Exploring Structural Variation in Tumor Evolution through Nanopore Sequencing

Student: Francisco José Villena González

MASTER'S DEGREE IN BIOINFORMATICS AND DATA SCIENCE FOR PRECISION PERSONALIZED MEDICINE AND HEALTH

2023–2024

Bioinformatics Unit of the Spanish National Cancer Research Centre

Master's Thesis Supervisor: Tomás Di Domenico

Submission Date: 15 de enero de 2025

Exploring Structural Variation in Tumor Evolution through Nanopore Sequencing

Abstract:

Structural variants (SVs) are genomic alterations encompassing deletions, insertions, and segment rearrangements, ranging from kilobases to entire chromosomes. Despite their significance as biomarkers in oncological diseases, these variants have remained relatively unexplored compared to single nucleotide variants, largely due to the inherent limitations of short-read sequencing technologies that have dominated large-scale genome sequencing projects. This scenario has undergone a transformative change with the advent of long-read sequencing technologies, which have enabled the achievement of the first truly complete human telomere-to-telomere reference genome, successfully filling gaps that short reads could not resolve. This project focuses on conducting a comprehensive performance evaluation of long-read-based structural variant callers, specifically in the context of tumor evolution analysis. To address the limited availability of appropriate datasets, we have developed specialized workflows leveraging high-performance computing resources for generating synthetic data with custom SVs, thus facilitating robust benchmarking of various structural variant detection methods. This computational approach enables systematic evaluation of SV detection algorithms under controlled conditions, providing valuable insights into their performance and reliability.

Key words:

Structural Variants, Cancer Genomics, Oxford Nanopore Sequencing, Long-Read Sequencing, High Performance Computing, Synthetic Data Generation, Variant Calling, Bioinformatics.

Acknowledgements:

Quiero empezar dando las gracias a Tomás, por haber sido y querer seguir siendo el mentor de este *bio* con interés por acercarse mucho más a lo *info*. También por romper mi maldición con los nanoporos, haciendo que este máster sobrepase todas mis expectativas.

Gracias a Fátima por incluirme entre sus filas en la Unidad de Bioinformática, y a todo el equipo, gracias a vosotros (y a los mixtos con huevo de la cafetería) la opción de teletrabajar se hace mucho menos atractiva.

Gracias a la Fundación Instituto Roche por la ayuda económica para cursar este máster, lo de que el dinero no da la felicidad es una mentira refutada por los que podemos seguir haciendo aquello que nos apasiona gracias a las becas.

Por último, a los culpables de que estos agradecimientos hayan sido en español, Nieves y Paco, mis padres. Gracias a sus esfuerzos he podido vivir la vida con el privilegio de también poder estudiarla.

Contents

1 Introduction **1**
1.1 Sequencing Technologies in Cancer Research 1
1.1.1 Limitations of Short-read Genome Assembly 2
1.1.2 Long-read Contributions to Genome Assembly 2
1.1.3 Structural Variation in Cancer 5

2 Objetives **7**

3 Materials and methods **9**
3.1 Computing resources 9
3.2 Software tools 9
3.2.1 Simulating long-read data and SV calling 10
3.2.2 Benchmarking of SV callers 13

4 Results **15**
4.1 Simulated data 15
4.2 SV calling 15
4.2.1 Computational demands 16
4.2.2 Calling performance 17

5 Discussion **19**
5.1 Computational and Resource Requirements 19
5.2 Simulation Design and Parameters 19
5.3 SV Caller Performance and Limitations 20
5.4 Clinical Relevance and Technical Challenges 20
5.5 Visualization Tools and Challenges 21
5.6 Future Directions 21

6 Conclusions **23**

A Appendix **33**

List of Figures

1.1 Limitations of short-read genome assembly 2
1.2 Gaps resolved by T2T assembly 3
1.3 Principle of nanopore sequencing 4
1.4 Evolution of ONT basecalling accuracy 4
1.5 Remission/relapse cycle of Multiple Myeloma 6

3.1 Simplified version of "visor-simulations" workflow for long-read simulation and SV calling 11
3.2 Simplified version of "bam-splitter" workflow for chromosomal splitting of BAM files 13

4.1 Computational resources demand of VISOR LASeR module 15
4.2 Computational resource demanded for SV detection methods 17
4.3 Performance of evaluated SV calling methods 18

A.1 R10 nanopore improvements over R9 34
A.2 Alignment-based SV Validation using GW 35

List of Tables

3.1 Technical specifications of computing nodes in CNIO's HPC cluster 9
3.2 Simulated chromosomal aberrations characteristic of Multiple Myeloma . . 12

A.1 Software tools used in this work 33
A.2 File sizes by coverage 34

1 Introduction

1.1 Sequencing Technologies in Cancer Research

Genomic sequences significantly influence organismal biology and provide insights into evolutionary history. The application of this genomic knowledge to human health has given rise to precision medicine, a discipline that leverages genetic and molecular markers to personalize medical treatments according to individual patient characteristics [1]. Advances in sequencing platforms and bioinformatics tools have enabled the identification of genetic variations that predispose individuals to specific diseases, facilitating the design of targeted therapies that enhance efficacy while reducing side effects [2].

In oncology, genomic analyses have revolutionized our understanding and therapeutic approach to cancer [3]. According to the National Cancer Institute (NCI), cancer encompasses more than 100 distinct types, arising when normal cells undergo genetic and/or epigenetic alterations leading to uncontrolled proliferation and compromised organismal homeostasis [4]. Cancer development represents a complex process resulting from interactions between individual genotype and various environmental factors, driving clonal evolution where distinct malignant cell subpopulations compete for resources and space [5]. Tumor DNA sequencing enables identification of patient-specific driver mutations, facilitating more effective targeted therapy selection and implementation [6].

DNA sequencing methods have evolved significantly over the years, with current technologies falling into three main categories:

- **Chain-termination method**: Also known as Sanger sequencing after its primary developer, this technique relies on controlled DNA synthesis termination, generating fragments of varying lengths that reveal the original sequence when size-separated. While largely superseded for genomic projects due to its low throughput, it remains valuable in research for verifying short reads such as PCR products from individual genes [7].
- **Short-read sequencing**: Technologies that perform massive parallel sequencing of clonally amplified DNA fragments (250-600 bp), delivering high sequencing depths with >99.9% accuracy [8]. Illumina dominates this category globally, though recent years have seen competition from MGI Tech, which promises faster, more cost-effective high-quality sequencing [9].
- **Long-read sequencing**: This approach also employs parallel sequencing strategies but generates individual reads spanning tens to thousands of kilobases. Currently, two main methods dominate this field: Single Molecule Real-Time (SMRT) sequencing,

commercialized by PacBio, and nanopore sequencing, pioneered by Oxford Nanopore Technologies (ONT) [8] with the recent appearence of MGI Tech as possible alternative [10].

1.1.1 Limitations of Short-read Genome Assembly

Cancer is largely driven by somatic changes in the genome, which can range from small nucleotide substitutions to chromosome-scale rearrangements. In this context, sequencing technologies play a crucial role, with predominant short-read sequencing having revolutionized our understanding of point mutations. However, this approach proves insufficient for resolving most large genomic alterations and generating gap-free assemblies. This limitation stems from the inherent inability of short reads (typically 150-300 base pairs) to span complex genomic regions, particularly those containing repetitive elements or large structural variations [11]. Furthermore, the fragmented nature of short reads complicates the accurate reconstruction of complex genomic architectures, often leading to ambiguous or incomplete assemblies (**Figure 1.1**).

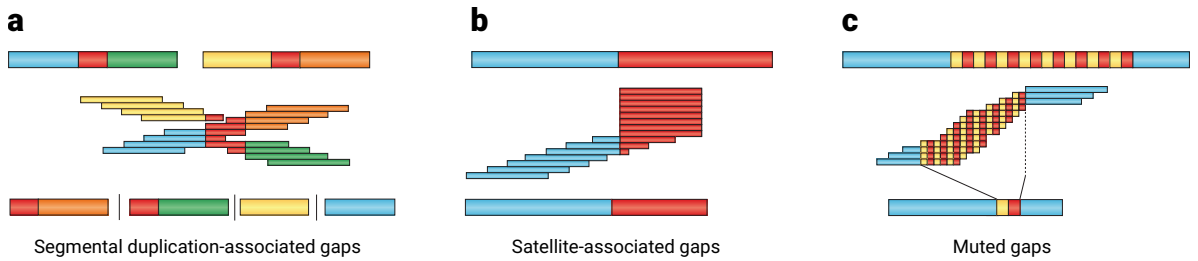

Figure 1.1: Limitations of short-read genome assembly. The upper bar of each figure shows regions to be resolved, with repetitive sequences highlighted in red. The middle displays short-read alignments, while the bottom bar shows the inferred sequence. (**a**) Large segmental duplications of high sequence identity (orange and green) create ambiguous read overlaps, resulting in multiple gaps flanking segmental duplications. (**b**) Satellite-associated gaps represent a special case causing read 'pileups' due to higher-order tandem arrays of repetitive sequences, primarily occurring in centromeric, acrocentric, and telomeric genomic regions. (**c**) Muted gaps occur when the assembled sequence appears shorter than the actual genome, typically in repetitive regions that are difficult to amplify or are toxic to bacterial cloning, such as simple tandem repeats. Adapted from [12].

1.1.2 Long-read Contributions to Genome Assembly

Long-read sequencing technologies emerged a decade ago, marking a turning point in sequence assembly by addressing short-read limitations. Initially, these technologies had much higher error rates (10%) compared to short reads ($< 1\%$) [13]. While this enabled routine bacterial genome assembly, it limited their application in human genomics, particularly for point mutation detection [14]. However, continuous improvements in long-read sequencing platforms have progressively reduced these error rates, ultimately enabling the generation of the first gap-free human reference genome, known as telomere-to-telomere (T2T) assembly [15]. This achievement resolved previously inaccessible genomic regions that remained incomplete in the Genome Reference Consortium's human reference version 38 (GRCh38), highlighted in red in **Figure 1.2**.

