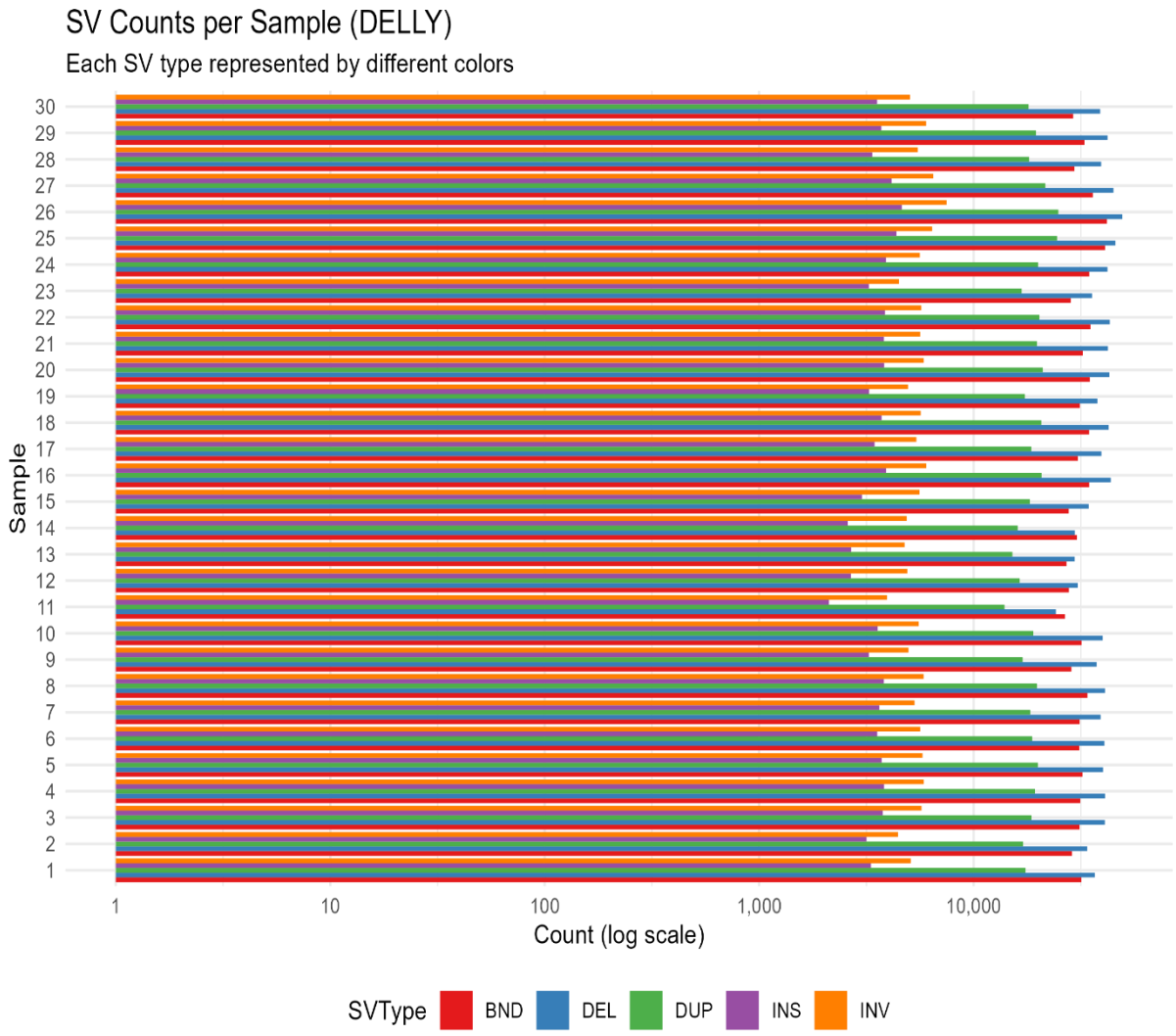


Figure 4. Distribution of Structural Variations by Type Across 30 Samples: *This bar chart categorizes the number of structural variations (SVs) identified in 30 samples using the DELLY SV caller. Each bar represents a single sample, with the length corresponding to the count of SVs detected. The structural variations are color-coded according to type, providing a comparative overview of the SVs.*

Across all samples, Deletions were the most common SV class (mean 39,574), representing genomic deletions. BND is the second most variant type (mean 31,573), followed by fewer duplications (mean 18,935) and insertions (mean 3,658) and inversions (mean 5,315). The ratio of deletions to duplications was ~2:1 in most samples.
The distribution and total number of structural variants (SVs) identified by Delly were assessed on a per chromosome basis (Figure 5).

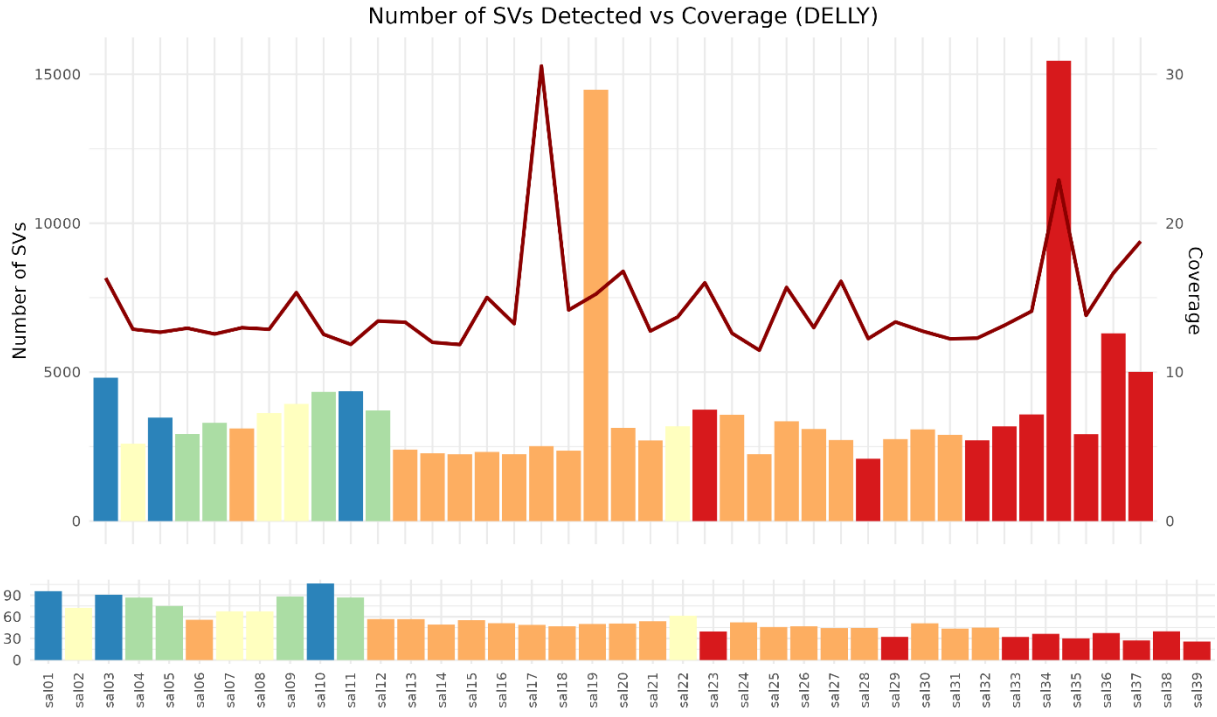


Figure 5. Structural variation distribution and coverage across chromosomes in a single sample *This graph illustrates the number of structural variations (SVs) detected across 39 chromosomes in a single sample alongside the corresponding sequencing coverage (line). The y-axis on the left quantifies the SV count pr chromosome, while the y-axis on the right measures the sequencing coverage depth. The bar graph's color gradation represents individual chromosomes, and the line graph overlays the coverage, offering insight into the density of SVs in relation to sequencing depth, as determined by DELLY analysis.*

Across all samples, chromosomes 19 and 36 consistently contained the greatest number of SVs compared to all other chromosomes, with an average SV count of 11,500 and 11,800 respectively versus 2500 for all other chromosomes combined.
In contrast, chromosomes 14, 29 and 33 showed the lowest SV numbers on average, ranging from 1,209 to 1,908 SVs across samples.
However, a subset showed more variable numbers of SVs, with some samples exhibiting up to 3-fold differences versus averages. For example, chromosome 17 ranged from 1,116 SVs in sample 11 to 3,040 SVs in sample 26. These findings are graphically presented in Appendix D, Figure D1, which provides a detailed visual comparison of SV frequencies across all chromosomes and samples.

Delly Structural Variant Merging

Delly was used to call SVs in each sample individually, and gave the total sum of SVs across all 30 samples as 2,885,011. Many of these calls result from the same SV being detected in multiple samples. We collapsed the raw SVs into a merged, consolidated dataset using the bcftools merge command which reduced the total Delly unique SV count to 1,550,299.

3.3 Manta, SV detection

Manta was utilized to detect structural variants (SVs) including deletions, duplications, insertions and breakends (Figure 6) . Between 17,185 (sample 23) to 30,946 (sample 26) total SVs were detected per individual fish genome.

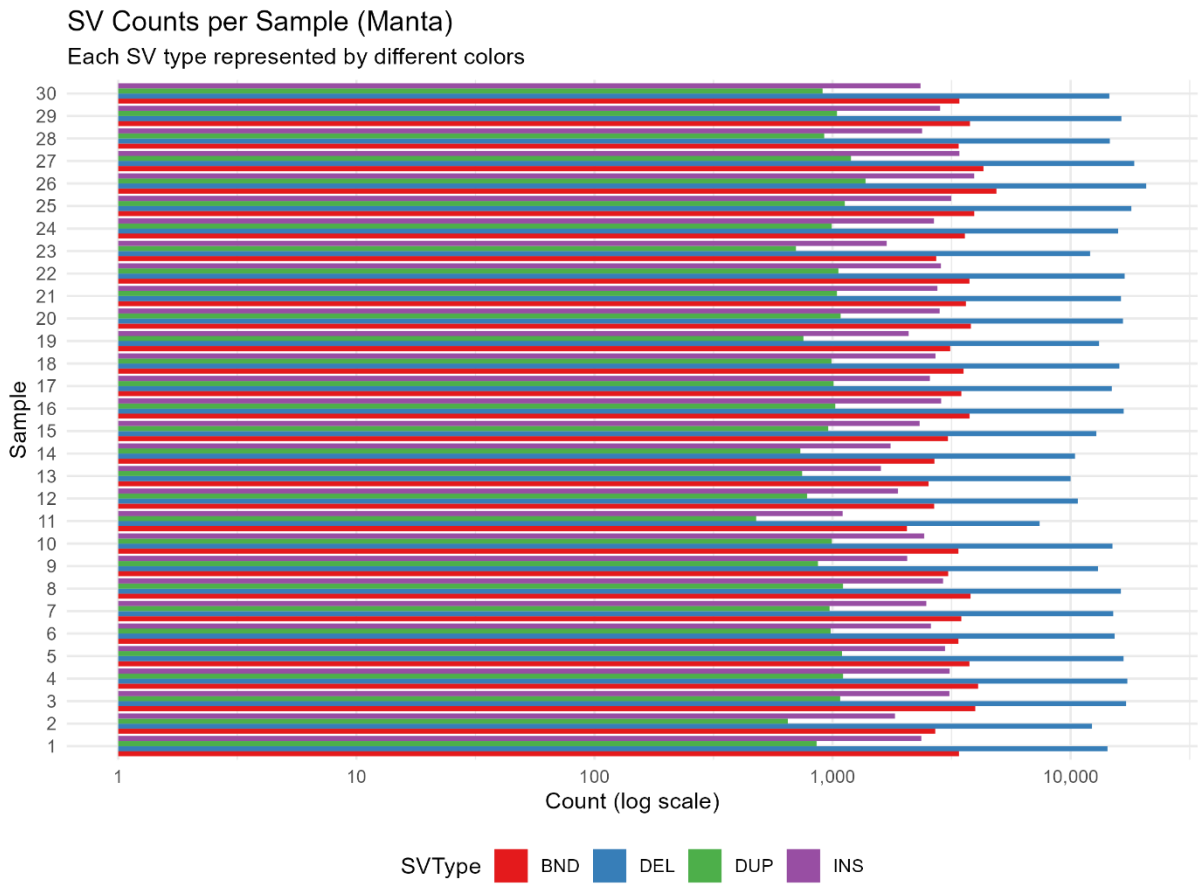


Figure 6. Distribution of detected structural variations by type across different samples*: This bar chart shows the number and types of structural variations (SVs) identified across 30 samples using Manta. Each bar represents a Sample, segmented by color to indicate the quantity of different SV types detected. The x-axis quantifies the number of SVs, while the y-axis lists the samples. The color key at the bottom corresponds to the different SV types, facilitating a comparative analysis of the structural variation distribution between samples..*

Deletions represented the predominant SV type, ranging from 7,407 (sample 11) to 20,756 (sample 26) per genome. Duplications, insertions and breakends occurred at lower but variable frequencies, together comprising 15-25% of SVs depending on sample. No inversions were identified across any sample by Manta. A subset of genomes showed substantially higher SV counts, including samples 25 and 26 containing 26,216 and 30,946 total SVs respectively. In contrast, samples 11 and 23 displayed

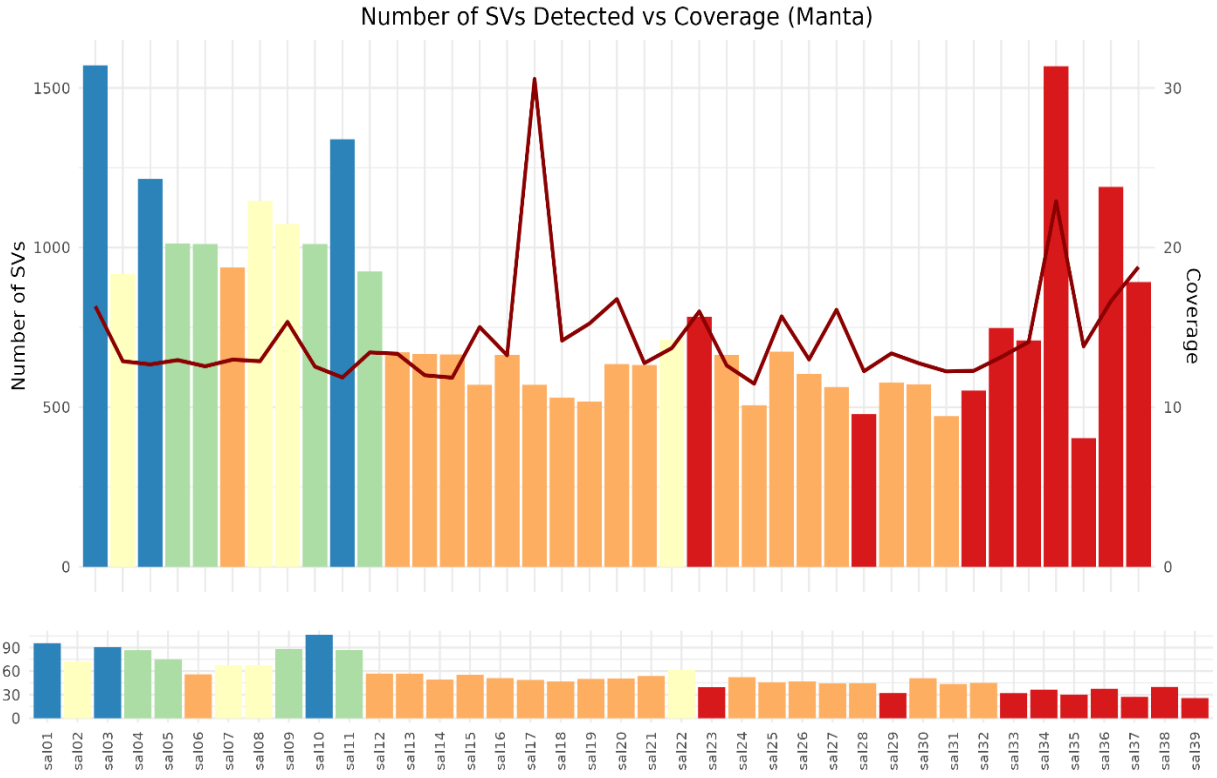


Figure 7. Correlation Between Structural Variations and Sequencing Coverage for Chromosomes in a Single Sample: *This figure represents a dual-axis chart where the bar graph depicts the number of structural variations (SVs) detected across chromosomes for a single sample, and the line graph shows the corresponding sequencing coverage. The x-axis mentions the chromosomes, while the left y-axis scales the number of SVs detected, and the right y-axis measures the sequencing coverage. The bars are color-coded to differentiate between the chromosomes, and the coverage line provides context to the SV detection efficiency, highlighting the variability and density of SVs in relation to the sequencing coverage.*

comparatively fewer SVs. The total SVs per chromosome were quantified to assess inter-chromosomal variability (Figure 7).

In Appendix D, Figure D2 provides a detailed visual summary of the structural variations observed across all samples. SVs showed non-random genomic distribution, with certain chromosomes emerging with high number of SVs. Significantly, chromosome 36 with the highest SV burden (Appendix E, Figure E1), ranging from 711 SVs (sample 11) to 1,568 SVs (sample 26) (Figure 7). Other SV-rich chromosomes included 1, 3, 4 and 38, each averaging 900-1,300 SVs across samples.

In contrast, chromosomes 13, 14, 25, 29 and 33 showed the lowest SV numbers, generally below 550 SVs per sample. While the chromosome with most SVs remain consistent, total SVs varied from 17,185 (sample 23) to 30,946 (sample 26) per fish. Samples 25 and 26 again displayed high SV content, implying heightened genomic variability.

Manta Structural Variant Merging

To determine number of unique SVs detected by Manta, individual outputs were consolidated and redundancy (re-calls of the same SV across samples) removed. While the un-consolidated collection totaled 652,365 SVs, the aggregated merged file reduced this number to 221,309. The merged Manta variants could then be integrated and contrasted with the merged 1.5 million Delly and Smoove callsets using Survivor, giving a broad portrait of structural mutation spectra in the population.

3.4 Smove, SV detection

Smoove can detect four classes of structural variation including deletions, duplications, inversions, and breakends (Figure 8). Between 7,746 (sample 11) and 20,789 (sample 26) total SVs were detected per fish genome.

Across all samples, a total of 449,767 SVs were identified. Deletions were the most prolific SV class, accounting for 85% of calls (n= 384,727). Duplications occurred at a 6-fold lower frequency (n=30,962), followed by much lower inversion (n=5,454) and breakend (n=28,624) numbers.

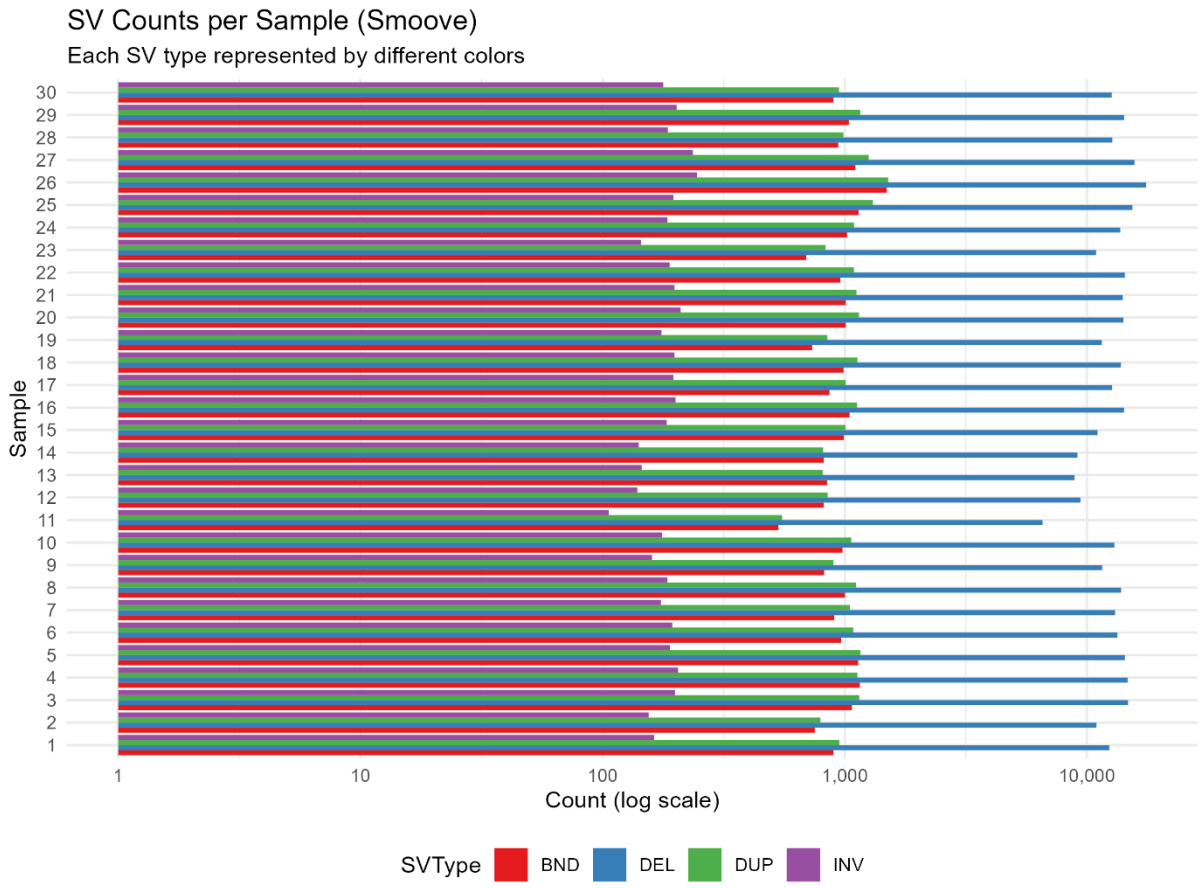


Figure 8. Distribution of Structural Variations by Type Across 30 Samples*: This figure provides an overview of the structural variations (SVs) detected across 30 samples by Smoove, with each sample represented by a group of four bars. Each bar within a group corresponds to a different type of SV, color-coded for distinction. The y-axis indicates samples, while the x-axis represents the count of each SV type. The visualization facilitates a comparative analysis of the prevalence and variety of SVs in each sample.*

Deletion counts per individual aligned with previous callers, ranging from 6,557 (sample 11) to 17,545 (sample 26). Samples 25 and 26 again displayed exceptional SV content, with 15,413 and 17,545 deletions called respectively. As with Delly and Manta, samples 11 and 23 showed markedly fewer deletions.

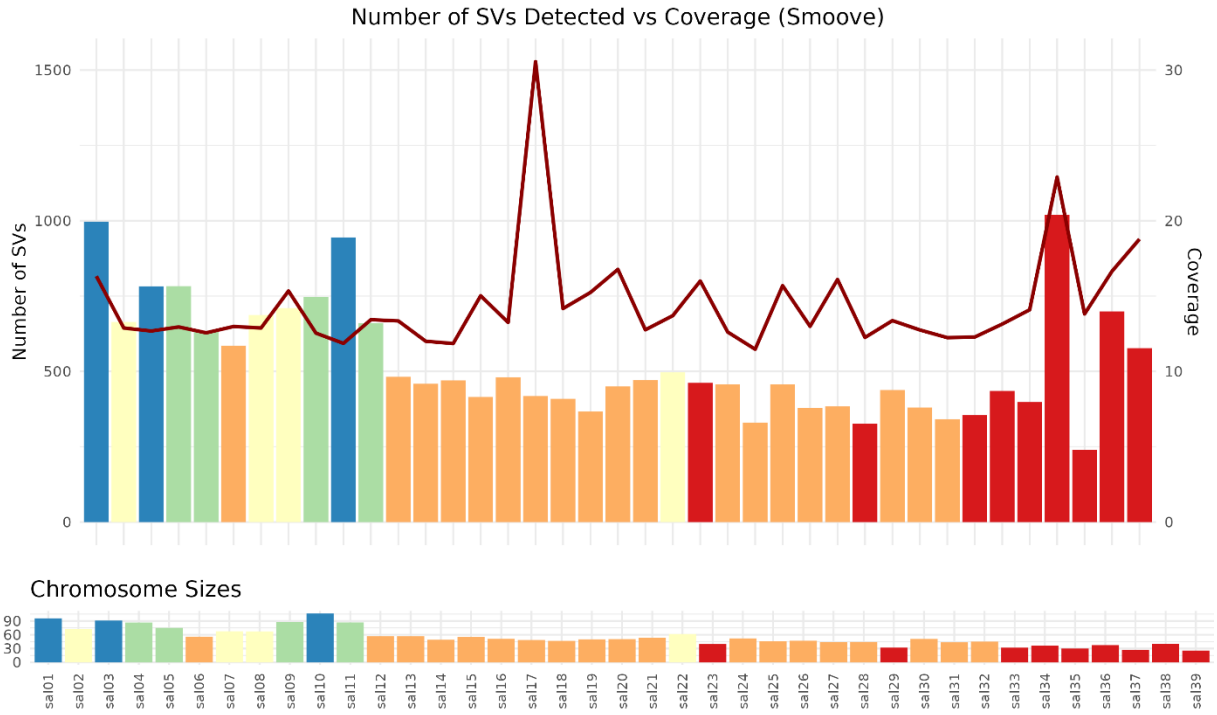


Figure 9. Number of Structural Variations and Sequencing Coverage per Chromosome in a Single Sample*: This graph illustrates the count of structural variations (SVs) detected across each chromosome in a single sample, detected using Smoove. The bar graph indicates the number of SVs per chromosome, with each bar's color representing a different chromosome on basis of their sizes, Blue are the largest, red are the smallest in size. Overlayed on the bar graph is a line plot that depicts sequencing coverage, allowing for a direct comparison between SV frequency and coverage depth on a per-chromosome basis. The bottom panel displays the relative sizes of each chromosome, providing context for the SV distribution in relation to chromosome length. This figure encapsulates the correlation between chromosomal architecture and the incidence of structural genomic variations within the sample.*

Prominently, sal 36 consistently emerged as an SV dense chromosome, with events ranging from 501 (sample 11) to 1,420 (sample 26), as illustrated in figure 9 for single sample, and in appendix D, figure D3 for all samples. Several other chromosomes also harbored extensive structural mutations, including chromosomes 1, 3, 10 and 38. Chromosome 1 displayed the top range, spanning 403 SVs (sample 11) up to 1,571 variants (sample 26). In contrast, chromosomes 13, 14 and 29 comprised SV deserts, typically containing well under 400 events.

Notably, while the ranking of chromosomal counts was stable across samples, the degree of SVs fluctuated. Particularly for high SV count regions like chromosome 36, samples 25 and 26 reached over 1,400 events, whereas sample 11 showed only 501 variants - a 3-fold difference. While exhibiting generally consistent patterns, Smoove reported lower raw SV numbers than prior methods. This likely reflects algorithmic differences in variant classification thresholds.

Smoove structural variant merging

Across all samples, the raw Smoove outputs contained 449,767 total SVs when aggregated across individual sample callsets. As with Delly and Manta, this initial tally double-counted signals shared between closely related fish. Using the same bcftools merge approach, the per-sample Smoove VCF files were consolidated into a unified variant set. Collapsing common variants reduced the overall events to 116,940 in the merged output.

The set of 116,940 non-redundant calls represents Smoove's highly specific portfolio of highest-confidence predictions, which can be intersected with the more sensitive Delly and Manta approaches.

3.5 Delly, Manta and Smoove VCF filter

Delly output underwent standardized filtering to mark variants as "PASS" or "LowQual" prior to comparison with Manta and Smoove. Across all Delly calls in the merged file, 270,406 (19%) variants were assigned a PASS filter status, while the remaining 1,279,893 (81%) calls were labelled as LowQual. Unlike Delly, Manta classified variants dichotomously as either "PASS" or non-PASS, without a specific "LowQual" designation. Manta calls comprised 370,334 (57%) PASS and 282,031 (43%) non-PASS variants. In the Smoove output VCF file, none of the 116,940 total variants were specifically designated with a "PASS" label in the FILTER column. Rather, all calls displayed alternative entries lacking a "LowQual" or failed designation.

Table 1. Quality Assessment of Structural Variant (SV). This table outlines the quality categorization of SVs identified by three tools: Delly, Manta, and Smoove. The 'Total SVs' column indicates the total number of SVs detected by each tool. 'PASS' represents the number of SVs classified as high quality, 'Low Quality' indicates SVs deemed of lower reliability, and 'No Quality Status' shows SVs without a specified quality assessment. The 'PASS Percentage' column provides the percentage of total SVs that met the high-quality criteria.

Tool	Total SVs	PASS	Low Quality	No Quality Status	PASS Percentage
Delly	1,550,299	270,406	1,279,893	-	18%
Manta	221,309	77,877	-	143,432	35%
Smoove	116,940	116,940	-	-	100%

3.6 SURVIVOR, Per-tool merging

The trio (Delly, Manta and Smoove) respectively contained 1,054,878 , 164,587 and 78,268 SVs (Table 2).

Table 2. Summary of Structural Variant (SV) Counts from Initial Detection to Post-Merge and SURVIVOR application. The table outlines the raw SV counts detected by each caller (Delly, Manta, Smoove), the counts following the merging process with BCFtools to eliminate redundancy, and the final counts after integration with the Survivor tool to establish a consensus across tools

SV Caller	Raw Output (SVs)	Post-Merge Output (SVs) by BCFtools	Integrated Callset (SVs) by Survivor
Delly	2,885,011	1,550,299	1,054,878
Manta	652,365	221,309	164,587
Smoove	449,767	116,940	78,268

Table 3. Comparative Analysis of SV callers. This table presents a summary of structural variants (SVs) detected by three tools: Delly, Manta, and Smoove, categorized into five types—Deletions (DEL), Duplications (DUP), Insertions (INS), Inversions (INV), and Translocations (TRA). The counts of each SV type are listed alongside the total number of SVs detected by each

tool. The 'Command Used for Merging' column specifies the SURVIVOR command parameters utilized for merging SVs from different files into a consolidated VCF file for each tool.

Tool	DEL	DUP	INS	INV	TRA	Total	Command Used for Merging
Delly	245,368	322,713	1,798	77,824	407,175	1,054,878	`SURVIVOR merge delly_files_list.txt 100 1 1 1 0 50 merged_delly.vcf`
Manta	92,421	10,349	17,331	9,676	34,831	164,608	`SURVIVOR merge manta_files_list.txt 100 1 1 1 0 50 merged_manta.vcf`
Smoove	59,245	9,845	-	5,385	3,793	78,268	`SURVIVOR merge smoove_files_list.txt 100 1 1 1 0 50 merged_smoove.vcf`

Furthermore, The outputs of the three structural variant detection tools Delly, Manta, and Smoove were compared across different bin sizes among each other ranging from 0-50 bp to over 1 million bp (Figure 10). In the 0-50 bp bin, Delly detected 407,175 variants (38.59% of its total) (Appendix B, Table 1), while Manta identified 34,852 (21.17% of its total) and Smoove found only 3,793 (4.85% of its total). As the bin size increased, the proportion of variants detected by Manta and Smoove increased, while Delly's proportion decreased. In the 50-100 bp bin, Manta identified the most variants (40,038, 24.32% of its total). Smoove detected the highest percentage of its total variants in the 1000-2500 bp range, finding 19,835 variants (25.35% of its total). Meanwhile, Delly maintained higher SV counts across all bins (Figure 10, A), but the proportion of its total variants detected in each bin declined steadily as bin size increased (Figure 10, B).

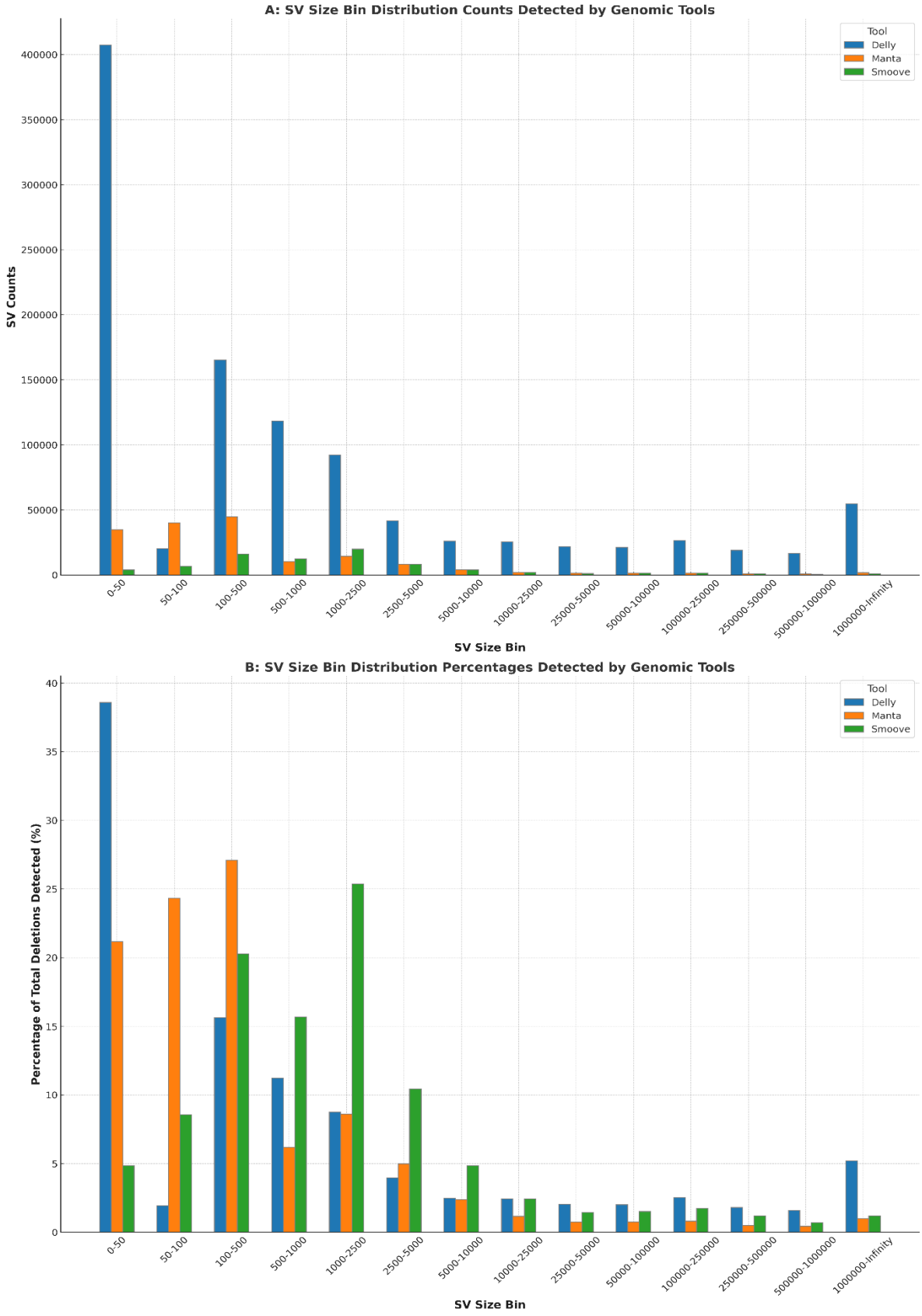


Figure 10. Distribution of structural variant (SV) with respect to their sizes detected *by Delly, Manta and Smoove: Every tool is differentiated by its specific color.*

And if we compare structural variant (SV) detection tools as per their types for example deletion revealed distinct performance characteristics across various SV size bins (Figure 11).