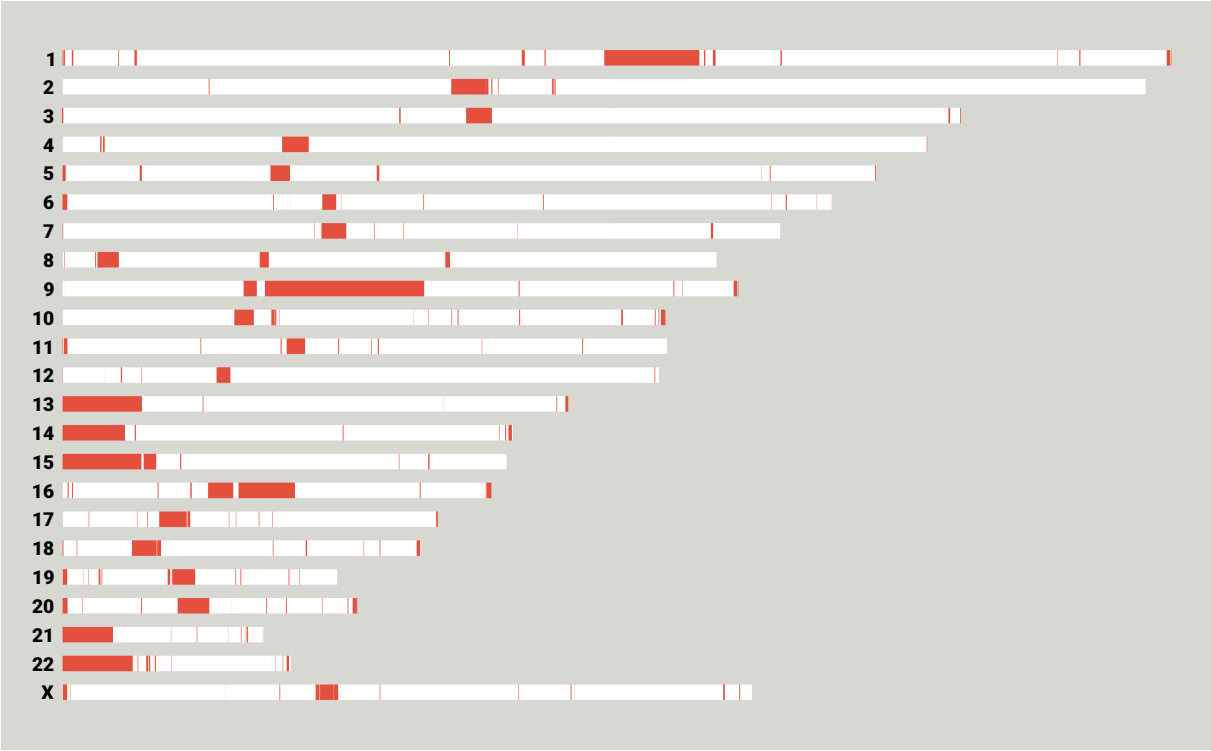

Figure 1.2: Gaps resolved by T2T assembly. Each bar represents a linear visualization of a chromosome, with chromosome numbers indicated on the left. Red segments denote previously missing sequences resolved by the T2T Consortium in 2022. The Y chromosome is not included, as its complex architecture, particularly its large, tandemly arrayed and inverted repeats (IRs), required additional analysis and was published separately in 2023 [16]. Adapted from [17].

Beyond sequencing methodology, significant differences exist between long-read platforms. Although PacBio pioneered the field in 2011 with high-throughput sequencing systems, their platforms have consistently required investments in the hundreds of thousands of dollars. In contrast, ONT's 2014 launch of MinION introduced a low-throughput but highly portable device requiring only a few thousand dollars investment. Subsequently, ONT expanded into high-throughput PromethION devices, notably the "P2 solo" model, which enables Whole Genome Sequencing (WGS) with a computer connection at under $20,000. This cost-effective solution makes long-read sequencing accessible to virtually any molecular biology laboratory [13], [18], [19].

ONT sequencing technology operates by measuring ionic current changes as single-stranded DNA molecules thread through nanoscale pores embedded in a membrane. Each nucleotide's unique shape creates distinctive current perturbations as it passes through the pore, enabling sequence determination and even modified bases detection. These nanopores are arranged in arrays across a flow cell, ONT's core consumable component, which contains thousands of individual sensing channels. Each flow cell integrates microfluidics for sample delivery, electronics for current measurement, and an application-specific integrated circuit (ASCI) that enables real-time data collection from multiple nanopores simultaneously (**Figure 1.3**).

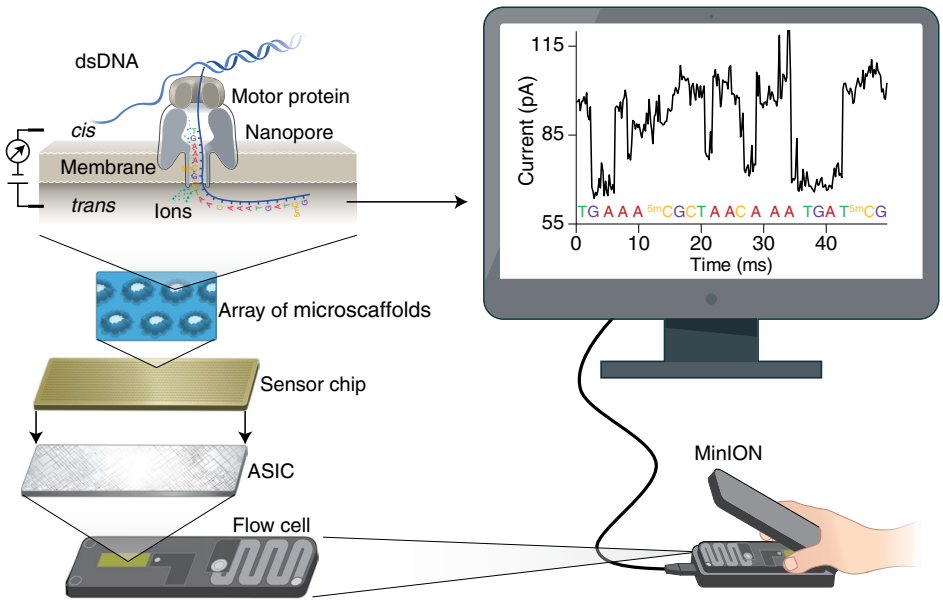

Figure 1.3: Schematic representation of nanopore sequencing using MinION technology. The diagram shows the key components: nanopore embedded in a membrane, array of microscaffolds, sensor chip, ASIC, and flow cell. The ionic current measurement graph displays the characteristic signal patterns produced as DNA strands pass through the nanopore. Adapted from [20].

Recent improvements in both the nanopore architecture of flow cells (transitioning from the R9 pore protein, with a 9 Å constriction, to the R10 variant featuring a longer barrel and 10 Å constriction) and basecalling algorithms have significantly enhanced sequence accuracy (**Figure 1.4**), making long-read sequencing a viable approach for the systematic analysis of cancer-associated variation [21]–[23].

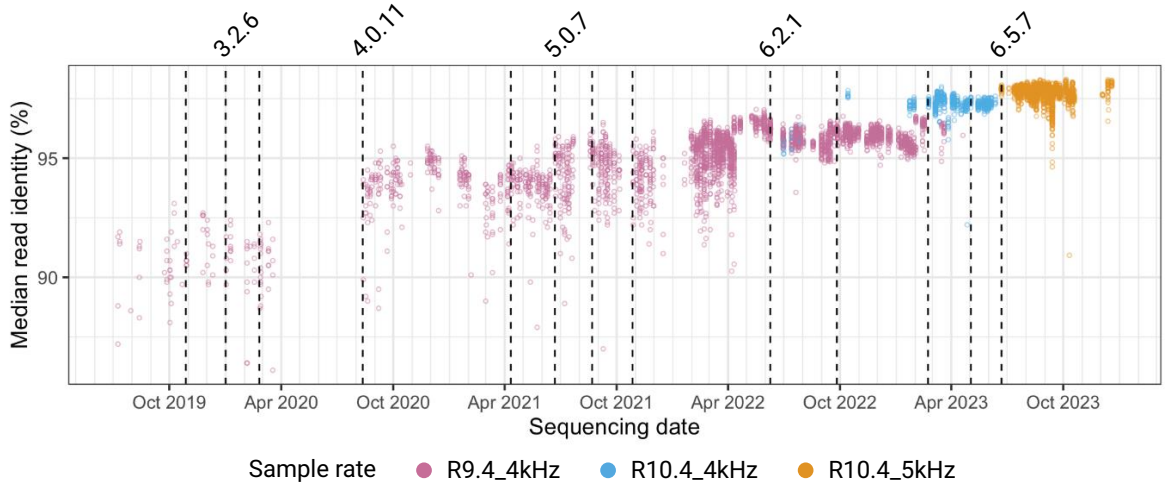

Figure 1.4: Temporal evolution of median read accuracy using Oxford Nanopore's Guppy basecaller versions shown above. Each data point represents a human WGS experiment at 20-30x coverage depth. Colors indicate the flowcell version and sampling rate: R9.4_4kHz (pink), R10.4_4kHz (blue), and R10.4_5kHz (orange), where sampling rate (kHz) represents the number of electrical measurements per second during sequencing (1 kHz = 1,000 measurements per second). Data from Genomics England's R&D Department, presented in ONT's Webinar "Unlocking comprehensive genome for large-scale projects", courtesy of Adam Giess and Melanie Tanguy [24].