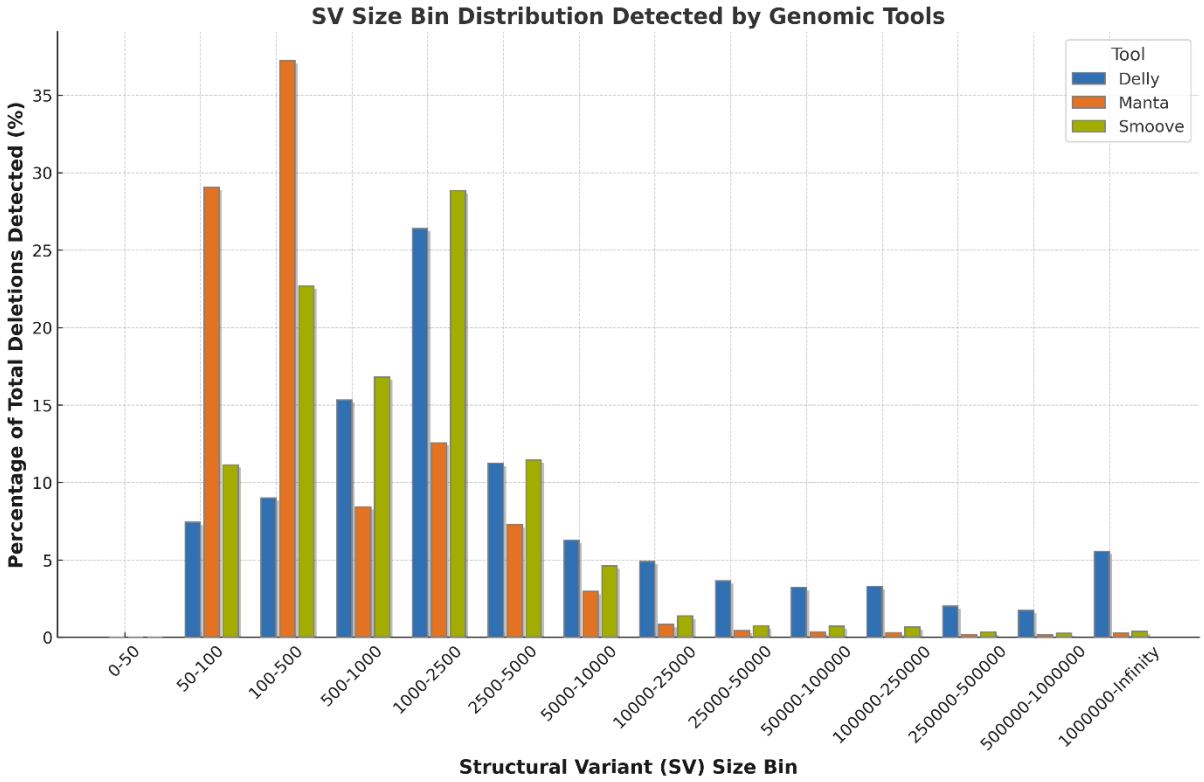


Figure 11. Distribution of structural variant (SV) with respect to their sizes detected by three tools (Deletions): This bar graph illustrates the percentage of total deletions detected across various SV size bins by three genomic tools: Delly, Manta, and Smoove. Each tool's performance is represented by a distinct color (Delly in blue, Manta in orange, and Smoove in green). The SV size bins are plotted along the x-axis, while the y-axis represents the percentage of total deletions detected within each bin.

The script and commands used for extracting deletions from per tool output and making the bins is available in Appendix A, Script A2 and to view the full data (Appendix B, Table 1 & 2).

3.7 Final SURVIVOR merging

The three merged, tool-specific VCF files from DELLY, Manta, and Smoove were integrated using a final SURVIVOR merge step. The same inclusive parameters were utilized with `100 1 1 1 0 50`:

Table 4. Structural variants overlap and unique *counts* from final SURVIVOR merged vcf. SUPP_VEC parameter represents a binary vector indicating the detection of SVs by each tool, where the order is DELLY (first digit), Manta (second digit), and Smoove (third digit).

Parameter (SUPP_VEC)	Description	Number of SVs
100	Detected by DELLY only	879,020
010	Detected by Manta only	90,282
001	Detected by Smoove only	7,156
011	Detected by both Manta and Smoove	3,956

101	Detected by both DELLY and Smoove	18,613
110	Detected by both DELLY and Manta	19,395
111	Detected by DELLY, Manta, and Smoove	47,966
Total		1,103,785

SV Type	**Description**	**Count**
DEL	Deletions	273,703
DUP	Duplications	309,482
INS	Insertions	18,603
INV	Inversions	81,374
TRA	Translocations	420,623
Total		1,103,785

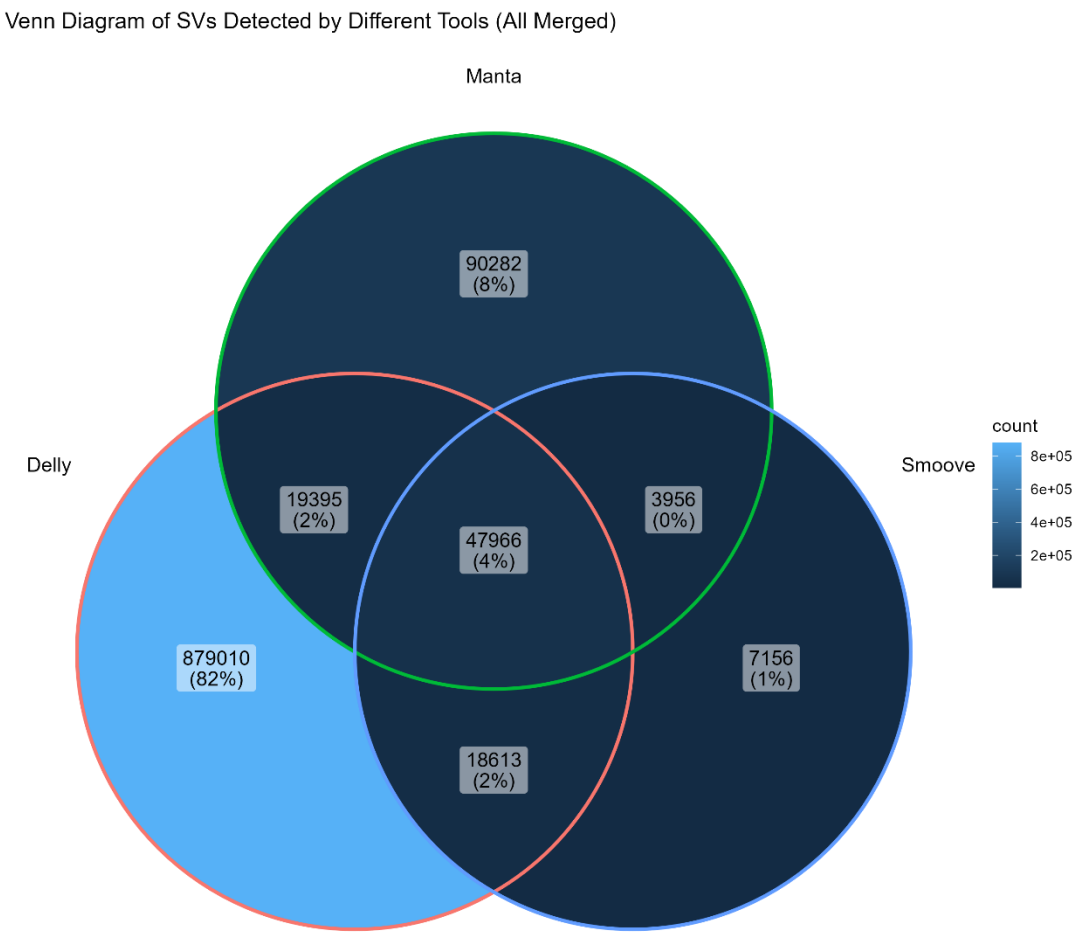


Figure 12. Overlap of detected structural variations between three SV callers: This Venn diagram illustrates the comparative results of structural variation (SV) detection across three tools, each represented by a circle. The numbers within the overlapping sections indicate the count of SVs identified by multiple methods, underscoring the consensus between the techniques. The percentages denote the proportion of shared SVs relative to the total detected by each method. The areas exclusive to each circle reflect the unique SVs identified by a single method, highlighting the method-specific sensitivity and specificity.

Integration of the structural variant (SV) callsets from DELLY, Manta, and Smoove using SURVIVOR merging resulted in 47,966 high-confidence SVs commonly identified by all three tools (Figure 12). Further characterization of these consensus SVs revealed their distribution across SV types per chromosome (Figure 13). The most prevalent SV type was deletions (DEL) at 41,549, followed by duplications (DUP) at 4,623, then inversions (INV) at 1,192 and translocations (TRA) at 602. Assessment of the SVs per chromosome showed higher numbers on the larger chromosomes, with sal 1 having the most at 2,131 total common SVs and sal10 having 2124 SVs. Chromosome 36 had 1,693 common SVs identified by three tools.

The chromosome with the least common SVs was chromosome 37, with only 530. The SV type proportions also differed between chromosomes - smaller chromosomes tended to have relatively higher proportions of INV and TRA compared to the larger chromosomes. These analyses help validate these regions as SV distributed.

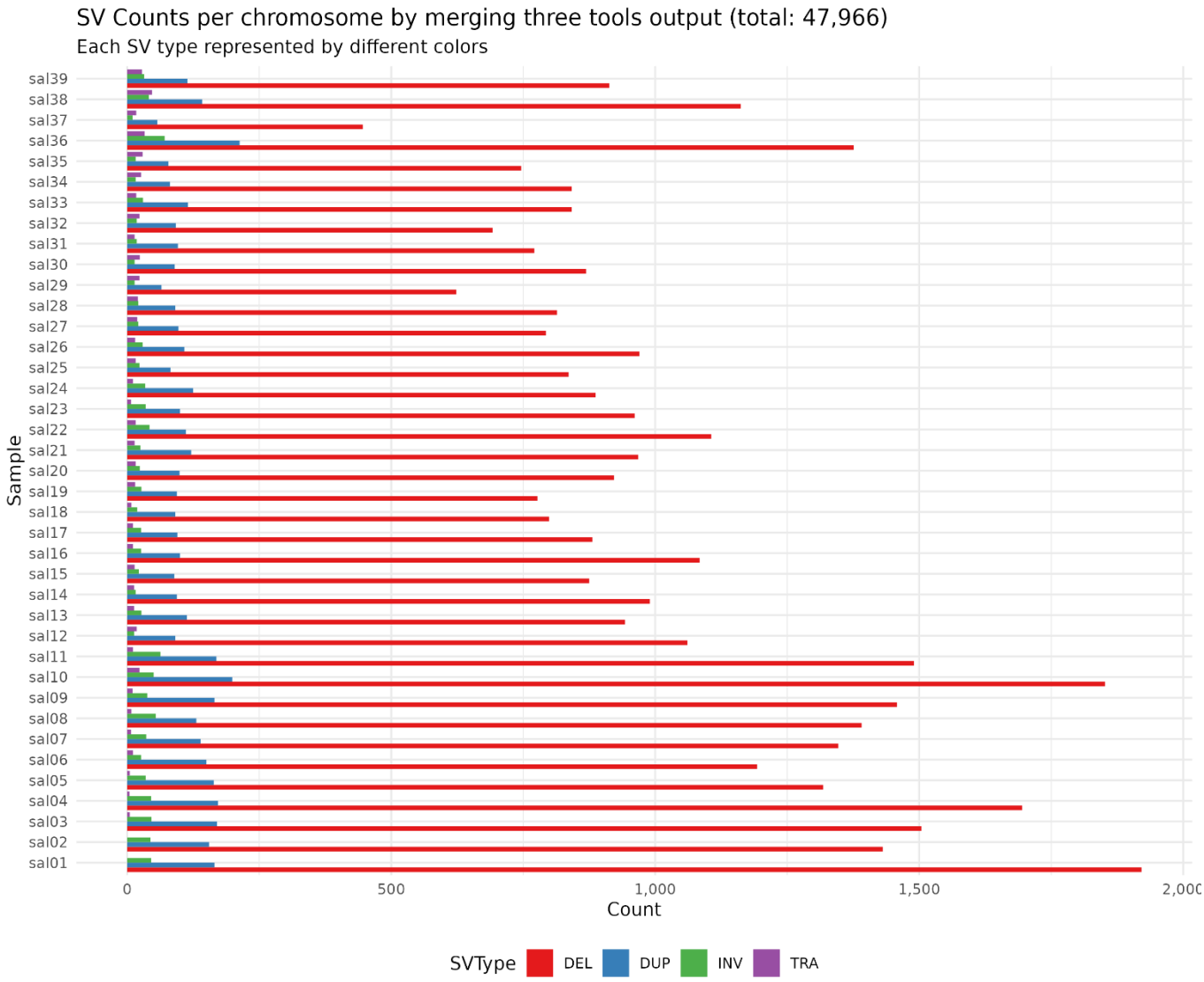


Figure 13. Distribution of structural variations by type at chromosomal level: This horizontal bar chart represents the number of SVs, and y axis labeled from sal01 to sal39 (chromosomes). Each group of horizontal bars corresponds to a chromosome and they are segmented into color-coded sections that represent different types of SVs. The length of each colored segment within a bar denotes the count of SVs of that particular type within the sample. The x-axis quantifies the number of SVs detected, facilitating a direct comparison of the variation load between chromosomes. This visualization provides an at-a-glance comparative analysis of the structural variation burden across chromosomes.

3.8 Breakdown of final merged file by SV types

Deletions

The three structural variation detection tools exhibited significant imbalance. DELLY identified the highest number of unique deletions at 167,338, followed by Manta at 35,541 and a markedly lower Smoove-specific of just 3,156.

However, despite highly uneven tool-specific counts, there was ample consensus where all algorithms corroborated shared variants. 41,549 deletions were unanimously called by DELLY, Manta and Smoove, representing majority concordance for Smoove and Manta, and DELLY (Figure 14).

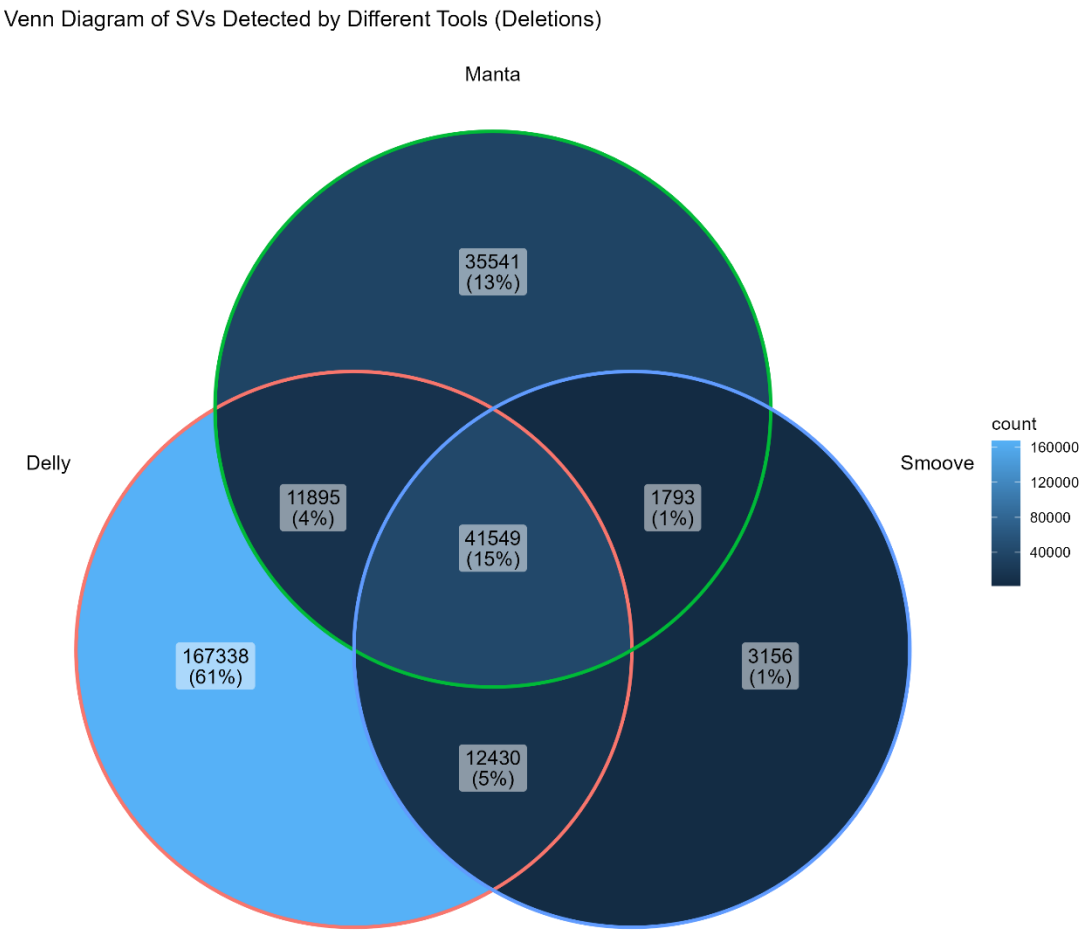


Figure 14. Overlap of detected deletions between three SV callers: This Venn diagram illustrates the comparative results of detected deletions across three tools, each represented by a circle. The numbers within the overlapping sections indicate the count of SVs detected by multiple methods, underscoring the consensus between the techniques. The percentages denote the proportion of shared SVs relative to the total detected by each method. The areas exclusive to each circle reflect the unique SVs identified by a single method, highlighting the method-specific sensitivity and specificity.

Duplication

The final merged VCF contained 309,482 total duplication events by the integrated DELLY, Manta and

Smoove caller approach (Figure 15).

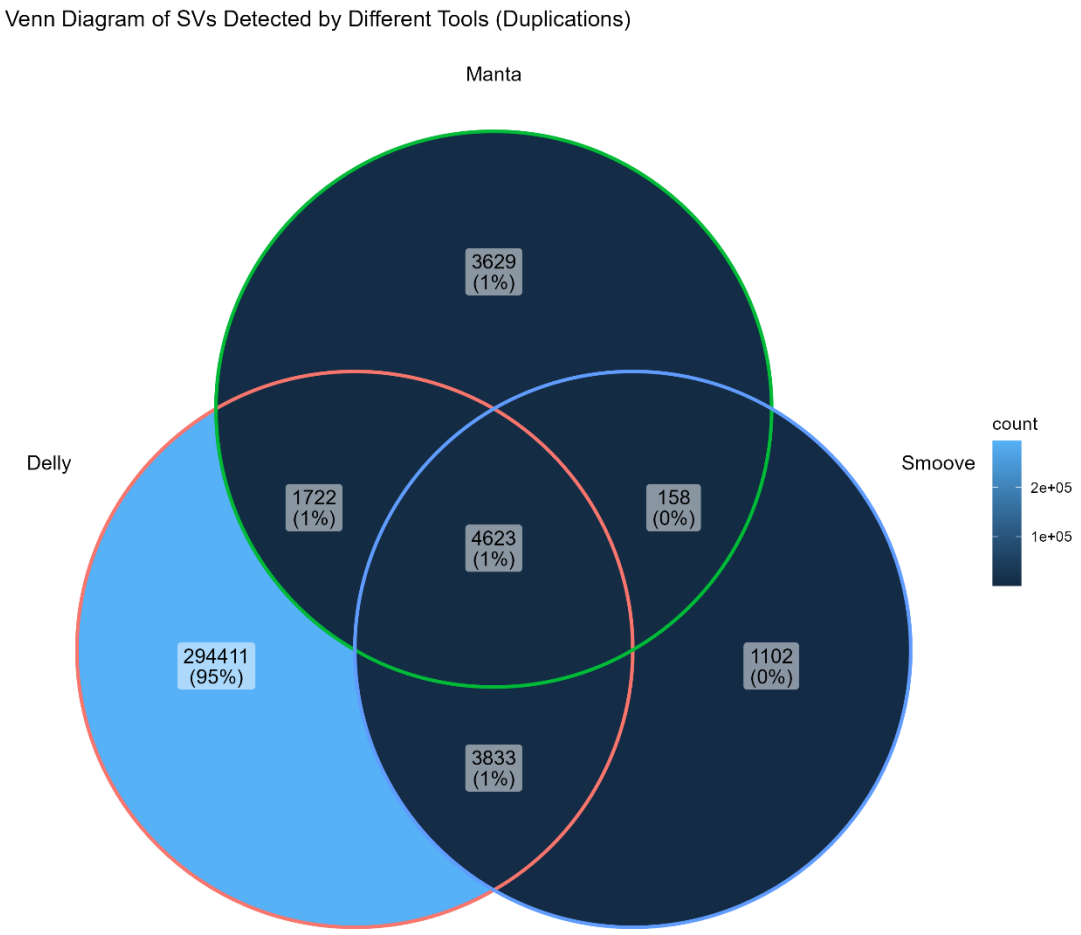


Figure 15. Overlap of detected duplications between three SV callers: *This Venn diagram illustrates the comparative results of detected duplications across three tools (DELLY, Manta and Smoove).*

Insertions

A total of 18,603 genomic insertion events were enumerated across the samples in the comprehensive final merged result. For insertions, little overlap was observed, with Delly and Manta 4, and Smoove do not detect any insertions (Figure 16).

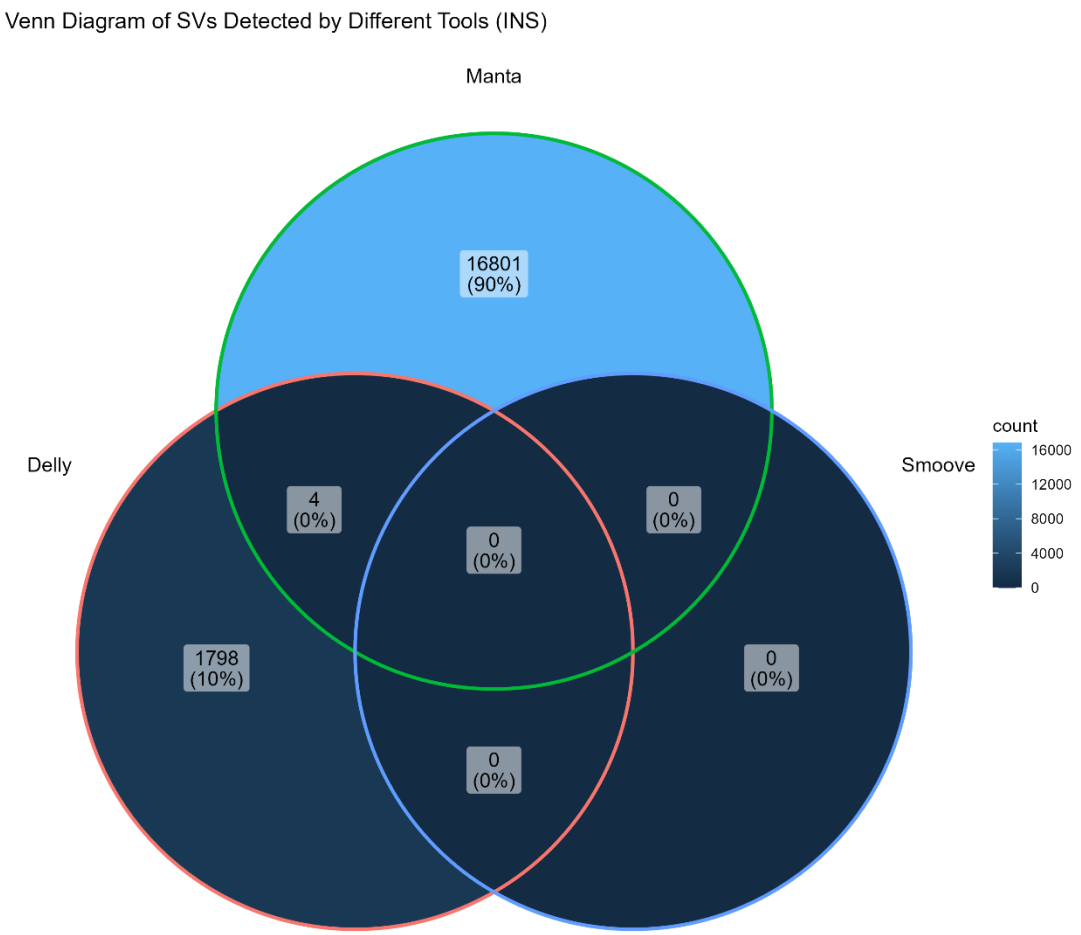


Figure 16. Venn diagram showing overlap and distribution of unique and shared insertion *calls* between DELLY, Manta and Smoove.

Inversions

The merged vcf contained 81,374 total genomic inversion events. It includes 1,192 high-confidence inversions supported by all 3 tools DELLY, Manta and Smoove. DELLY contributed the vast majority of inversion calls, though substantial subsets were discretely identified by Manta and Smoove. A modest fraction displayed unanimous multi-tool support (Figure 17).

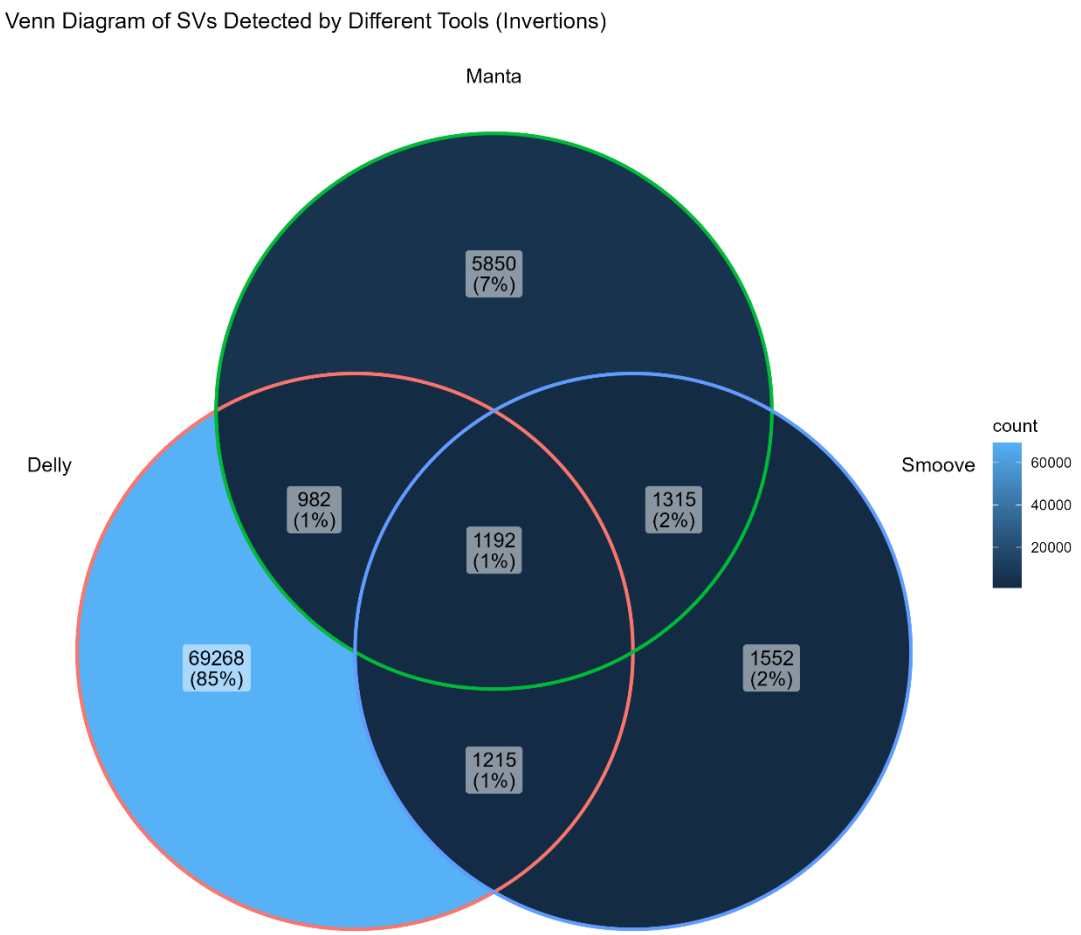


Figure 17. Venn diagram *inversions* showing overlap and distribution of unique and shared inversion calls between DELLY, Manta and Smoove.

Breakends

Lastly, breakends/translocations contained 420,623 BND/Tanslocations.

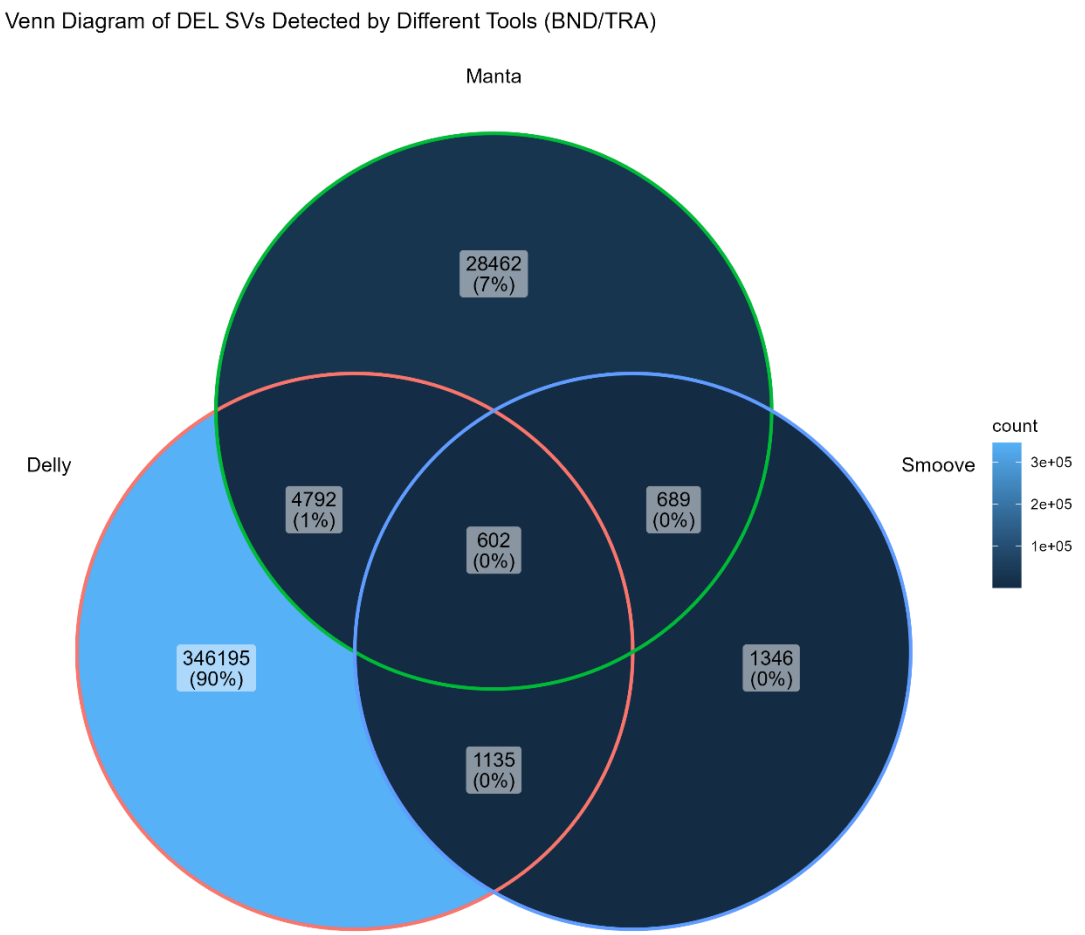


Figure 18. Venn diagram BND/TRA depicting overlap between unique and shared translocation calls ascertained by DELLY, Manta and Smoove.