The assembly of the first T2T genome required inputs from multiple platforms: PacBio for long and accurate HiFi data (15–25 kb at 99.5% accuracy), ONT for ultra-long (UL) data (>100 kb at 95% accuracy), and Illumina' short-reads (0.15 kb at 99.99% accuracy) [25]. While this combination of data types proved effective, it complicated data generation and limited accessibility. Currently, ONT provides all three sequencing modes on a single instrument using R10.4 PromethION flow cells, specific assembly chemistry kits, and associated bioinformatic workflows, resulting in automated assemblies with base accuracy exceeding 99.999% and near-perfect continuity [26]. This is particularly promising as it opens up the possibility of generating personalized human genomes using PromethION sequencers for precision medicine research purposes.

1.1.3 Structural Variation in Cancer

Structural variants (SVs) are genomic alterations ranging from 50 pb to whole-chromosome events, including deletions, insertions, and segment rearrangements. The significance of SVs as a hallmark of cancer is becoming increasingly evident. A recent analysis of 2,658 tumor genomes revealed that approximately 50% of driver mutations overlap with SVs, highlighting their crucial role in cancer development [27], [28]. Despite their importance as biomarkers in oncological diseases, our understanding of SVs remains limited compared to single nucleotide variants (SNVs). This knowledge gap stems primarily from the technical limitations of short-read sequencing, which has dominated large-scale genome sequencing projects. Using short reads, SV detection mainly relies on copy number variation (CNV) estimates, an indirect measure that fails to capture copy-number neutral events, inversions, and balanced translocations. Consequently, many cancer-driving SVs escape discovery using traditional short-read sequencing [29], [30].

The emergence of long-read sequencing technologies has marked a turning point in SV analysis, enabling improved variation detection in cancer genomes. A pioneering study using medulloblastoma cells, a primary childhood brain tumor, analyzed samples at diagnosis and post-treatment using long-read sequencing. This approach led to the identification of a novel mutational pattern called templated insertion (TI) thread, characterized by short (<1 kb) insertions that self-concatenate into highly amplified structures up to 50 kbp in size. This pattern was subsequently confirmed in other cancer types, showing particular prevalence in liposarcomas and frequent co-occurrence with chromothripsis, a catastropic mutational event where chromosomes shatter and reassemble chaotically [31].

Long-read sequencing has also enabled the characterization of novel mutation mechanisms driving genomic rearrangements in cancer. A notable example is the discovery of loss-translocation-amplification (LTA) chromothripsis in osteosarcoma. This mechanism, initiated by a single double-strand break, triggers simultaneous TP53 inactivation and oncogene amplification through breakage-fusion-bridge cycles. LTA chromothripsis appears to be uniquely prevalent in osteosarcoma, distinguishing it from other TP53-driven cancers [32].

Multiple Myeloma as a Model for SVs Detection

Multiple myeloma (MM) is a neoplasm of terminally differentiated B cells (plasma cells) characterized by frequent chromosomal translocations that place oncogenes under the control of immunoglobulin enhancers. Unlike most hematopoietic cancers, MM exhibits complex chromosomal abnormalities. Some of these SVs, first detected more than two

decades ago, remain major prognostic factors and are currently used for risk stratification in MM patients, primarily through Fluorescence in situ hybridization (FISH) detection [33].

MM follows a characteristic remission/relapse cycle where cancer cells progressively acquire resistance to different lines of treatment (**Figure 1.5**) [34]. In this context, ONT nanopore sequencing combined with automated SV calling methods could enhance our understanding of disease mechanisms and potentially identify new prognostic markers and treatment targets, ultimately contributing to improved patient care and survival outcomes. However, evaluating these strategies remains challenging due to two limitations: the lack of a golden standard for SV calling and the absence of curated, open datasets.

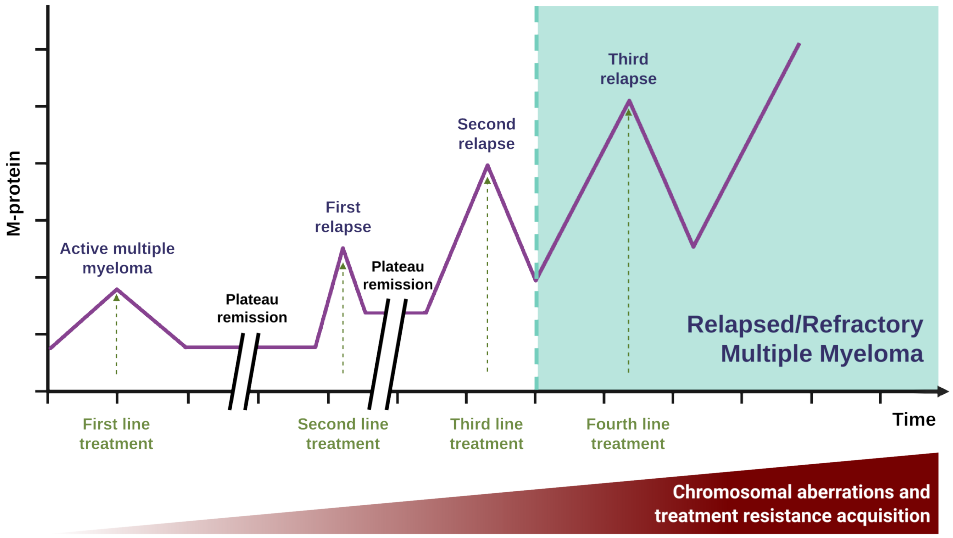

Figure 1.5: Remission/relapse cycle of MM. The disease trajectory is characterized by serial cycles of response, remission, and relapse in the presence of treatment, typically monitored through M protein levels (an abnormal antibody produced by malignant plasma cells). Through successive relapses, MM cells acquire new chromosomal abnormalities, leading to clonal evolution with diminished depth and duration of response over time. Courtesy of Álvaro Otero Sobrino, researcher of Hospital 12 de Octubre's Hematological Malignancies group.

2
Objetives

The primary aim of this research is to evaluate cutting-edge long-read sequencing approaches to track and analyze somatic structural variations across cancer genomes evolution. To this end, we have established the following specific objectives:

1. Assess the technical feasibility of detecting large-scale structural variants using synthetic long-read sequencing data that simulates current ONT flow cell protocols.
2. Evaluate and compare current SV callers, considering both their detection accuracy and computational efficiency, to identify optimal solutions for analysis workflows.
3. Validate the possibility of detecting clinically relevant Multiple Myeloma markers using long-read sequencing approaches, identifying both their potential and current limitations for clinical applications.

3 Materials and methods

3.1 Computing resources

Code development for this project was performed on a VANT MOOVE15 laptop with an Intel® Core™ i5-1235U processor, 64 GB DDR4 RAM, and 2 TB NVMe SSD, running Ubuntu 24.04.1 LTS.

Due to the computational demands of the tasks required to achieve the proposed objectives, the CNIO High-Performance Computing (HPC) cluster was utilized. The cluster currently features 12 compute nodes with configurations as detailed in **Table 3.1**.

Table 3.1: Technical specifications of computing nodes in CNIO's HPC cluster [35].

Count	Node names	CPU cores	RAM	GPUs
1	bc001	24	32 GB	–
6	bc00[2-7]	52	512 GB	–
3	bc00[8-10]	128	1 TB	–
1	hm001	224	2 TB	–
1	gp001	112	768 GB	3 x Nvidia A100 80 GB

The cluster's storage resources include 52 TB of standard storage space for user home directories, complemented by 512 TB of high-performance storage optimized for compute job input and output operations. From this high-performance storage, 30 TB was specifically allocated as project space for code execution and data generation.

3.2 Software tools

Visual Studio Code served as the primary interface for cluster access via SSH and code development. The implementation primarily utilized Bash, Python, and R programming languages.

The cluster operates under Slurm Workload Manager, a Linux/Unix-based system for HPC resource management. Workflow integration was accomplished through Snake-

make, a Python-based workflow manager that enables isolated software environments for each workflow step using Conda, thus avoiding node-specific installations and potential compatibility issues.

Miniforge, a minimal Conda distribution preconfigured with conda-forge as the default software, was installed to manage Conda environments. The Bioconda software repository was added to access specialized bioinformatics packages.

A detailed compilation of all software tools utilized throughout this research is presented in **Table A.1**. The subsequent subsections detail how these tools were strategically integrated into comprehensive workflows to address the project's specific objectives.

3.2.1 Simulating long-read data and SV calling

In the absence of curated long-read gold-standard datasets, testing SV calling methods with real biological datasets demands complex experimental procedures that are time-consuming, expensive, and potentially biased. In contrast, *in silico* approaches provide an efficient and accurate alternative to evaluate SV caller performance, where the ground truth is known.

Data simulator

VISOR toolkit was selected for haplotype-specific simulations of simple and complex SVs. Their VISOR LASeR module uses an error model trained on ONT R10.4.1 reads from 2023, provided by Badread [36].

SV callers

A set of SV calling tools was selected based on two key criteria: ONT long-read compatibility and somatic SV detection ability. Each selected caller provides distinct features valuable for this analysis:

- **SAVANA**: Implements a machine learning model trained on tumor-normal paired samples to detect somatic SVs and copy number aberrations (CNAs) in clinical cases.
- **Severus**: Specializes in tumor/normal comparative analysis, supporting multi-tumor samples and employing breakpoint graph frameworks for complex chromosomal rearrangement detection.
- **Sniffles2**: A pioneering ONT long-read SV caller since 2018, maintaining continuous development and regular updates.
- **SVision-pro**: Employs a neural network-based approach that converts genomic features from paired samples into image representations for comparative SV detection.

Generation and calling of SVs

A comprehensive workflow for VISOR-based simulations and SV calling analysis was developed. The complete code and documentation are available in the following repository: https://github.com/villena-francis/visor-simulations. **Figure 3.1** illustrates the key workflow steps.