DELLY contributed the most sample-specific translocation calls (Figure 18). More moderate subsets were discretely identified by Manta and minor contributions by Smoove. A subset of 602 events displayed unanimous multi-tool support.

Chapter 4 Discussion

A holistic overview of samples with mean coverage

4.1 Chromosome 36: A repeat-rich region

In my analysis, chromosome 36 stands out as the chromosome with the most SVs detected. This tiny chromosome represents just 1.75% of the total DNA content. Yet despite this, between 4.9% (Smoove) to 22.9% (Delly) of all detected structural variants mapped specifically to chromosome 36 (Figure 6,8,10). The density of structural variations (SVs) per chromosome length is also higher versus other chromosomes (Appendix E, Figure E1). Notably, chromosome 36 has the highest density of SVs. So, maybe these complex, repeat-rich regions could present multiple targets for erroneous rearrangements via recombination or replication errors. The sample-specific increase in chromosome 36's mean sequence coverage provides further evidence for interspersed structural variations. We therefore can hypothesize the high SV levels in chromosome 36 mainly derive from this chromosome specific structure, probably rich in repeated sequences The repetitive nature of chromosome 36 can potentially promote diverse DNA breakage and rearrangement mechanisms. Despite the overall

hotspot trend.
Although, the absolute SV counts differ substantially between the 3 algorithmic approaches. Yet they converge on the same patterns of chromosome 36 instability. This is also strengthening the hypothesis of higher detected SVs on chromosome 36.

Certain samples (like Samples 25 & 26) show far higher SV density than others (Appendix E). If we see chromosome 36's mean coverage, it is substantially elevated versus the average of ~10X. Across samples, chromosome 36 mean coverage ranges from 15.14 to 22.9. This directly fits expectations of higher interspersed SVs on chromosome 36.

In summary, the coverage signals reinforce the patterns of higher SV density and its abundance seen through independent SV analysis. By integrating these orthogonal measures, we can achieve heightened confidence in deducing the genomic architecture and evolutionary dynamics of tricky regions like chromosome 36. Both the SV and coverage data solidify chromosome 36's identity as an SV-laden hotspot within the arctic char genome.

4.2 Chromosome 17's genomic Architecture

We observed an intriguing pattern where chromosome sal36 shows the highest levels of structural variants across multiple detection algorithms,that could partially be explained by a higher coverage than the average 10x, yet sal17 has even higher sequence coverage estimates (spanning 16.4X-30.6X) based on aligned BAM files. This seemingly contradicts initial hypotheses that increased in coverage providing more substrate for rearrangement events and elevated SV counts.

Ultimately, high sequence coverage does not guarantee corresponding surges in SVs. The architecture of specific repeats and duplications likely plays a key modulating role. While sal17 might be abundant with interspersed duplications that is driving up its coverage, these expanded segments may remain in 'stable' orientations that resist rearrangement. In contrast, the SVs on sal36 could exist in inverted or otherwise 'unstable' orientations - increasing probability of erroneous deletion, inversion, or shifted insertion events during repair or replication (B. Z. Li et al., 2020). The repetitive motifs may also differ substantially in length, complexity, or percent match - further influencing structural variability (Carvalho & Lupski, 2016).

In addition, high-identity repeats pose bioinformatic challenges, with assemblies based on short-read sequencing often collapsing such regions into single contigs (Treangen & Salzberg, 2012; Wang et al., 2021). This can obscure the full extent of underlying SVs detectable from long-read or linked-read technologies. Nevertheless, the multiple algorithms applied here reliably detect elevated SVs on sal36 relative to other regions.

We conclude that chromosome architecture (repeat density/orientation/identity repeats) can modulate more SVs than simple coverage metrics. Further high-resolution characterization of sal17's versus sal36's repetitive motifs will clarify the differential impacts on variant accumulation.

4.3 Delly, Manta and Smoove, VCF filter

The LowQual categorization indicates lower confidence structural variations. Sources of uncertainty include technical artifacts, false signals from repetitive regions, or insufficient supporting evidence meeting callers' statistical filters. In contrast, obtaining a PASS designation marks the validated variant.. However, true events can still be present among the more abundant LowQual class depending on balance of precision versus sensitivity. As the Delly results made up 84% of total SVs in the merged outputs are low quality, the PASS/LowQual ratios for this caller weigh heavily on overall quality designations.

The skew towards LowQual calls reinforces the continued challenges of comprehensively capturing structural variations in complex genomes. Nevertheless, through integration with complementary experiments, even ambiguous initial predictions can lead to discovery of novel variants refining the full mutation spectrum of arctic charr diversity.
The non-PASS calls encompass those failing certain quality or technical filters. As with Delly, sources of uncertainty include artifacts, repetitive regions or statistical thresholds. Nevertheless, obtaining a PASS label indicates the highest-confidence validated variant set per Manta's probabilistic algorithm. Despite no standalone LowQual group, applying orthogonal checks could still help demarcate true versus false signals among the 43% non-PASS fraction.

In comparison to Delly's 19% PASS rate, Manta showed greater relative confidence, with over half its calls achieving PASS status. However, Manta's total output was far lower, likely contributing to stricter thresholds for passing filters. Integrating the PASS/non-PASS designations from both Manta and Delly could help stratify confirmation approaches depending on initial quality grades. Additionally, the lack of extreme imbalance in Manta filters highlights its value for substantiating the abundant LowQual variants uncovered uniquely by Delly's sensitive approach.

4.4 Comparative SV: Merged Structural Variant Analysis

4.4.1 Survivor per tool merging

Our analysis focused on evaluating the performance of three structural variant detection tools, across different SV size bins. The 3 output merged files generated by using the SURVIVOR merge tool with a minimum SV length filter parameter set to 50 bp. Consequently, the 0-50 bp bin primarily captured translocations and other SVs where breakpoint did not allow accurate size determination. Delly demonstrated the greatest sensitivity overall for detecting large number of predicted translocations and SV breakpoints. Its capacity for capturing these imprecise variants suggests an advantage in detection specificity.

Manta showed particular strength in detecting small to mid-sized SVs in the 50-500 bp range. Its robust performance in these categories points to high sensitivity in this size of SV.

Smoove revealed lower overall counts across most size bins, indicating limitations in sensitivity relative to Delly and Manta. However, it provide most detection of SVs in length from 100 to 2500 bp (Appendix B, Table B1).
Moreover, I have selected the main concerning SV type deletion and upon normalization of the number of deletions per bin to their respective total per tool, expressed the data in percentage terms to facilitate a more direct comparison (Appendix B, Table B2). The analysis highlighted that Manta was particularly able at identifying deletions of 200 base pairs (bp) or larger, with its higher SV counts observed in the 50-100 bp (29.06%) and 100-500 bp (37.23%) size bins. Smoove showed a comparable advantage in the mid-size range, specifically from 200 bp to 10,000 bp, with its highest performance in the 1000-2500 bp bin at 28.84%. Delly's performance was more evenly distributed, with a notable proportion of deletions (26.39%) detected in the 1000-2500 bp size bin, suggesting a proficiency in identifying larger SVs.

The findings indicate a stratified efficiency of the tools with respect to SV size, with Manta tending to capture a higher percentage of smaller SVs, Smoove showing balanced detection across a mid-size range, and Delly demonstrating a relative uniformity across a broad SV size spectrum, but with an emphasis on SVs greater than 1000 bp. These insights into the performance distribution across SV size bins underscore the importance of tool selection based on the SV size range of interest in genomic studies.

4.4.2 Final merging

As shown in table 3, The final merged VCF file has 1,103,785 total structural variants (SVs) detected by unifying three VCFs from three softwares. The SUPP_VEC tag was utilized by SURVIVOR to categorize unique and overlapping SVs between the three tools based on the following designations: 100 for DELLY-only calls, 010 for Manta-only, 001 for Smoove-only, 110 for DELLY+Manta, 101 for DELLY+Smoove, 011 for Manta+Smoove, and 111 for events detected by all three tools. As summarized in Table 3, the vast majority of SVs in the final dataset were contributed by DELLY (916,411). Comparatively fewer calls were unique to Manta (90,288) and Smoove (7,156). High-confidence SVs supported by all three callers comprised 47,966 events. Furthermore, in our study, we detected a substantially higher number of high-confidence structural variants (47,966 SVs) in Arctic charr compared to the 15,483 SVs reported by Bertolotti et al. (2020) in Atlantic salmon. This difference raises several key points for discussion from both biological and technical perspectives. The higher SV count in Arctic charr may reflect genuine species-specific genomic differences, given the distinct evolutionary trajectories and adaptations of these two salmonid species. Arctic charr genomes may have greater underlying genomic complexity, mosaicism, and structural dynamism. However, we must also critically evaluate methodological differences including sequencing, reference genomes used, and the SV detection and filtering algorithms employed. Our study have used less stringent filtering to have a broader view of SVs, leading to a higher but less accurate SV count. Rigorous false positive filtering by Bertolotti et al. (2020) likely improved precision and lower their count of high confident SVs.

The final dataset contains 273,703 deletions, 309,482 duplications and a range of other structural variation types.

A relatively small but significant fraction of 4.6k duplications demonstrated unanimous support. As with deletions, the distribution spotlights complementarity between the callers in detecting duplications for a holistic overview.
while DELLY and Smoove contributed minor tool-specific insertion subsets, Manta overwhelmingly dominated insertion calls in the ensemble result. This underscores the specialized capacity of Manta for sensitively detecting this subclass of structural variation.
Before merging through SURVIVOR, Manta outputs from the 30 Samples showed no identifiable inversions, instead finding deletions, duplications, insertions and breakends. Contrastingly, DELLY and Smoove both reported inversion subsets in their discrete VCF callsets.

However, the final merged SV dataset contained 81,374 total inversion events - including 69k DELLY-specific, 5.8k Manta-specific, 982 calls supported by both tools and 1.3k calls supported by Manta and Smoove. This implies that Survivor enabled interconversion of complex breakpoint-associated events initially classified differently. For instance, a complex DELLY event may have been re-designated as an inversion after intersecting and merging with another tool's callset. Concordant multi-tool support also conferred higher validation confidence for subclass reassignment.

While the standalone Manta outputs lacked clear inversions, integration with DELLY and Smoove events via SURVIVOR merging permitted subclass relabeling - eliciting a more complete view of the inversion landscape.
Before merging VCF files by SURVIVOR, their preliminary analyses revealed that Manta outputs lacked explicitly defined inversion events, instead finding deletions, duplications, insertions and breakends. Meanwhile, DELLY and Smoove outputs contained recognizable inversion subgroups.

Interestingly, the final integrated SV dataset obtained via merging using SURVIVOR contained over 81k labeled inversions including major DELLY and Manta contributions. This implies SURVIVOR enabled redesignation of complex variants by intersecting the tools' calls. For instance, a Manta breakend intersecting a DELLY deletion may have been interpreted as an inversion after merging. Stringent criteria would require multi-tool corroboration to enable subclass conversion.

A similar flow was observed for translocations. While the initial DELLY, Manta and Smoove VCFs lacked explicitly defined translocations rather they were having BND, SURVIVOR merging elicited over 420k translocation calls - predominantly consisting of DELLY and Manta variants likely reclassified from other Rearrangement types.

In summary, SURVIVOR merging empowered variant subclass conversion, eliciting a more complete overview of inversions and translocations by re-interpreting tools' existing call data.

Finally, In terms of tools' performance, my findings were aligned with the broader trends noted in recent studies. Delly's robust and broad performance across various SV types is consistent with its application in diverse genomic studies. Delly uses paired-end (DP) and split-reads (SR) in a stepwise manner to detect SVs, which is useful for both germline and somatic SV detection, as highlighted in a study by van Belzen et al. (2021). This aligns well with my observation of Delly's sensitivity in detecting large numbers of translocations and SV breakpoints.

Manta which called SVs on paired-end base, its effectiveness in detecting small to mid-sized SVs aligns with its design, which is optimized for precise SV calling. Manta consistently shows strong performance in studies that evaluate multiple SV detection tools, including its ability to work effectively in an integrated approach with other tools (Coutelier et al., 2022). While Delly demonstrates considerable detection capabilities across diverse structural variant (SV) types in Arctic charr, it may lack sufficient sensitivity for smaller SVs that fall below its optimal detection range. On the other hand, Manta's optimization specifically for precise calling of small- to mid-sized SVs could provide depth in cataloging smaller events to complement Delly's range. Integrating these two algorithms by capitalizing on their complementary strengths - Delly providing detection breadth in larger SVs and Manta supplying precision in smaller variants could enable more comprehensive and detailed construction of an Arctic charr SV catalog across a wide size spectrum. Additionally, Manta's efficacy in our SV size range of interest could help reveal crucial genetic markers for breeding programs in aquaculture. Therefore, a multi-tool approach harnessing the combined detection capabilities of Manta for smaller SVs and Delly for larger variants promises the most complete overview of genomic structural variations in this species, serving both the key objectives outlined in this thesis effectively.

Smoove, although exhibiting lower overall counts, was found to be effective in a specific SV size range. This is in line with the literature suggesting that no single algorithm can call every type of SV with high precision and recall. This necessitates using multiple algorithms for full-spectrum SV detection, as suggested by (Kosugi et al., 2019).
Furthermore, if restricted to selecting one structural variant (SV) detection tool, Manta appears the most suitable match given its precision and sensitivity in SV size range of interest. Its efficacy in calling small to mid-sized SVs could significantly aid efforts to construct a detailed and comprehensive catalog of SVs in Arctic charr - a key goal of this thesis. Additionally, Manta's performance in detecting SVs within our targeted range can be invaluable in revealing specific genetic markers for selective breeding programs in Arctic charr aquaculture. However, while Manta excels in these aspects, relying on any single algorithm has inherent limitations in detecting SVs across the full spectrum of size and type distributions. As consistently advocated in genomic literature(Cameron et al., 2019; Gong et al., 2021; van Belzen et al., 2021), integrated approaches utilizing multiple complementary tools provide the

most complete and accurate SV detection results. Therefore, this thesis strongly recommends implementing an ensemble strategy combining Manta's precision in small- to mid-size SVs, Delly's detection strengths in larger structural variants, and other specialized tools. This multi-algorithm approach will serve the objectives of developing a comprehensive Arctic charr SV catalog and identifying genetic markers for aquaculture breeding most effectively - thus advancing the overarching goals of this research substantially.

4.5 Conclusion

This study presents a comprehensive evaluation of structural variation in aquaculture strains of Arctic charr, a cold-water salmonid fish of emerging economic importance. Integrated genome-wide analysis was performed using multiple specialized algorithms Delly, Manta and Smoove to show patterns of genomic rearrangements.

The results reveal key insights into factors influencing accuracy of SV detection from short-read sequences. Despite overall moderate consensus, most calls remained discordant (false positives) across approaches. This finding reiterates the ubiquity of tradeoffs between sensitivity and precision in variant discovery. Applying orthogonal confirmation and combining complementary methods would likely improve performance.

Nevertheless, intersecting callsets probably captures validated variants for advancing genomic selection. The aggregated outputs pointed to exceptional samples and chromosomes, indicating instability arising under aquaculture conditions. Exploring SV allele distributions, family segregation, and trait correlations promises to reveal its significance.

Overall, this work enhances knowledge of structural variation spectra differentiating Arctic charr strains and benchmarking SV tools. The resources and guidelines will broadly empower SV analysis in other non-model species. Furthermore, responsibly incorporating genomic insights could augment selective breeding programs.

Altogether, the multifaceted genomic investigation provides a springboard for elucidating rearrangement mechanisms shaping phenotypic diversity across taxa. Continued technological and analytical advances will undoubtedly further unravel the evolutionary importance of structural variations.

References

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., Buchanan, J., Chappell, J., Daniels, W., Dong, S., Dunham, R., Durland, E., Elaswad, A., Gomez-Chiarri, M., Gosh, K., Guo, X., … Zhou, T. (2017). Aquaculture genomics, genetics and breeding in the United States: Current status, challenges, and priorities for future research. *BMC Genomics*, *18*(1), 1–23. https://doi.org/10.1186/s12864-017-3557-1

Adams, C. E., Fraser, D., Huntingford, F. A., Greer, R. B., Askew, C. M., & Walker, A. F. (1998). Trophic polymorphism among Arctic charr from Loch Rannoch, Scotland. *Journal of Fish Biology*, *52*(6), 1259–1271. https://doi.org/10.1006/jfbi.1998.0676

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

Allal, F., & Nguyen Hong Nguyen. (2022). Genomic Selection in Aquaculture Species. In J. Ahmadi, N., Bartholomé (Ed.), *Genomic Prediction of Complex Traits* (pp. 469–491). Humana Press Inc.

AquaIMPACT. (2023). *Implementing Genomic Selection in Aquaculture Industry: Reducing Genotyping Costs by Low-Density Genotyping Panels*. https://projects.luke.fi/aquaimpact/2023/05/25/implementing-genomic-selection-in-aquaculture-industry-reducing-genotyping-costs-by-low-density-genotyping-panels/

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Research*, *46*(5), 2159–2168. https://doi.org/10.1093/nar/gky066

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. M. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019

Bachtrog, D., & Charlesworth, B. (2022). Towards a complete sequence of the human Y chromosome. *Genome Biology*, *2*(5), 1–47.

Balachandran, P., & Beck, C. R. (2020). Structural variant identification and characterization. *Chromosome Research*, 31–47. https://doi.org/10.1007/s10577-019-09623-z

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., … Macqueen, D. J. (2020). The structural variation landscape in 492 Atlantic salmon genomes. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18972-x

Bourret, V., Kent, M. P., Primmer, C. R., Vasemägi, A., Karlsson, S., Hindar, K., McGinnity, P., Verspoor, E., Bernatchez, L., & Lien, S. (2013). SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology*, *22*(3), 532–551. https://doi.org/10.1111/mec.12003

Brunner, P. C., Douglas, M. R., Osinov, A., Wilson, C. C., & Bernatchez, L. (2001). Holarctic phylogeography of arctic charr (Salvelinus alpinus L.) Inferred from mitochondrial DNA sequences. *Evolution*, *55*(3), 573–586. https://doi.org/10.1111/j.0014-3820.2001.tb00790.x

Burssed, B., Zamariolli, M., Bellucco, F. T., & Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, *15*(1), 1–15. https://doi.org/10.1186/s13039-022-00600-6

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), 1–11. https://doi.org/10.1038/s41467-019-11146-4

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

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

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1–16. https://doi.org/10.1038/s41467-018-08148-z

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19). https://doi.org/10.1093/nar/gkw654

Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, *18*(8), 495–506. https://doi.org/10.1038/nrm.2017.48

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)*, *32*(8), 1220–1222. https://doi.org/10.1093/BIOINFORMATICS/BTV710

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., Ferguson, M. M., Stadnik, A., Devlin, R. H., Muzzerall, R., Edwards, M., Davidson, W. S., & Koop, B. F. (2021). Retraction: The Arctic charr (Salvelinus alpinus) genome and transcriptome assembly. *PLoS ONE*, *16*(2 February), 247083. https://doi.org/10.1371/journal.pone.0247083

Correa, K., Lhorente, J. P., López, M. E., Bassini, L., Naswa, S., Deeb, N., Di Genova, A., Maass, A., Davidson, W. S., & Yáñez, J. M. (2015). Genome-wide association analysis reveals loci associated with resistance against Piscirickettsia salmonis in two Atlantic salmon (Salmo salar L.) chromosomes. *BMC Genomics*, *16*(1), 1–9. https://doi.org/10.1186/s12864-015-2038-7

Coutelier, M., Holtgrewe, M., Jäger, M., Flöttman, R., Mensah, M. A., Spielmann, M., Krawitz, P., Horn, D., Beule, D., & Mundlos, S. (2022). Combining callers improves the detection of copy number variants from whole-genome sequencing. *European Journal of Human Genetics*, *30*(2), 178–186. https://doi.org/10.1038/s41431-021-00983-x

Cretu Stancu, M., Van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., De Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., Korzelius, J., De Bruijn, E., Cuppen, E., Talkowski, M. E., Marschall, T., De Ridder, J., & Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*, *8*(1), 1–13. https://doi.org/10.1038/s41467-017-01343-4

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

Diblasi, C., Barson, N., & Marie, S. (2023). Resolving large-scale genome evolution in the high-throughput sequencing era: structural variants, genome rearrangement, and karyotype dynamics in animals. *EcoEvoRxiv*, *0*.

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Fan, X., Chaisson, M., Nakhleh, L., & Chen, K. (2017). HySA: A hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. *Genome Research*, *27*(5), 793–800. https://doi.org/10.1101/gr.214767.116

FAO. (1995). Selective breeding programmes for medium-sized fish farms. In *FAO Fisheries Technical Paper* (Vol. 352).

Ferguson, M. M., Danzmann, R. G., & Hutchings, J. A. (1991). Incongruent estimates of population differentiation among brook charr, Salvelinus fontinalis, from Cape Race, Newfoundland, Canada, based upon allozyme and mitochondrial DNA variation. *Journal of Fish Biology*, *39*, 79–85. https://doi.org/10.1111/j.1095-8649.1991.tb05070.x

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

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

Gjedrem, T., Robinson, N., & Rye, M. (2012). The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. *Aquaculture*, *350–353*, 117–129. https://doi.org/10.1016/j.aquaculture.2012.04.008

Gong, T., Hayes, V. M., & Chan, E. K. F. (2021). Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics*, *22*(3), 1–15. https://doi.org/10.1093/bib/bbaa056

Gonzalez-Pena, D., Gao, G., Baranski, M., Moen, T., Cleveland, B. M., Brett Kenney, P., Vallejo, R. L., Palti, Y., & Leeds, T. D. (2016). Genome-wide association study for identifying loci that affect fillet yield, carcass, and body weight traits in rainbow trout (Oncorhynchus mykiss). *Frontiers in Genetics*, *7*(NOV). https://doi.org/10.3389/fgene.2016.00203

Gutierrez, A. P., Turner, F., Gharbi, K., Talbot, R., Lowe, N. R., Peñaloza, C., McCullough, M., Prodöhl, P. A., Bean, T. P., & Houston, R. D. (2017). Development of a medium density combined-species SNP array for pacific and european oysters (Crassostrea gigas and Ostrea edulis). *G3: Genes, Genomes, Genetics*, *7*(7), 2209–2218. https://doi.org/10.1534/g3.117.041780

Helgadóttir, G., Renssen, H., Olk, T. R., Oredalen, T. J., Haraldsdóttir, L., Skúlason, S., & Thorarensen, H. Þ. (2021). Wild and Farmed Arctic Charr as a Tourism Product in an Era of Climate Change. *Frontiers in Sustainable Food Systems*, *5*(August), 1–10. https://doi.org/10.3389/fsufs.2021.654117

Hollox, E. J., Zuccherato, L. W., & Tucci, S. (2022). Genome structural variation in human evolution. *Trends in Genetics*, *38*(1), 45–58. https://doi.org/10.1016/j.tig.2021.06.015

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., & Robledo, D. (2020). Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics*, *21*(7), 389–409. https://doi.org/10.1038/s41576-020-0227-y

Houston, R. D., Taggart, J. B., Cézard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S. C., Archibald, A. L., Bron, J. E., Penman, D. J., Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics*, *15*(1), 1–13. https://doi.org/10.1186/1471-2164-15-90

Huddleston, J., & Eichler, E. E. (2016). An incomplete understanding of human genetic variation. *Genetics*, *202*(4), 1251–1254. https://doi.org/10.1534/genetics.115.180539

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345.

https://doi.org/10.1038/nbt.4060

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(May 2016), 1–11. https://doi.org/10.1038/ncomms14061

Jiang, T., Liu, S., Cao, S., & Wang, Y. (2022). Structural Variant Detection from Long-Read Sequencing Data with cuteSV. *Methods in Molecular Biology*, *2493*, 137–151. https://doi.org/10.1007/978-1-0716-2293-3_9/COVER

Jonsson, B., & Jonsson, N. (2001). Polymorphism and speciation in Arctic charr. *Journal of Fish Biology*, *58*(3), 605–638. https://doi.org/10.1006/jfbi.2000.1515

Kapralova, K. H., Morrissey, M. B., Kristjánsson, B. K., Lafsdóttir, G. Á., Snorrason, S. S., & Ferguson, M. M. (2011). Evolution of adaptive diversity and genetic connectivity in Arctic charr (Salvelinus alpinus) in Iceland. *Heredity*, *106*(3), 472–487. https://doi.org/10.1038/hdy.2010.161

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

Klemetsen, A. (2010). The Charr Problem Revisited: Exceptional Phenotypic Plasticity Promotes Ecological Speciation in Postglacial Lakes. *Freshwater Reviews*, *3*(1), 49–74. https://doi.org/10.1608/frj-3.1.3

Klemetsen, A., Amundsen, P. A., Dempson, J. B., Jonnson, B., Jonnson, 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*, 1–59.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, *20*(1), 8–11. https://doi.org/10.1186/s13059-019-1720-5

Koufariotis, L., Hayes, B. J., Kelly, M., Burns, B. M., Lyons, R., Stothard, P., Chamberlain, A. J., & Moore, S. (2018). Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-35698-5

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS ONE*, *12*(5), 1–20. https://doi.org/10.1371/journal.pone.0177459

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C. J., Wang, C., Zamani, N., Grant, B. R., Grant, P. R., Webster, M. T., & Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), 371–375. https://doi.org/10.1038/nature14181

Leung, H. C. M., Yu, H., Zhang, Y., Leung, W. S., Lo, I. F. M., Luk, H. M., Law, W. C., Ma, K. K., Wong, C. L., Wong, Y. S., Luo, R., & Lam, T. W. (2022). Detecting structural variations with precise breakpoints using low-depth WGS data from a single oxford nanopore MinION flowcell. *Scientific Reports*, *12*(1), 1–8. https://doi.org/10.1038/s41598-022-08576-4

Li, B. Z., Putnam, C. D., & Kolodner, R. D. (2020). Mechanisms underlying genome instability mediated by formation of foldback inversions in saccharomyces cerevisiae. *ELife*, *9*, 1–122. https://doi.org/10.7554/ELIFE.58223

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Liu, S., Vallejo, R. L., Palti, Y., Gao, G., Marancik, D. P., Hernandez, A. G., & Wiens, G. D. (2015). Identification of single nucleotide polymorphism markers associated with bacterial cold water disease resistance and spleen size in rainbow trout. *Frontiers in Genetics*, *6*(SEP), 1–10. https://doi.org/10.3389/fgene.2015.00298

Liu, Z. J., & Cordes, J. F. (2004). Erratum: DNA marker technologies and their applications in aquaculture genetics (Aquaculture (2004) 238 (1-37) PII: S0044-8486(04)00285-6 and DOI: 10.1016/j.aquaculture.2004.05.027. *Aquaculture*, *242*(1–4), 735–736. https://doi.org/10.1016/j.aquaculture.2004.08.022

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., Jiang, C., Sun, L., Wang, R., Zhang, Y., Zhou, T., Zeng, Q., Fu, Q., Gao, S., Li, N., Koren, S., Jiang, Y., Zimin, A., Xu, P., … Waldbieser, G. C. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms11757

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

Mani, R. S., & Chinnaiyan, A. M. (2010). Triggers for genomic rearrangements: Insights into genomic, cellular and environmental influences. *Nature Reviews Genetics*, *11*(12), 819–829. https://doi.org/10.1038/nrg2883

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends in Ecology and Evolution*, *35*(7), 561–572. https://doi.org/10.1016/j.tree.2020.03.002

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Moccetti, P., Siwertsson, A., Kjær, R., Amundsen, P. A., Præbel, K., Tamayo, A. M. P., Power, M., & Knudsen, R. (2019). Contrasting patterns in trophic niche evolution of polymorphic Arctic charr populations in two subarctic Norwegian lakes. *Hydrobiologia*, *840*(1), 281–299. https://doi.org/10.1007/s10750-019-3969-9

Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J., & Schmitz, U. (2023). The application of long-read sequencing in clinical settings. *Human Genomics*, *17*(1), 1–13. https://doi.org/10.1186/s40246-023-00522-3

Palti, Y., Gao, G., Liu, S., Kent, M. P., Lien, S., Miller, M. R., Rexroad, C. E., & Moen, T. (2015). The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources*, *15*(3), 662–672. https://doi.org/10.1111/1755-0998.12337

Pappas, F., Kurta, K., Vanhala, T., Jeuthe, H., Hagen, Ø., Beirão, J., & Palaiokostas, C. (2023). Whole-genome re-sequencing provides key genomic insights in farmed Arctic charr (Salvelinus alpinus) populations of anadromous and landlocked origin from Scandinavia. *Evolutionary Applications*, *16*(4), 797–813. https://doi.org/10.1111/eva.13537

Pappas, F., & Palaiokostas, C. (2021). Genotyping strategies using ddrad sequencing in farmed arctic charr (Salvelinus alpinus). *Animals*, *11*(3), 1–14. https://doi.org/10.3390/ani11030899

Pedersen, B. S. (2020). *Smoove: structural-variant calling and genotyping with existing tools*. https://github.com/brentp/smoove

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

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 (Oxford, England)*, *28*(18). https://doi.org/10.1093/BIOINFORMATICS/BTS378

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*(8), 1450–1477. https://doi.org/10.1111/jeb.13047

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Sae-Lim, P., Gjerde, B., Nielsen, H. M., Mulder, H., & Kause, A. (2016). A review of genotype-by-environment interaction and micro-environmental sensitivity in aquaculture species. *Reviews in Aquaculture*, *8*(4), 369–393. https://doi.org/10.1111/raq.12098

Sakamoto, Y., Zaha, S., Suzuki, Y., Seki, M., & Suzuki, A. (2021). Application of long-read sequencing to the detection of structural variants in human cancer genomes. *Computational and Structural Biotechnology Journal*, *19*, 4207–4216. https://doi.org/10.1016/J.CSBJ.2021.07.030

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). 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

Steensma, M. J., Lee, Y. L., Bouwman, A. C., Pita Barros, C., Derks, M. F. L., Bink, M. C. A. M., Harlizius, B., Huisman, A. E., Crooijmans, R. P. M. A., Groenen, M. A. M., Mulder, H. A., & Rochus, C. M. (2023). Identification and characterisation of de novo germline structural variants in two commercial pig lines using trio-based whole genome sequencing. *BMC Genomics*, *24*(1), 1–11. https://doi.org/10.1186/s12864-023-09296-3

Stenløkk, K. S. R. (2023). *Genomic structural variations as drivers of adaptation in salmonid fishes*.

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R. E., Persengiev, S., Antonacci, F., Ventura, M., Prado-Martinez, J., Project, G. A. G., Marques-Bonet, T., & Eichler, E. E. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, *23*(9), 1373–1382. https://doi.org/10.1101/gr.158543.113

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J. S., Soudi, S., Burge, D. O., Huang, K., Ostevik, K. L.,

Drummond, E. B. M., Imerovski, I., Lande, K., Pascual-Robles, M. A., Nanavati, M., Jahani, M., Cheung, W., Staton, S. E., Muños, S., Nielsen, R., Donovan, L. A., … Rieseberg, L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, *584*(7822), 602–607. https://doi.org/10.1038/s41586-020-2467-6

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36–46. https://doi.org/10.1038/nrg3117

Tsai, H. Y., Hamilton, A., Guy, D. R., Tinch, A. E., Bishop, S. C., & Houston, R. D. (2015). The genetic architecture of growth and fillet traits in farmed Atlantic salmon (Salmo salar). *BMC Genetics*, *16*(1). https://doi.org/10.1186/s12863-015-0215-y

Vallejo, R. L., Leeds, T. D., Gao, G., Parsons, J. E., Martin, K. E., Evenhuis, J. P., Fragomeni, B. O., Wiens, G. D., & Palti, Y. (2017). Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. *Genetics Selection Evolution*, *49*(1), 1–13. https://doi.org/10.1186/s12711-017-0293-6

van Belzen, I. A. E. M., Schönhuth, A., Kemmeren, P., & Hehir-Kwa, J. Y. (2021). Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *Npj Precision Oncology*, *5*(1), 1–11. https://doi.org/10.1038/s41698-021-00155-6

Wang, P., Meng, F., Moore, B. M., & Shiu, S. H. (2021). Impact of short-read sequencing on the misassembly of a plant genome. *BMC Genomics*, *22*(1), 1–18. https://doi.org/10.1186/s12864-021-07397-5

Wellenreuther, M., & Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology and Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Yáñez, J. M., Barría, A., López, M. E., Moen, T., Garcia, B. F., Yoshida, G. M., & Xu, P. (2023). Genome-wide association and genomic selection in aquaculture. *Reviews in Aquaculture*, *15*(2), 645–675. https://doi.org/10.1111/raq.12750

Yang, L. (2020). A Practical Guide for Structural Variation Detection in the Human Genome. *Current Protocols in Human Genetics*, *107*(1), 1–17. https://doi.org/10.1002/cphg.103

Yu, Y., Zhang, X., Yuan, J., Li, F., Chen, X., Zhao, Y., Huang, L., Zheng, H., & Xiang, J. (2015). Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp Litopenaeus vannamei. *Scientific Reports*, *5*(October), 1–14. https://doi.org/10.1038/srep15612

Zenger, K. R., Khatkar, M. S., Jones, D. B., Khalilisamani, N., Jerry, D. R., & Raadsma, H. W. (2019). Genomic selection in aquaculture: Application, limitations and opportunities with special reference to marine shrimp and pearl oysters. *Frontiers in Genetics*, *10*(JAN). https://doi.org/10.3389/fgene.2018.00693

Zhang, J., Nie, C., Li, X., Zhao, X., Jia, Y., Han, J., Chen, Y., Wang, L., Lv, X., Yang, W., Li, K., Zhang, J., Ning, Z., Bao, H., Zhao, C., Li, J., & Qu, L. (2022). Comprehensive analysis of structural variants in chickens using PacBio sequencing. *Frontiers in Genetics*, *13*(October), 1–11. https://doi.org/10.3389/fgene.2022.971588