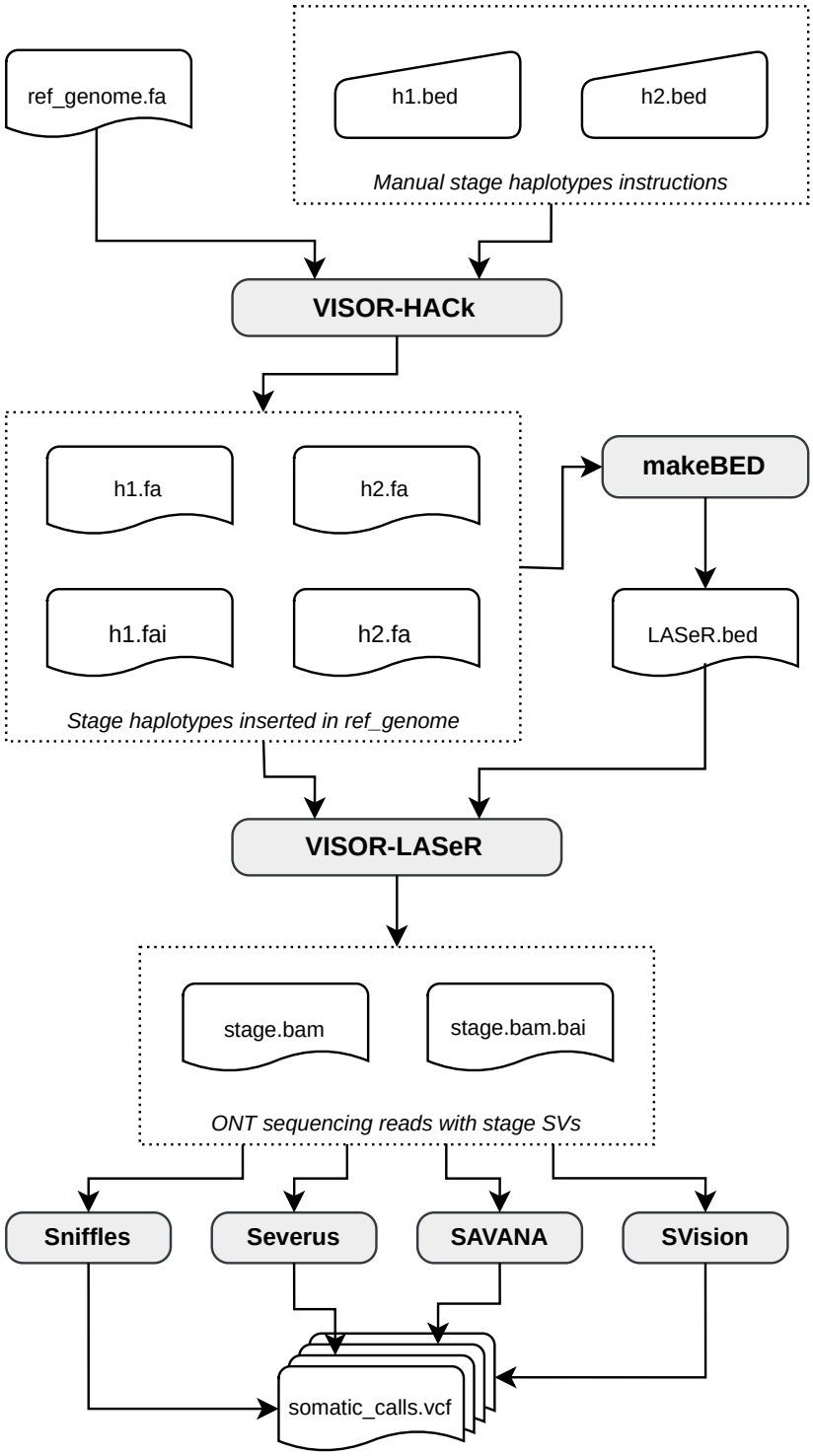

Figure 3.1: Simplified version of "visor-simulations" workflow for long-read simulation and SV calling. VISOR-HACk generates FASTA files (reference sequences) with incorporated SVs using a reference genome and BED-formatted haplotype instructions (tab-delimited genomic coordinates). The makeBED script creates a BED file (genomic intervals) from maximum chromosome sizes extracted from haplotype FASTAs. VISOR-LASeR then generates BAM files (aligned sequencing reads) and their indexes (.bai) using these files as input. each generating its corresponding variant call file (VCF) containing the identified SVs. The workflow parallelizes simulations across stages using configuration files and wildcards, producing multiple replicates at various sequencing coverages with corresponding normal samples.

Simulation stages

Each stage incorporates specific SVs into the GRCh38 reference genome, simulated at four coverage levels: 30x, 50x, 100x, and 200x. The initial stage generated one normal sample and three tumor replicates per coverage level, totaling 16 BAM simulations. Subsequent stages reused the normal samples, requiring only 12 simulations each through the automation provided by the "visor-simulations" workflow.

The stage generated for this project (v1) incorporated six chromosomal aberrations characteristic of multiple myeloma, including tandem duplications, deletions, and various types of translocations (**Table 3.2**). Breakpoint coordinates for these structural variants were approximately determined using the UCSC Genome Browser, as specific locations were not detailed in the clinical literature.

Table 3.2: Simulated chromosomal aberrations characteristic of Multiple Myeloma [37]. Size values represent the final length in base pairs (bp) after genome insertion. Tandem duplication consists of a 2,030,586 bp fragment repeated four times. Input files for VISOR read simulation available at https://github.com/villena-francis/visor-simulations/tree/main/resources/v1

SV type	Description	Size (pb)
Tandem duplication	Amplification of chromosome 1q (1q21+), representing one of the most frequent structural cytogenetic abnormalities in Multiple Myeloma, occurring in approximately 40% of cases .	10152930
Deletion	Deletion of chromosome 17p, present in approximately 10% of Multiple Myeloma cases, serves as a significant poor prognostic indicator. The minimally deleted region at locus 17p13 encompasses the tumor suppressor gene TP53	234564
Translocation (reciprocal)	Chromosomal rearrangement involving the immunoglobulin heavy chain (IGH) gene at 14q32, a hallmark genetic event in Multiple Myeloma pathogenesis.	1502367
Translocation (cut-paste)	Rearrangement involving CDKN2A, a crucial tumor suppressor gene whose inactivation through mutations or deletions is among the most frequent alterations in human cancers, second only to TP53 alterations.	31271
Translocation (copy-paste)	Structural variation affecting the KRAS oncogene region, whose alterations are frequently associated with Multiple Myeloma progression.	45683
Inversion	Chromosomal inversion at 6q25.1, included as a control variant to validate the detection capabilities of structural variant analysis pipelines, although not typically characteristic in Multiple Myeloma.	3600001

3.2.2 Benchmarking of SV callers

Data subsampling for local device processing

Visual inspection of BAM file reads is essential for validating simulation quality and SV caller accuracy. Due to cluster limitations with graphical interfaces, this analysis requires local processing. We developed a workflow to split whole-genome BAM files by chromosome, available at https://github.com/villena-francis/bam-splitter. **Figure 3.2** illustrates the workflow's key components.

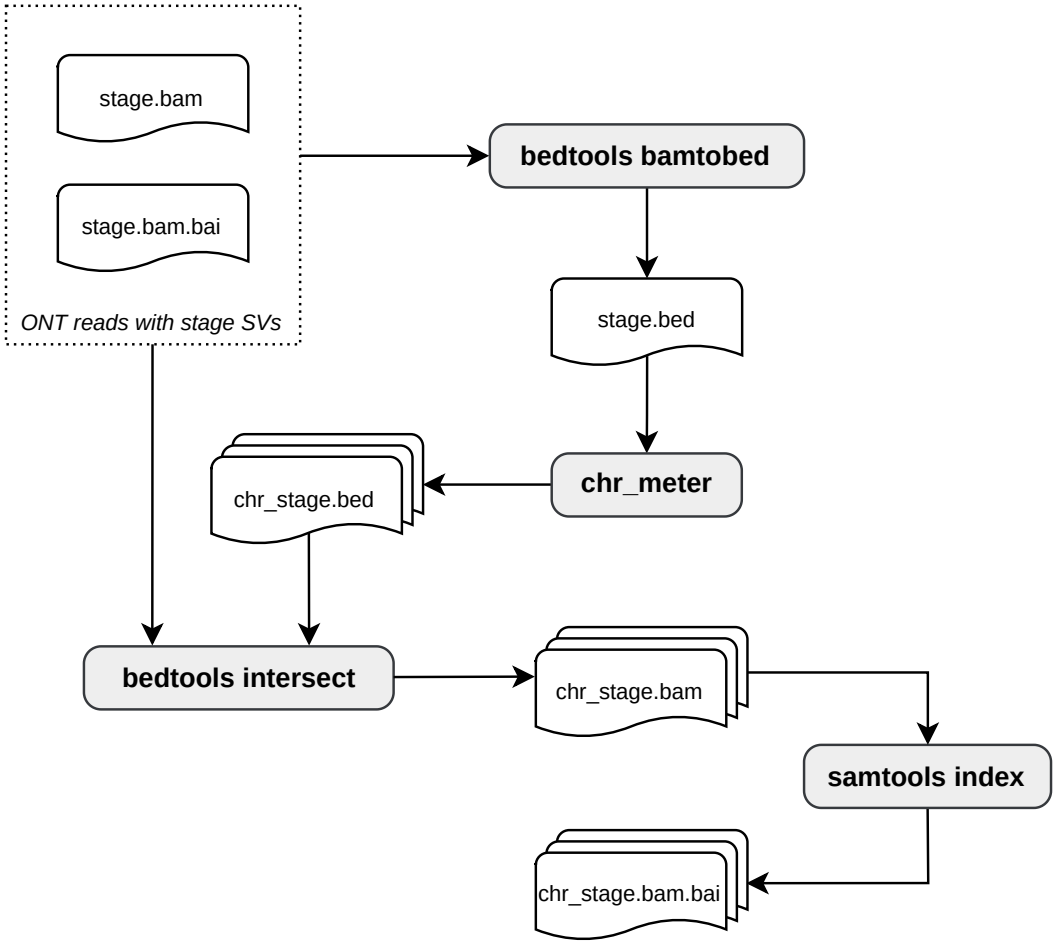

Figure 3.2: Simplified version of "bam-splitter" workflow for chromosomal splitting of BAM files. The process begins with bedtools bamtobed generating a comprehensive BED file of chromosome coordinates. The chr_meter script then creates individual BED files for selected chromosomes, which bedtools intersect uses to produce chromosome-specific BAM files. Samtools generates corresponding indexes for each BAM file. The workflow employs configuration files and wildcards to parallelize chromosome processing.

Visualization of long read alignments

Chromosomal BED visualization was conducted using Genome-Wide (GW), an advanced, ultra-fast genome browser capable of exploring extensive genomic regions and complete chromosomes. GW's specialized features, particularly its VCF-based manual curation capability, facilitated rapid validation of SV caller predictions.

Classification of SV calling results

Performance evaluation of SV callers employed a binary classification framework with the following criteria:

- **True Positives (TP)**: Successfully detected simulated SVs.
- **False Positives (FP)**: Caller-identified SVs absent in simulation, verified through read inspection.
- **False Negatives (FN)**: Simulated SVs present in reads but undetected by caller.

True Negatives (TN) were excluded from this analysis due to their inapplicability in SV calling evaluation, given the vast genomic space and nature of SV detection methods.

Metrics used

Due to the impossibility of quantifying true negatives in SV calling, we employed TN-independent metrics:

1. **Recall**: Proportion of correctly identified positive cases

$$Recall = \frac{TP}{TP + FN} \tag{3.1}$$

2. **Precision**: Accuracy of positive predictions

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

3. **F1 Score**: Harmonic mean of precision and recall

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.3}$$

Computational efficiency was evaluated using Slurm-generated statistics: CPU/GPU utilization (cores), RAM consumption (GB), and execution time (minutes).

Data Visualization and Statistical Analysis

Metric analysis and visualization were implemented through R scripts available at https://github.com/villena-francis/master_thesis/tree/main/data/cluster_bmk.

4 Results

4.1 Simulated data

VISOR simulations generated a total of 14 TB of simulated ONT long-read sequencing data. Of this volume, approximately 5.6 TB corresponded to reference genome-aligned reads in BAM format and their respective indices, while the remaining data volume consisted of unaligned reads in FASTQ format (**Table A.2**). The computational resources required for the simulation process were monitored, with CPU utilization, RAM consumption, and generation times detailed in **Figure 4.1**.

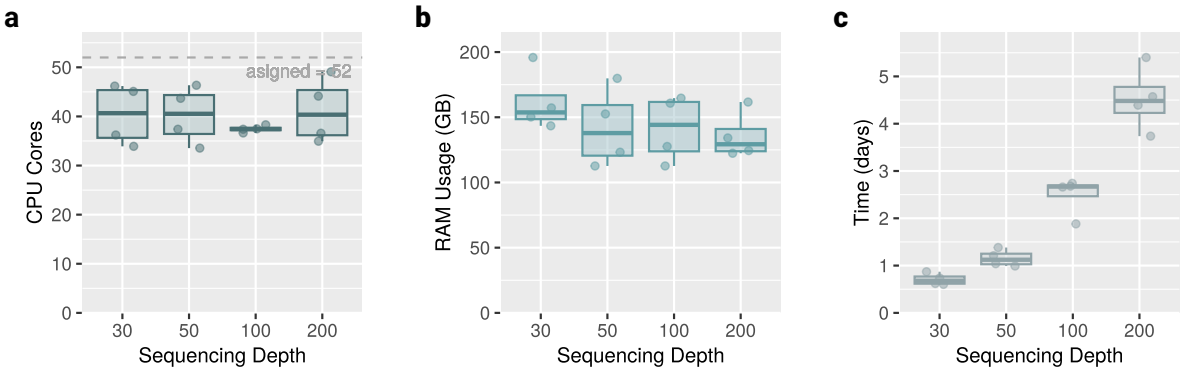

Figure 4.1: Computational resources demand of VISOR LASeR module by sequencing depth, which represents the simulation bottleneck due to its CPU cores (**a**) and RAM memory (**b**) consumption over time (**c**). For each sequncing deph, three "tumour" and one "normal" WGS technical replicas were obtained by simulating long reads. Calculations for VISOR HACk module were not performed since it runs only once for a few minutes with low resource consumption to introduce the SV set into the reference genome that will feed VISOR LASeR. The raw data used for all calculations is available in REPO.

4.2 SV calling

SV callers generate VCF files similar to those produced by SNV callers. However, in SV calling, each row represents a breakpoint, which defines the boundary of a structural variant. Simple SVs, such as insertions or deletions, require the identification of two breakpoints, while complex variants like reciprocal translocations necessitate the detection of four breakpoints. To illustrate the final output files generated by each caller, results from a 50x coverage simulated sample are available at https://github.com/villena-francis/master_thesis/tree/main/data/vcf_examples.

Standard VCF Outputs

- **SAVANA**. A significant limitation in SAVANA's VCF output, compared to other variant callers, is its inability to classify SV types. The tool only identifies and correlates breakpoints without providing information about the specific type of SV present.

- **Severus**. Beyond conventional SV detection, Severus incorporates specialized processing for variable number tandem repeats (VNTRs), enabling precise annotation of structural variants within these regions. The tool provides comprehensive output including standard VCF files, detailed quality metrics in log files, and graphical representations of chromosomal rearrangements through visualization plots.

- **Sniffles2**. This caller generates standard VCF output exclusively, without additional supporting files or visualizations.

- **SVision-pro**: Despite leveraging an innovative image-based encoding approach to analyze genomic features and detect inter-genome variations between tumor-normal paired samples, SVision-pro's output is limited to standard VCF files. The visual representations used in its internal processing pipeline are not accessible in the final output.

4.2.1 Computational demands

While computational requirements for analyzing real samples may be higher due to the larger number of SVs typically detected in actual cancer genomes, our simulation-based analysis provides proportional estimates of computational demands across different sequencing coverages for each of the four callers evaluated. Core utilization, RAM consumption, and execution times are illustrated in **Figure 4.2**.

A key distinction in processing unit allocation lies with SVision-pro, whose neural network-based SV detection model is GPU-optimized, whereas SAVANA, Severus, and Sniffles2 rely on CPU-based computation. We allocated 24 cores to GPU-based callers, as preliminary testing showed peak utilization remained below 20 cores.

The initial RAM allocation of 40 GB proved sufficient for Severus, Sniffles2, and SVision-pro, with memory usage increasing linearly with sequencing depth. However, SAVANA consistently exceeded this limit across all sequencing coverages, demonstrating significantly higher memory requirements compared to other callers.

SVision-pro demonstrates the lowest execution times across all sequencing depths, followed closely by Sniffles2. Both callers show efficient performance scaling, with execution times increasing sub-linearly with sequencing depth. SAVANA shows the highest execution times at the three lower sequencing coverages (30x, 50x and 100x), while Severus requires the longest processing time at maximum coverage (200x).

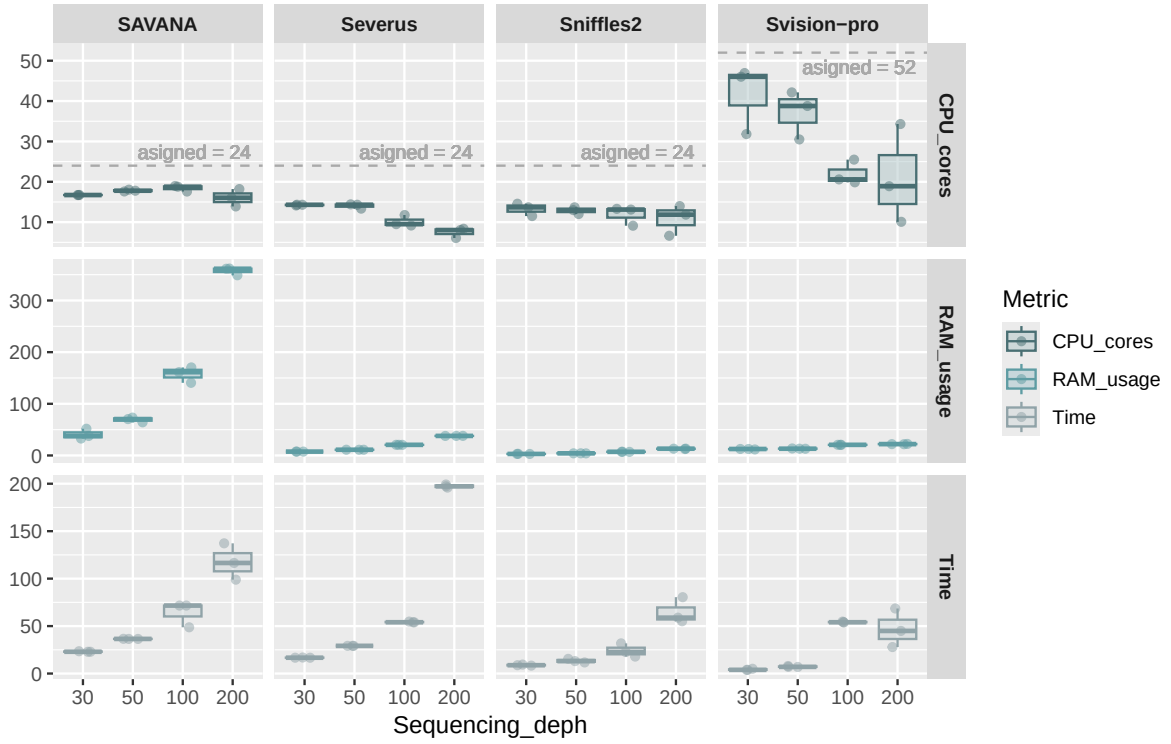

Figure 4.2: Computational resources demanded for SV detection methods of SAVANA, Severus, Sniffles2, and SVision-pro, measured in CPU cores, RAM memory (GB), and execution time (minutes). For each sequencing depth, three measurements per metric were obtained by comparing a single normal sample against three technical replicates of the tumor sample. The complete raw data used for all calculations is available in https://github.com/villena-francis/master_thesis/tree/main/data/cluster_bmk/hpc_data.