Zong, W., Wang, J., Zhao, R., Niu, N., Su, Y., Hu, Z., Liu, X., Hou, X., Wang, L., Wang, L., & Zhang, L. (2023). Associations of genome-wide structural variations with phenotypic differences in cross-bred Eurasian pigs. *Journal of Animal Science and Biotechnology*, *14*(1), 1–20. https://doi.org/10.1186/s40104-023-00929-x

Appendices

Appendix A: Some commands and scripts used in Methods and Materials

Script A.1: bcftools view Command for SV Length Filtering. *This script is utilized for filtering structural variants in VCF files, retaining only those variants with lengths of 50 base pairs or greater, thus ensuring a focused analysis of significant SVs in the study.*

```
bcftools view -i 'SVLEN>=50' input.vcf -o filtered_output.vcf
```

Script A.2: SV Length Distribution Analysis in VCF Files. This script is designed to analyze the length distribution of deletion structural variants (SVs) in a VCF file. It utilizes AWK for parsing and calculating SV counts within specified bin ranges. The script efficiently segments SV lengths into bins and outputs a CSV file summarizing the count of deletions per bin, facilitating a detailed understanding of SV length distribution in the dataset.
Initialization: Setting file paths and declaring bin sizes.

```
vcf_path="/path/to/merged_smoove.vcf"

output_file="output.csv"

# Define bin sizes

declare -a bins=(50 100 500 1000 2500 5000 10000 25000 50000 100000
250000 500000 1000000)

# Prepare the output file

echo "Bin,Count" > "$output_file"

# Function to process each bin

process_bin() {
    local start=$1
    local end=$2
    local count
  # Count deletions (DEL) in the bin range using awk
    count=$(awk -v start=$start -v end=$end \
        'BEGIN{FS="\t"; count=0}
        $1 !~ /^#/ && $8 ~ /SVTYPE=DEL/ && $8 ~ /SVLEN=/ {
            match($8, /SVLEN=-?([0-9]+)/, arr);
            svlen=arr[1];
            if (svlen >= start && svlen < end) count++
        }
```

```
        END{print count}' "$vcf_path")
  # Write to output file
    echo "${start}-${end},$count" >> "$output_file"
}
# Process each bin
prev_bin=0
for bin in "${bins[@]}"; do
    process_bin $prev_bin $bin
    prev_bin=$bin
done
# Handle SVs larger than the last bin
process_bin $prev_bin "Infinity"
# Display the output
cat "$output_file"
```

Appendix B: Extra Tables

Table B1. This table shows the total number of SVs detected and distributed per bin and their proportions

#	Bin Sizes	Count (Delly)	Count (Manta)	Count (Smoove)	Percentage (Delly)	Percentage (Manta)	Percentage (Smoove)
1	0-50	407,175	34,852	3,793	38.59%	21.17%	4.85%
2	50-100	20,146	40,038	6,689	1.91%	24.32%	8.55%
3	100-500	164,981	44,583	15,852	15.64%	27.08%	20.25%
4	500-1000	118,220	10,160	12,269	11.21%	6.17%	15.68%
5	1000-2500	92,122	14,162	19,835	8.73%	8.60%	25.35%
6	2500-5000	41,516	8,174	8,155	3.94%	4.97%	10.42%
7	5000-10000	26,054	3,910	3,778	2.47%	2.37%	4.83%
8	10000-25000	25,576	1,887	1,886	2.42%	1.15%	2.41%

9	25000-50000	21,491	1,178	1,116	2.04%	0.72%	1.43%
10	50000-100000	21,003	1,238	1,174	1.99%	0.75%	1.50%
11	100000-250000	26,352	1,341	1,351	2.50%	0.81%	1.73%
12	250000-500000	18,877	778	917	1.79%	0.47%	1.17%
13	500000-1000000	16,610	696	540	1.57%	0.42%	0.69%
14	1000000-Infinity	54,755	1,611	913	5.19%	0.98%	1.17%
	Total	1,054,878	164,608	78,268			

Table B2. This table presents the data of distribution and proportion of Structural Variants (SVs) across each bin but only for deletions

#	Bin sizes	Count (delly)	Count (manta)	Count (smoove)	Percentage (delly)	Percentage (manta)	Percentage (smoove)
0	0-50	0	0	0	0.00	0.00	0.00
1	50-100	18210	26853	6599	7.42	29.06	11.14
2	100-500	22080	34409	13434	9.00	37.23	22.68
3	500-1000	37599	7792	9952	15.32	8.43	16.80
4	1000-2500	64743	11598	17088	26.39	12.55	28.84
5	2500-5000	27553	6722	6785	11.23	7.27	11.45
6	5000-10000	15395	2735	2733	6.27	2.96	4.61
7	10000-25000	12008	761	814	4.89	0.82	1.37
8	25000-50000	8953	402	436	3.65	0.43	0.74
9	50000-100000	7920	321	422	3.23	0.35	0.71
10	100000-250000	8066	284	392	3.29	0.31	0.66
11	250000-500000	4945	158	202	2.02	0.17	0.34
12	500000-1000000	4285	135	157	1.75	0.15	0.27
13	1000000-Infinity	13611	251	231	5.55	0.27	0.39
	Total	245,368	92,421	59,245			

Appendix C: Detailed Chromosomal Coverage Depth for All Samples

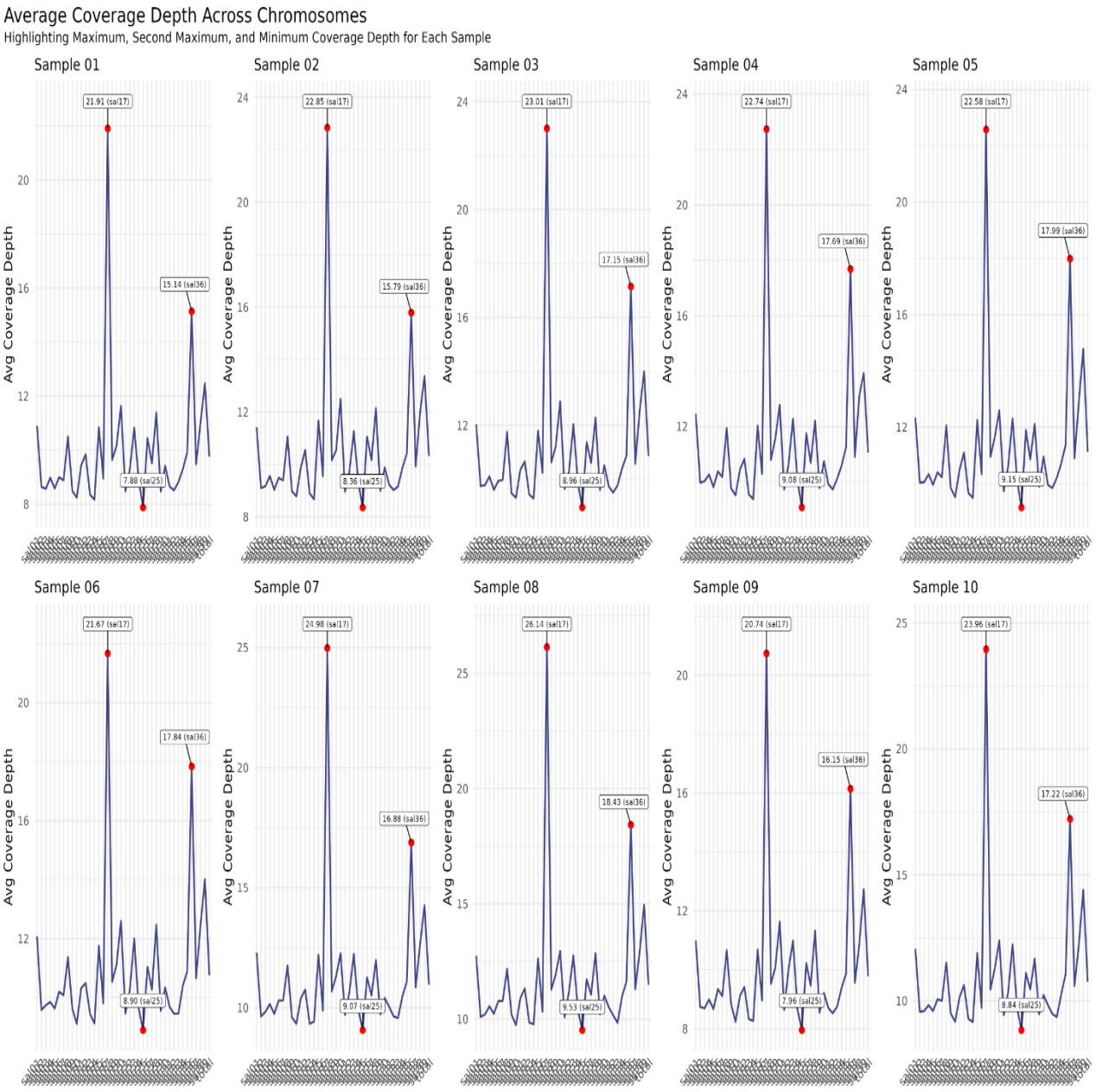


Figure C1. Comprehensive chromosomal coverage depth analysis for Samples 01 to 10: *This figure illustrates the average coverage depth across chromosomes for Samples 01 to 10. Each panel represents an individual sample, with the y-axis denoting the average coverage depth and the x-axis corresponding to the individual chromosomes. Red dots shows the chromosomes of the maximum coverage depth, the second highest coverage depth, and the minimum coverage depth within each sample, providing critical insights into coverage variability across the genomic landscape.*

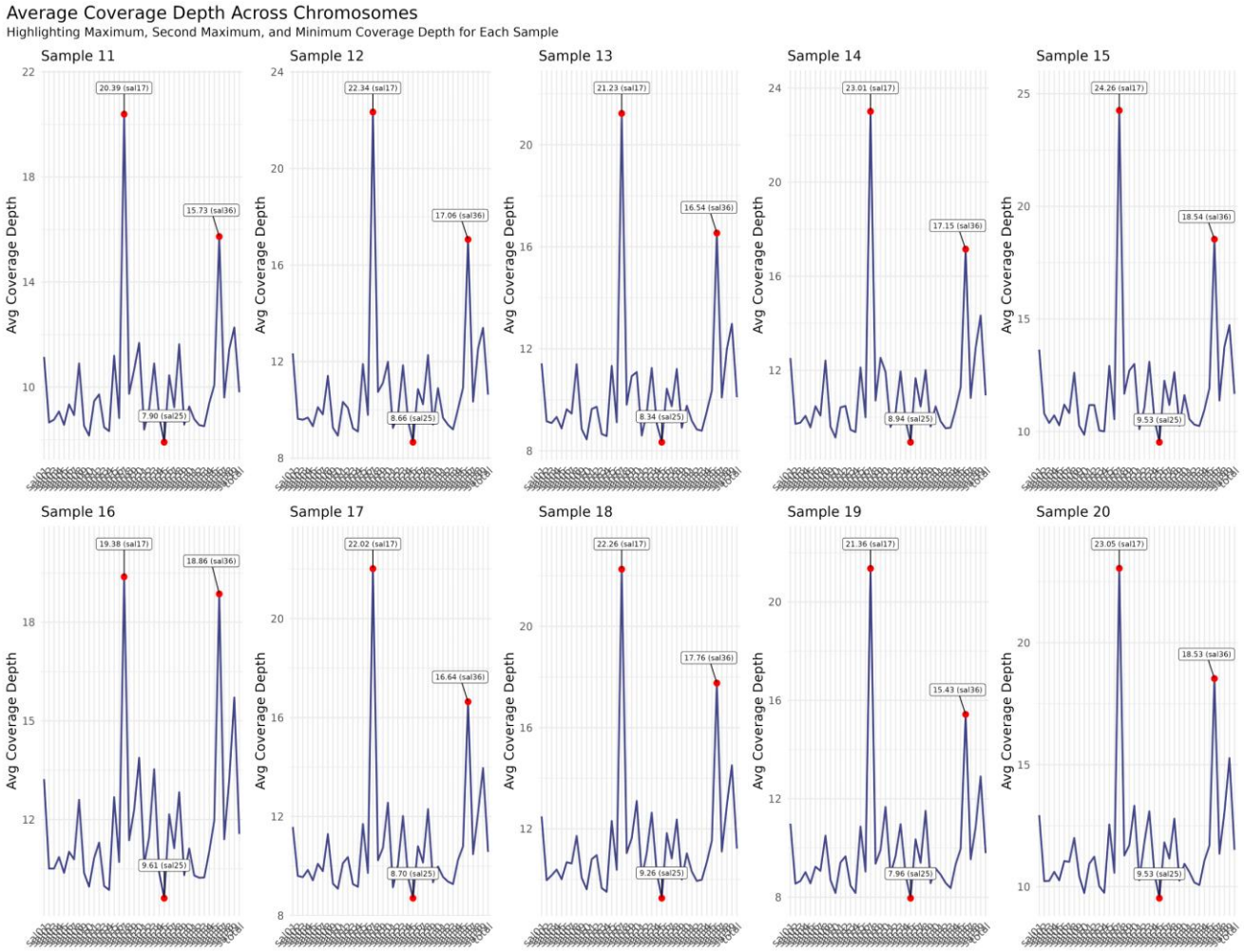


Figure C2. Sequential Chromosomal Coverage Depth Profiles for Samples 11 to 20: *Displayed are the average coverage depths across chromosomes for Samples 11 to 20. For each sample, the coverage depth is plotted along the y-axis against the chromosome number on the x-axis. Red dots are employed to mark the specific chromosomal points exhibiting the maximum, second maximum, and minimum coverage depths, thereby showing the distribution and range of coverage across samples.*

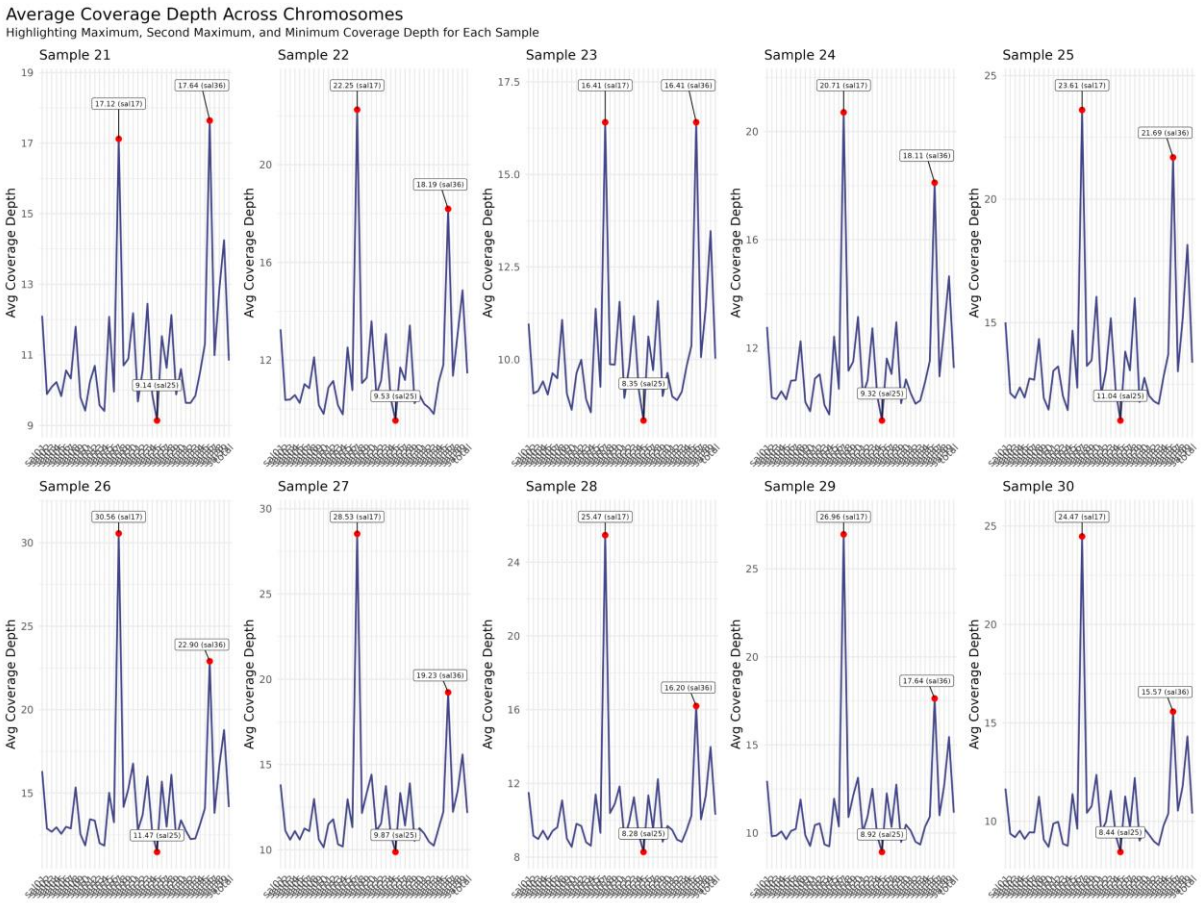


Figure C3. Average coverage depth across chromosomes for samples 21-30*, illustrating both the typical range and notable outliers. The red dots indicate the maximum, second maximum, and minimum coverage depths for each sample. Chromosome sal17 consistently shows higher coverage across all depicted samples.*