4.2.2 Calling performance

From the proposed SVs in synthetic data, SAVANA and Severus successfully identified all instances of tandem duplication, deletion, inversion, and reciprocal translocation. Although cut-paste and copy-paste translocations were included in the simulation instructions, only the deletion component of cut-paste events was successfully reconstructed in the BAM files, while the translocation component and copy-paste events failed to be reconstructed.In contrast, neither Sniffles2 nor SVision-pro detected any of the simulated SVs, explaining the stark differences in Precision, Recall, and F1 scores shown in **Figure 4.3**.

Visual inspection through GW of VISOR-generated BED files confirmed the presence of SVs detected by SAVANA and Severus, while validating the absence of undetected variants. This verification ruled out false negatives, resulting in a recall of 1.0 for both callers. Notably, Severus reported additional findings from its specialized VNTR analysis, which, although not explicitly introduced in simulation instructions, were present as artifacts from VISOR-LASeR's error model.

While SAVANA and Severus achieved identical recall values, SAVANA demonstrates higher precision due to Severus reporting slightly more false positives. Despite generating significantly more variant calls than both SAVANA and Severus, neither Sniffles2 nor SVision-pro detected any of the expected simulated SVs. Extensive visual inspection of their findings revealed no overlap with validated variants, resulting in their classification as false positives and consequently zero precision scores.

As a balanced measure combining precision and recall, F1 score reflects the overall performance of each caller. SAVANA achieves the highest F1 score due to its perfect recall and superior precision. Severus follows closely, with its slightly lower F1 score attributed solely to its marginally higher false positive rate, as it maintained perfect recall. Both Sniffles2 and SVision-pro yield F1 scores of zero, as expected from their complete absence of true positive findings and high number of false positive calls.

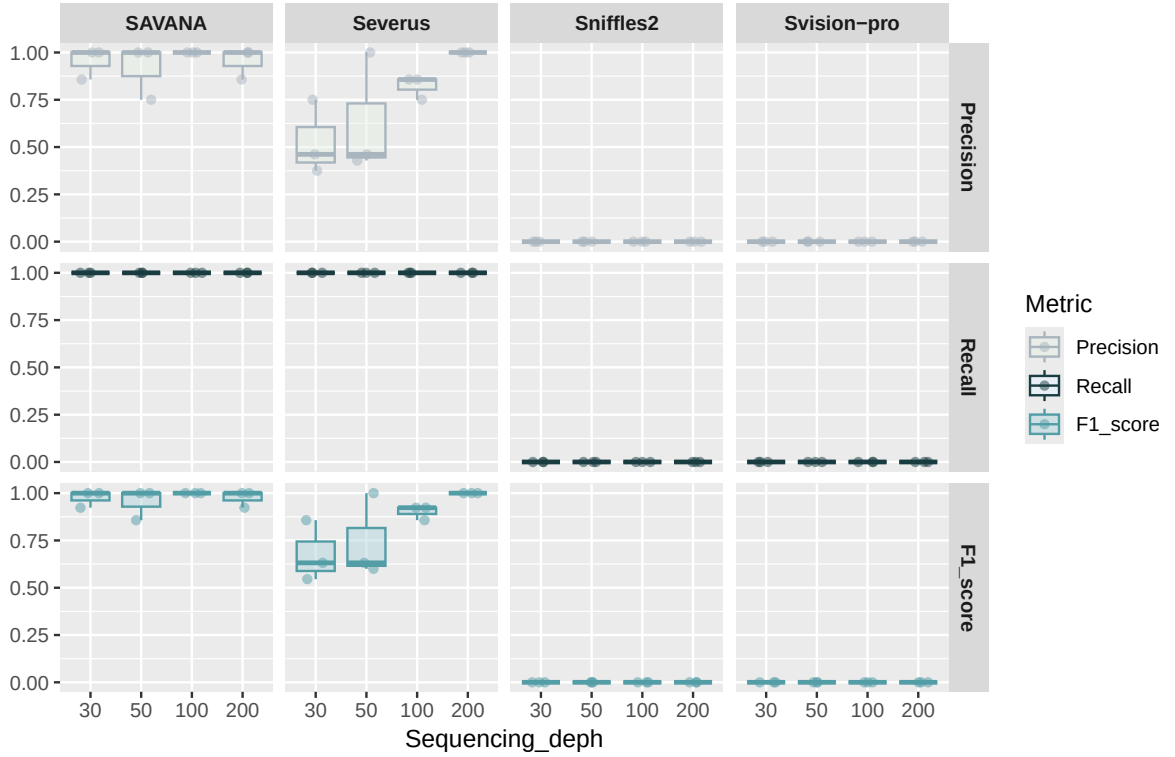

Figure 4.3: Performance of SV calling methods of SAVANA, Severus, Sniffles2 and SVision-pro, based on precision, recall, and F1 score metrics. For each sequencing depth, three results per metric were obtained by comparing one normal sample against three technical replicates of the tumor sample. The raw data used for all calculations is available in https://github.com/villena-francis/master_thesis/tree/main/data/cluster_bmk/calls_data.

5 Discussion

5.1 Computational and Resource Requirements

The synthetic data analysis configuration in this project demanded substantial computational resources: 14 TB of storage (2.73% of CNIO's HPC cluster capacity) and 832 cores for VISOR-LASeR execution (52 cores per sample, exceeding 114% of available cores). Furthermore, the single-job-per-GPU restriction for SVision-pro execution led to extended processing times. This intensive resource utilization presented significant challenges given the cluster's shared nature among multiple CNIO research groups and projects. These findings emphasize that computational tool selection should consider not only accuracy but also efficient resource management based on available computing resources as critical evaluation criteria.

5.2 Simulation Design and Parameters

The selection of SVs for simulation was influenced by ongoing research in the CNIO Bioinformatics Unit's long-reads group, specifically their collaboration with Hospital 12 de Octubre's Hematological Malignancies group on WGS of MM patient samples, comparing pre-treatments and relapses. Consequently, the SV set includes both characteristic MM variants and more speculative cancer-related SVs to cover a broader spectrum of SVs. Despite the availability of T2T genomes, GRCh38 was chosen as the reference genome due to its extensive use in research and the substantial accumulation of annotation-associated findings throughout its trajectory.

The SV simulation was configured to represent bulk sequencing of a homogeneous cell population containing a single clone, with all variant allele frequencies (VAFs) set to 0.5. This configuration ensures variants are present on one allele and represented in half of the generated sequencing reads. Such design aimed to provide clear variant representation in the reads, theoretically ensuring reliable detection by callers while facilitating visual validation through GW. However, this idealized scenario differs from real tumor samples, where variant allele frequencies can vary significantly due to tumor heterogeneity and normal cell contamination [38].

Reads were simulated with a mean length of 15,000 bp and a standard deviation of 13,000 bp, using VISOR-LASeR default settings. These parameters align with realistic values obtainable through actual sequencing using the PromethION platform's protocol for 10 kb human DNA with Ligation Sequencing Kit V14 [39], which has been selected for sequencing the MM patient samples. Notably, this protocol aims to generate ∼30-40x

genome coverage. Our benchmarking results at higher coverages (100x and 200x) showed no substantial performance improvements for any caller in detecting simulated SVs, suggesting that the additional costs and efforts associated with using multiple flow cells and preparing larger sample quantities may not be justified. However, Severus detected unplanned VNTR anomalies at lower coverages (30x and 50x), introduced by the Badread error model trained on ONT R10.4.1 reads. This observation is crucial for real sample sequencing: while such artifacts were more prevalent in R9.4 reads (see **Figure A.1**), they might still influence SV calling results, as even a small number of reads containing these sequencing artifacts could be considered representative by callers at the coverage levels currently achievable with a single flowcell.

5.3 SV Caller Performance and Limitations

Performance metrics position SAVANA as the top performer, yet its limitation to breakpoint correlation without SV classification represents a significant drawback. This limitation, combined with substantially higher RAM consumption and execution times compared to other callers, impacts its practical utility. Severus, despite slightly lower precision, matches SAVANA's recall while maintaining lower RAM usage and provides comprehensive SV classification alongside visual representations of chromosomal rearrangements. These characteristics likely influenced Severus's integration into the latest EPI2ME release, Oxford Nanopore's open-source platform designed to provide wet lab scientists with a user-friendly interface for data analysis without requiring advanced bioinformatics skills. However, despite its integration, EPI2ME's documentation lacks comparative analyses justifying Severus's selection over other SV callers, possibly due to its target audience [40].

Our evaluation employed minimum argument sets for all callers, relying on developer-configured default parameters for complex adjustments. This standardized approach may explain the poor performance of Sniffles2 and SVision-pro, particularly in detecting large SVs, as default settings might not be optimized for such variants. Notably, following our analysis, Sniffles2 version 2.5 released improvements specifically targeting detection of large deletions and duplications ($>$ 50 kb) [41]. Interestingly, the better-performing tools in our analysis (SAVANA and Severus) remain available only in preprint servers, while Sniffles2 and SVision-pro are published in peer-reviewed journals.

5.4 Clinical Relevance and Technical Challenges

VISOR toolkit proved valuable for evaluating SV callers' capability to identify characteristic large structural events in Multiple Myeloma using ONT long reads, particularly those with diagnostic significance. Through synthetic data generation, we successfully validated the detection of two of the most frequent chromosomal aberrations in Multiple Myeloma, notably the largest SVs in our simulation set, using SAVANA and Severus: the tandem amplification of 1q21+ (100 Mb) and an IGH-involving translocation (1,5 Mb). These structural variants, currently verified in clinical settings through FISH due to the limitations of short-read sequencing assembly, represent critical diagnostic markers that could potentially be identified through long-read sequencing approaches.

The inability to simulate certain SVs provides valuable insights into technical limita-

tions. While VISOR-HACk module unambiguously inserts all chromosomal abnormalities into a FASTA file using BED-formatted instructions and the reference genome as a template, challenges emerge in the VISOR-LASeR module, which generates and aligns reads to the reference genome using minimap2. These limitations manifest in two ways: standard read lengths may be insufficient for reconstructing certain events, suggesting the potential need for ultra-long read protocols capable of generating sequences up to 4 Mb, and minimap2 alignment accuracy may be compromised for complex variants. For instance, in the copy-paste translocation designed to generate a proximal KRAS duplicate, the aligner appears to have defaulted to mapping all reads to the original gene position. Similarly, for the cut-paste translocation, only the deletion component was detected, while the inverted sequence insertion failed to be properly positioned, leaving its corresponding reads unaccounted for in the alignment.

5.5 Visualization Tools and Challenges

Initial visualization attempts of synthetic data were made using the widely-adopted Integrative Genomics Viewer (IGV) version 2.17.3. However, despite using chromosome-specific BAM files generated through the "bam-splitter" workflow, IGV's performance proved inadequate, exhibiting slow loading times and frequent crashes. Through email correspondence, Severus's lead developer shared similar experiences with long-read BAM files in IGV, suggesting a workaround of generating smaller BAM files containing only the SV regions with 1 Mb upstream and downstream sequences. While this approach would have been feasible for planned SVs, it proved impractical for investigating additional findings like VNTR anomalies due to automation limitations.

GW emerged as a capable alternative, efficiently handling both individual chromosome and whole-genome BAM files. Its VCF compatibility created an interactive index for rapid SV coordinate navigation, dramatically accelerating event verification. However, GW requires terminal-based operation in conjunction with an alignment view display, demanding more advanced computational skills compared to IGV's graphical user interface **Figure A.2**. Nevertheless, GW's developer demonstrated strong responsiveness to error reports and feature requests through Github. Our project-specific experience led to reporting Conda installation [42] and loading genome anotation files [43] issues, and requesting vector format export capabilities for alignment visualizations [44].

5.6 Future Directions

The development of robust tools for SV analysis in cancer genomes requires extensive testing and validation. While this study evaluated SV detection capabilities using synthetic data across different sequencing coverages, it primarily demonstrates the value of generating simulated datasets for benchmarking existing tools, developing new algorithms, and training machine learning models. These computational advances ultimately contribute to improving cancer genomics analysis in clinical settings.

A primary challenge lies in identifying structural events that, despite existing in the template genome, remained undetected in alignments. The unaligned reads in FASTQ format will be valuable for expanding benchmarks to include long-read aligners beyond

minimap2, enabling more comprehensive tool evaluation. This expansion is crucial for improving the reliability of genomic analysis in clinical settings.

To enhance the exploration of SVs in cancer, future work should investigate several key aspects. First, examining the impact of read length on Sniffles2 and SVision-pro's calling capabilities, starting with sizes successfully identified in their published work. Second, introducing tumor heterogeneity through multiple clone combinations in synthetic sequencing data would better reflect real tumor complexities and allow evaluation of how varying VAFs affect SV detection capabilities. Third, exploring ultra-long read protocol parameters could help reconstruct previously undetected SVs. Based on our findings and resource efficiency considerations, these tests could focus on 30x and 50x sequencing coverages, aligning with practical clinical sequencing depths.

Data collection automation represents another critical area for improvement in clinical implementation. Currently, computational performance statistics are manually extracted from Slurm logs, a time-consuming process that could be streamlined through automated scripts. In anticipation of future evaluations, we have requested VISOR toolkit enhancements to log all parameters, both specified and default, facilitating automated tracking of simulation characteristics [45]. Additionally, The implementation of Truvari, a comprehensive toolkit for benchmarking, merging, and annotating structural variants from VCF files, would enhance the workflow [46]. Its capabilities make it particularly suitable for analyzing tumor genome evolution through longitudinal sequencing and SV calling analysis.

Continuous communication with tool developers remains essential for optimizing clinical applications, sharing synthetic data experiences, and facilitating potential improvements. In this context, frequent requests for GW enhancements, particularly regarding data visualization and high-quality figure export capabilities, will support better clinical result interpretation and documentation.

Most critically, the experience gained through synthetic data analysis must be validated on ONT-sequenced tumor-normal paired patient samples, integrating structural variation analysis with SNVs and methylation data. This comprehensive genomic characterization approach aims to provide clinicians with more accurate and complete information for personalized cancer treatment decisions, ultimately improving patient outcomes through better-informed therapeutic strategies.

6
Conclusions

Experimental data is essential for creating and improving computational methods, whose outputs not only provide new knowledge but can also serve as a foundation for refining future experiments. Synthetic data can catalyze this cycle, generating information faster and with fewer resource expenditures. Based on the results of this work, we can conclude the following:

1. Long-read WGS data obtainable with a single ONT PromethION flow cell enables the reconstruction of large-scale SVs, including insertions, deletions, translocations, and inversions spanning complete chromosomal bands.
2. Severus emerges as a balanced solution for SV detection, combining comprehensive variant classification capabilities with efficient resource utilization, making it suitable for routine analysis workflows.
3. Long-read sequencing has the potential detect of clinically significant MM markers, promising to enhance current diagnostic capabilities despite some technical limitations in SV detection that remain to be addressed.

These findings provide valuable guidance for SV detection implementation while establishing a framework for synthetic data generation that can be used to benchmark tools and train future machine learning models. This work also highlights areas requiring further development in long-read sequencing analysis.

Bibliography

[1] F. S. Collins and H. Varmus, "A New Initiative on Precision Medicine," en, *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, Feb. 2015, ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMp1500523. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMp1500523.

[2] F. Carrasco-Ramiro, R. Peiró-Pastor, and B. Aguado, "Human genomics projects and precision medicine," en, *Gene Therapy*, vol. 24, no. 9, pp. 551–561, Sep. 2017, Publisher: Nature Publishing Group, ISSN: 1476-5462. DOI: 10.1038/gt.2017.77. [Online]. Available: https://www.nature.com/articles/gt201777.

[3] C. M. Johannessen and J. S. Boehm, "Progress towards precision functional genomics in cancer," *Current Opinion in Systems Biology*, Regulatory and metabolic networks • Cancer and systemic diseases, vol. 2, pp. 74–83, Apr. 2017, ISSN: 2452-3100. DOI: 10.1016/j.coisb.2017.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452310017300410.

[4] *What Is Cancer? - NCI*, en, cgvArticle, Archive Location: nciglobal,ncienterprise, Sep. 2007. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[5] S. Turajlic, A. Sottoriva, T. Graham, and C. Swanton, "Resolving genetic heterogeneity in cancer," en, *Nature Reviews Genetics*, vol. 20, no. 7, pp. 404–416, Jul. 2019, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/s41576-019-0114-6. [Online]. Available: https://www.nature.com/articles/s41576-019-0114-6.

[6] J. K. Sicklick, S. Kato, R. Okamura, *et al.*, "Molecular profiling of cancer patients enables personalized combination therapy: The I-PREDICT study," en, *Nature Medicine*, vol. 25, no. 5, pp. 744–750, May 2019, Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: 10.1038/s41591-019-0407-5. [Online]. Available: https://www.nature.com/articles/s41591-019-0407-5.

[7] S. Y. Moorcraft, D. Gonzalez, and B. A. Walker, "Understanding next generation sequencing in oncology: A guide for oncologists," *Critical Reviews in Oncology/Hematology*, vol. 96, no. 3, pp. 463–474, Dec. 2015, ISSN: 1040-8428. DOI: 10.1016/j.critrevonc.2015.06.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1040842815001250.

[8] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, "Long-read human genome sequencing and its applications," en, *Nature Reviews Genetics*, vol. 21, no. 10, pp. 597–614, Oct. 2020, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/s41576-020-0236-x. [Online]. Available: https://www.nature.com/articles/s41576-020-0236-x.

[9] S. A. Jeon, J. L. Park, S.-J. Park, *et al.*, "Comparison between MGI and Illumina sequencing platforms for whole genome sequencing," en, *Genes & Genomics*, vol. 43, no. 7, pp. 713–724, Jul. 2021, ISSN: 2092-9293. DOI: 10.1007/s13258-021-01096-x. [Online]. Available: https://doi.org/10.1007/s13258-021-01096-x.

[10] J.-Y. Zhang, Y. Zhang, L. Wang, *et al.*, *A single-molecule nanopore sequencing platform*, en, Pages: 2024.08.19.608720 Section: New Results, Aug. 2024. DOI: 10.1101/2024.08.19.608720. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.08.19.608720v1.

[11] S. S. Ho, A. E. Urban, and R. E. Mills, "Structural variation in the sequencing era," en, *Nature Reviews Genetics*, vol. 21, no. 3, pp. 171–189, Mar. 2020, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/s41576-019-0180-9. [Online]. Available: https://www.nature.com/articles/s41576-019-0180-9.

[12] M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler, "Genetic variation and the de novo assembly of human genomes," en, *Nature Reviews Genetics*, vol. 16, no. 11, pp. 627–640, Nov. 2015, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/nrg3933. [Online]. Available: https://www.nature.com/articles/nrg3933.

[13] E. Espinosa, R. Bautista, R. Larrosa, and O. Plata, "Advancements in long-read genome sequencing technologies and algorithms," *Genomics*, vol. 116, no. 3, p. 110 842, May 2024, ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2024.110842. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888754324000636.

[14] N. J. Loman, J. Quick, and J. T. Simpson, "A complete bacterial genome assembled de novo using only nanopore sequencing data," en, *Nature Methods*, vol. 12, no. 8, pp. 733–735, Aug. 2015, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/nmeth.3444. [Online]. Available: https://www.nature.com/articles/nmeth.3444.