4.6 Appendix D: Comparative Analysis of Structural Variations and Sequencing Coverage Across 30 Samples

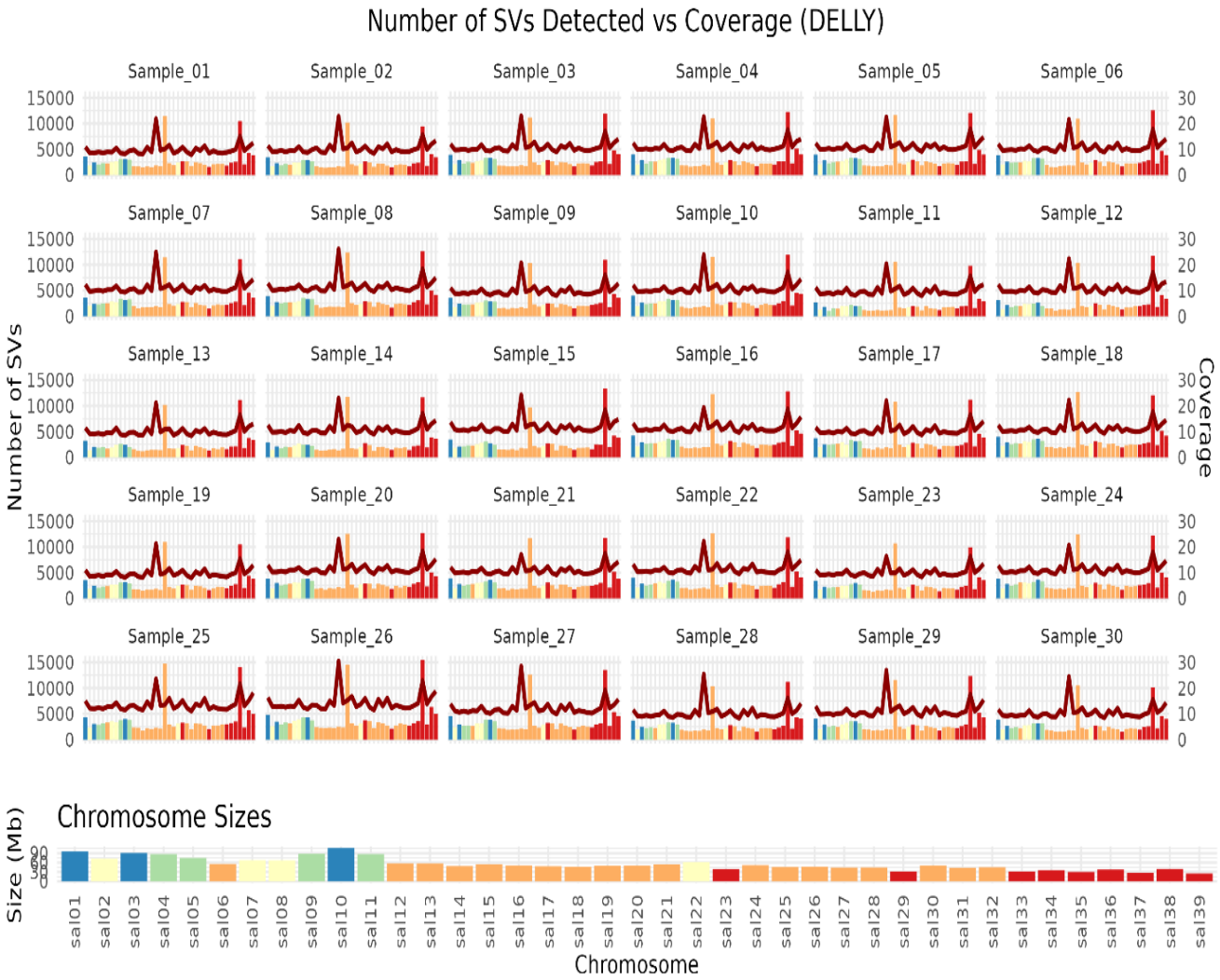


Figure D1. Comparative Overview of Structural Variations and Coverage across 30 Samples: *This figure presents a side-by-side comparison of the number of structural variations (SVs) detected and the sequencing coverage for each of the 30 samples. Each subfigure corresponds to a unique sample and is plotted with the number of SVs on the left y-axis and sequencing coverage on the right y-axis, across chromosomes. The bottom panel provides a color-coded reference for chromosome sizes, facilitating a correlation between chromosomal length, coverage, and SV detection. This comprehensive visual compilation allows for cross-sample comparisons and highlights the genomic architechure.*

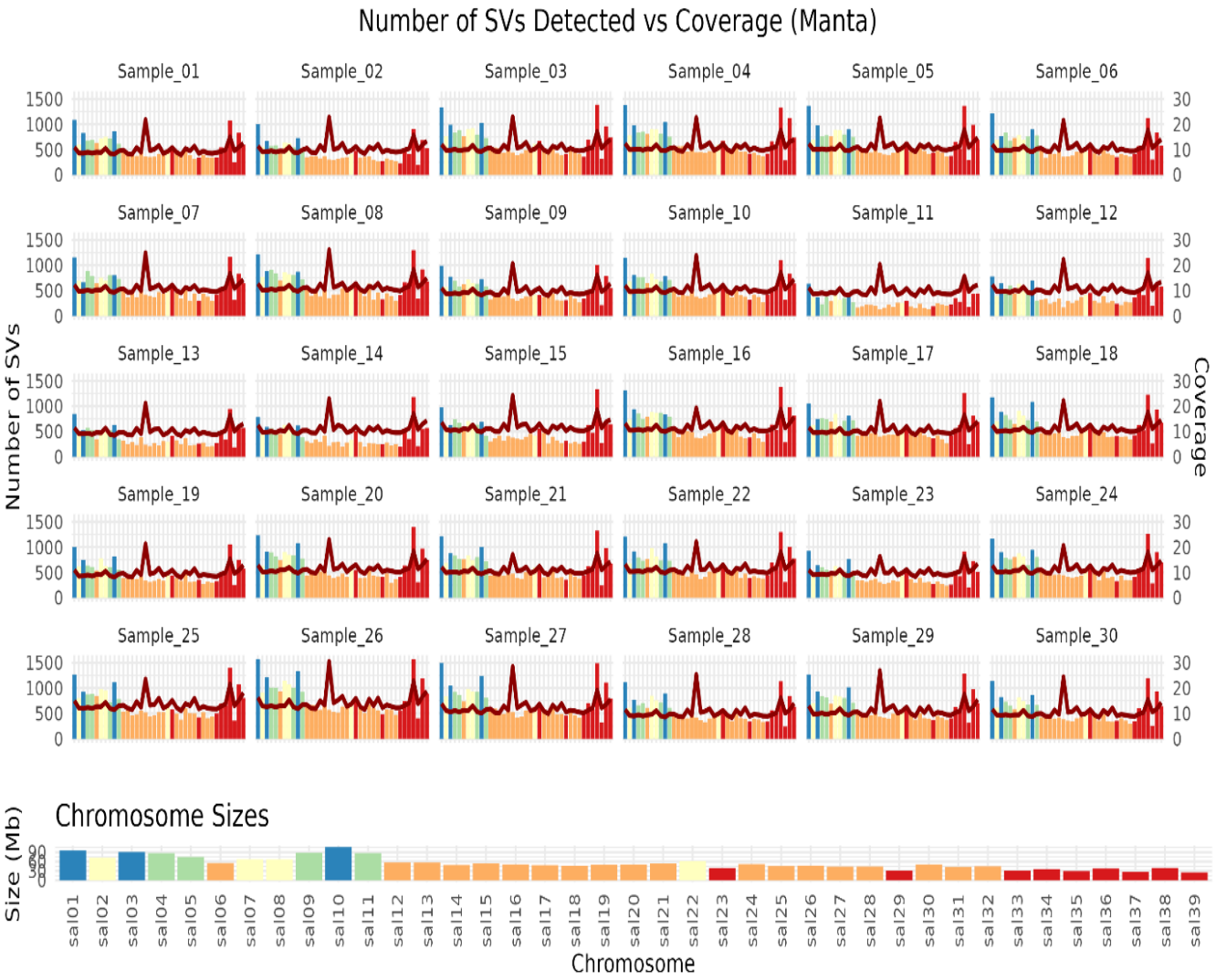


Figure D2. Structural Variations (SVs) and Sequencing Coverage Across 30 Samples*: This multi-faceted figure illustrates the number of SVs detected by Manta at chromosomal level. Each subplot represents the number of SVs and coverage data for one sample, with the number of SVs displayed by the bar graph (left y-axis) and sequencing coverage shown by the line graph (right y-axis) across the entire chromosomal span. The bottom of the figure features a color-coded key that corresponds to the sizes of the chromosomes.This collective visualization allows for cross-sample comparisons of SV distribution patterns and coverage metrics.*

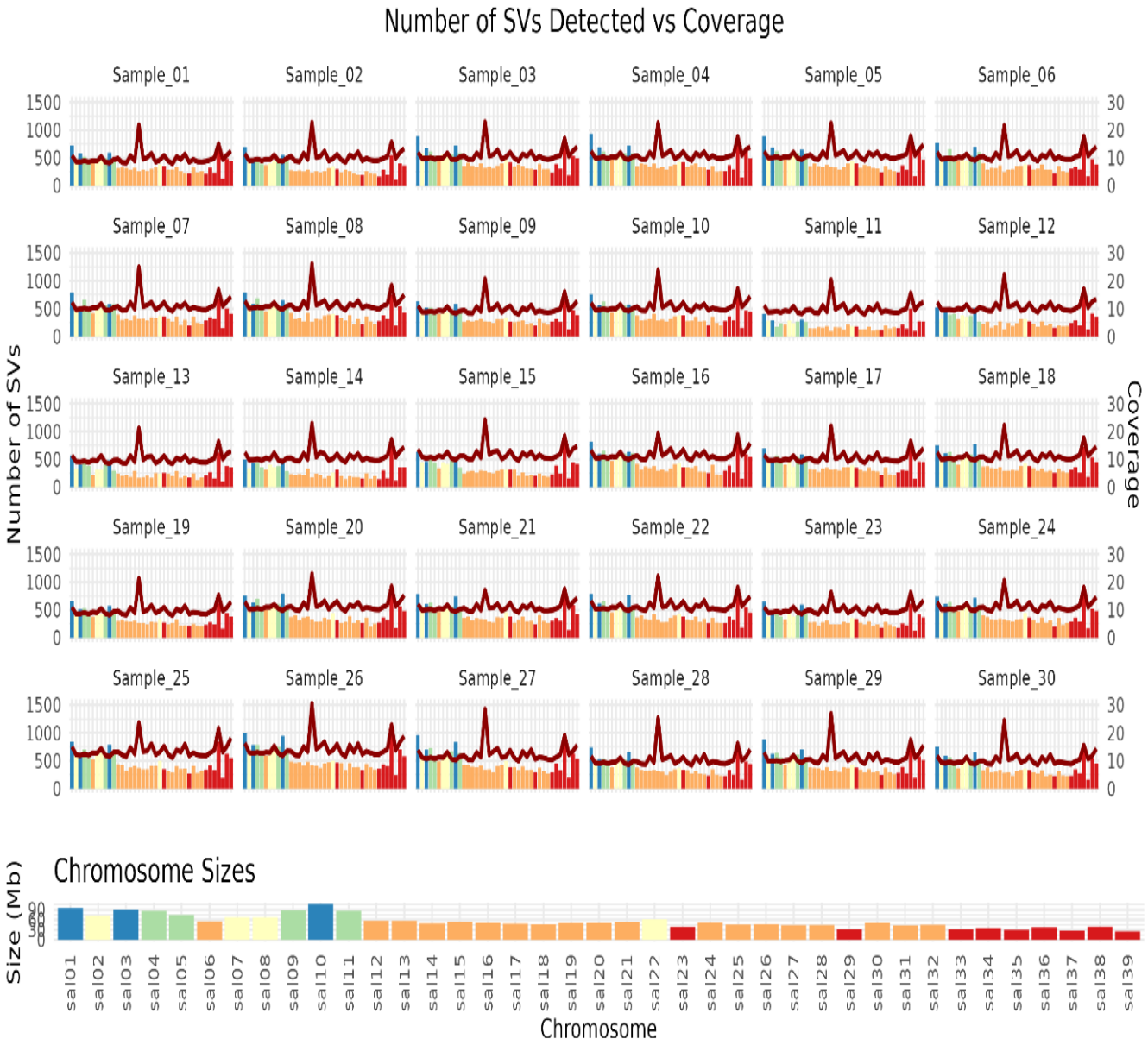
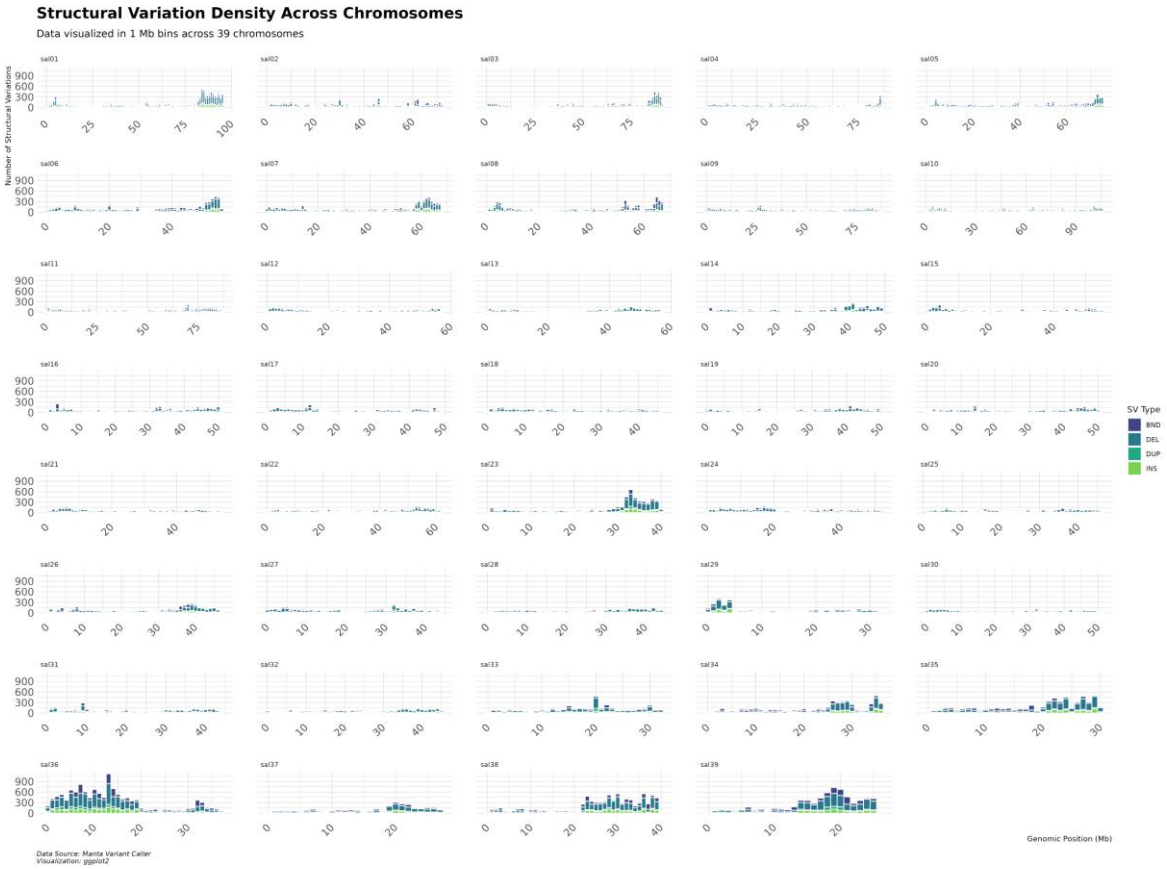


Figure D3. Multi-Sample Analysis of Structural Variations (SVs) and Sequencing Coverage*: This figure depicts the number of structural variations (SVs) detected alongside the sequencing coverage across 30 samples, with each sub plot representing a sample. The bar graph in each sub plot shows the count of SVs across the chromosomal level, while the overlaid line graph indicates the corresponding sequencing coverage depth. Chromosome sizes are provided as a reference in a color-coded key at the bottom, facilitating a comparison of SV distribution relative to chromosome length.*

Appendix E: SV density per chromosomal length

Figure E1. SV density per chromosomal length



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

Postboks 5003
NO-1432 Ås
Norway