[15] S. Nurk, S. Koren, A. Rhie, *et al.*, "The complete sequence of a human genome," EN, *Science*, Apr. 2022, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.abj6987. [Online]. Available: https://www.science.org/doi/10.1126/science.abj6987.

[16] A. Rhie, S. Nurk, M. Cechova, *et al.*, "The complete sequence of a human Y chromosome," en, *Nature*, vol. 621, no. 7978, pp. 344–354, Sep. 2023, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-023-06457-y. [Online]. Available: https://www.nature.com/articles/s41586-023-06457-y.

[17] L. M. Zahn, "Filling the gaps," *Science*, vol. 376, no. 6588, pp. 42–43, Apr. 2022, Publisher: American Association for the Advancement of Science. DOI: 10.1126/science.abp8653. [Online]. Available: https://www.science.org/doi/10.1126/science.abp8653.

[18] *Vega benchtop system*, en-US. [Online]. Available: https://www.pacb.com/vega/.

[19] Oxford Nanopore Technologies, *Nanopore store: PromethION 2 Solo*. [Online]. Available: https://store.nanoporetech.com/eu/p2-solo.html.

[20] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, "Nanopore sequencing technology, bioinformatics and applications," en, *Nature Biotechnology*, vol. 39, no. 11, pp. 1348–1365, Nov. 2021, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-021-01108-x. [Online]. Available: https://www.nature.com/articles/s41587-021-01108-x.

[21] M. Kolmogorov, K. J. Billingsley, M. Mastoras, *et al.*, "Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation," en, *Nature Methods*, vol. 20, no. 10, pp. 1483–1492, Oct. 2023, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-023-01993-x. [Online]. Available: https://www.nature.com/articles/s41592-023-01993-x.

[22] Y. Sakamoto, S. Miyake, M. Oka, *et al.*, "Phasing analysis of lung cancer genomes using a long read sequencer," en, *Nature Communications*, vol. 13, no. 1, p. 3464, Jun. 2022, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-022-31133-6. [Online]. Available: https://www.nature.com/articles/s41467-022-31133-6.

[23] W. Schaal, A. Ameur, U. Olsson-Strömberg, M. Hermanson, L. Cavelier, and O. Spjuth, "Migrating to Long-Read Sequencing for Clinical Routine BCR-ABL1 TKI Resistance Mutation Screening," *Cancer Informatics*, vol. 21, p. 11 769 351 221 110 872, Jul. 2022, ISSN: 1176-9351. DOI: 10.1177/11769351221110872. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9290162/.

[24] Oxford Nanopore Technologies, *Unlocking comprehensive genome for large-scale projects*, Sep. 2024. [Online]. Available: https://www.youtube.com/watch?v=nRmm13IAFcg.

[25] H. Li and R. Durbin, "Genome assembly in the telomere-to-telomere era," en, *Nature Reviews Genetics*, vol. 25, no. 9, pp. 658–670, Sep. 2024, Publisher: Nature Publishing Group, ISSN: 1471-0064. DOI: 10.1038/s41576-024-00718-w. [Online]. Available: https://www.nature.com/articles/s41576-024-00718-w.

[26] S. Koren, Z. Bao, A. Guarracino, *et al.*, *Gapless assembly of complete human and plant chromosomes using only nanopore sequencing*, en, Pages: 2024.03.15.585294 Section: New Results, Mar. 2024. DOI: 10.1101/2024.03.15.585294. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.03.15.585294v2.

[27] Y. Li, N. D. Roberts, J. A. Wala, *et al.*, "Patterns of somatic structural variation in human cancer genomes," en, *Nature*, vol. 578, no. 7793, pp. 112–121, Feb. 2020, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1913-9. [Online]. Available: https://www.nature.com/articles/s41586-019-1913-9.

[28] F. Menghi, F. P. Barthel, V. Yadav, *et al.*, "The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations," *Cancer Cell*, vol. 34, no. 2, 197–210.e5, Aug. 2018, ISSN: 1535-6108. DOI: 10.1016/j.ccell.2018.06.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1535610818302654.

[29] D. L. Cameron, L. Di Stefano, and A. T. Papenfuss, "Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software," en, *Nature Communications*, vol. 10, no. 1, p. 3240, Jul. 2019, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-019-11146-4. [Online]. Available: https://www.nature.com/articles/s41467-019-11146-4.

[30] H. J. Abel, D. E. Larson, A. A. Regier, *et al.*, "Mapping and characterization of structural variation in 17,795 human genomes," en, *Nature*, vol. 583, no. 7814, pp. 83–89, Jul. 2020, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2371-0. [Online]. Available: https://www.nature.com/articles/s41586-020-2371-0.

[31] T. Rausch, R. Snajder, A. Leger, *et al.*, "Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures," *Cell Genomics*, vol. 3, no. 4, p. 100 281, Apr. 2023, ISSN: 2666-979X. DOI: 10.1016/j.xgen.2023.100281. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666979X23000411.

[32] J. E. Valle-Inclan, S. D. Noon, K. Trevers, *et al.*, "Ongoing chromothripsis underpins osteosarcoma genome complexity and clonal evolution," English, *Cell*, vol. 0, no. 0, Jan. 2025, Publisher: Elsevier, ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2024.12.005. [Online]. Available: https://www.cell.com/cell/abstract/S0092-8674(24)01418-1.

[33] W. M. Kuehl and P. L. Bergsagel, "Multiple myeloma: Evolving genetic events and host interactions," en, *Nature Reviews Cancer*, vol. 2, no. 3, pp. 175–187, Mar. 2002, Publisher: Nature Publishing Group, ISSN: 1474-1768. DOI: 10.1038/nrc746. [Online]. Available: https://www.nature.com/articles/nrc746.

[34] S. E. Kurtin, "Relapsed or Relapsed/Refractory Multiple Myeloma," en, no. 6, 2013.

[35] *Usage - CNIO Bioinformatics Unit documentation*. [Online]. Available: https://cnio-bu.github.io/hpc/usage/.

[36] R. Wick, "Badread: Simulation of error-prone long reads," *Journal of Open Source Software*, vol. 4, no. 36, p. 1316, Apr. 2019, ISSN: 2475-9066. DOI: 10.21105/joss.01316. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01316.

[37] A. Y. Aksenova, A. S. Zhuk, A. G. Lada, *et al.*, "Genome Instability in Multiple Myeloma: Facts and Factors," en, *Cancers*, vol. 13, no. 23, p. 5949, Jan. 2021, Number: 23 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-

6694. DOI: 10.3390/cancers13235949. [Online]. Available: https://www.mdpi.com/2072-6694/13/23/5949.

[38] I. Dagogo-Jack and A. T. Shaw, "Tumour heterogeneity and resistance to cancer therapies," en, *Nature Reviews Clinical Oncology*, vol. 15, no. 2, pp. 81–94, Feb. 2018, Publisher: Nature Publishing Group, ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2017.166. [Online]. Available: https://www.nature.com/articles/nrclinonc.2017.166.

[39] *Ligation sequencing gDNA V14 — human sample (N50 10 kb) on PromethION (SQK-LSK114)*, en-GB, Oct. 2022. [Online]. Available: https://nanoporetech.com/document/ligation-sequencing-gdna-v14-10-kb-human-sample-on-promethion-sqk-lsk114.

[40] Oxford Nanopore Technologies, *EPI2ME Desktop*, en. [Online]. Available: https://labs.epi2me.io/about/.

[41] *Releases · fritzsedlazeck/Sniffles/releases/tag/v2.5*, en. [Online]. Available: https://github.com/fritzsedlazeck/Sniffles/releases/tag/v2.5.

[42] *Error: Skia GrGLInterface was not valid · Issue #38 · kcleal/gw*. [Online]. Available: https://github.com/kcleal/gw/issues/38.

[43] *Display IDs instead of gene names when loading .gff3 and .gtf files · Issue #43 · kcleal/gw*. [Online]. Available: https://github.com/kcleal/gw/issues/43.

[44] *Saving to pdf from the main program · Issue #59 · kcleal/gw*. [Online]. Available: https://github.com/kcleal/gw/issues/59.

[45] *Add parameter logging to VISOR LASeR output · Issue #42 · davidebolo1993/VISOR*. [Online]. Available: https://github.com/davidebolo1993/VISOR/issues/42.

[46] A. C. English, V. K. Menon, R. A. Gibbs, G. A. Metcalf, and F. J. Sedlazeck, "Truvari: Refined structural variant comparison preserves allelic diversity," *Genome Biology*, vol. 23, no. 1, p. 271, Dec. 2022, ISSN: 1474-760X. DOI: 10.1186/s13059-022-02840-6. [Online]. Available: https://doi.org/10.1186/s13059-022-02840-6.

[47] J. Köster and S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts480.

[48] D. Bolognini, A. Sanders, J. O. Korbel, A. Magi, V. Benes, and T. Rausch, "VISOR: A versatile haplotype-aware structural variant simulator for short- and long-read sequencing," *Bioinformatics*, vol. 36, no. 4, pp. 1267–1269, Feb. 2020, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz719. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz719.

[49] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty191.

[50] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of SAMtools and BCFtools," *GigaScience*, vol. 10, no. 2, giab008, Feb. 2021, ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. [Online]. Available: https://doi.org/10.1093/gigascience/giab008.

[51] H. Elrick, C. M. Sauer, J. E. Valle-Inclan, *et al.*, *SAVANA: Reliable analysis of somatic structural variants and copy number aberrations in clinical samples using long-read sequencing*, en, Pages: 2024.07.25.604944 Section: New Results, Jul. 2024. DOI: 10.1101/2024.07.25.604944. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.07.25.604944v1.

[52] A. Keskus, A. Bryant, T. Ahmad, *et al.*, *Severus: Accurate detection and characterization of somatic structural variation in tumor genomes using long reads*, en, Pages: 2024.03.22.24304756, Mar. 2024. DOI: 10.1101/2024.03.22.24304756. [Online]. Available: https://www.medrxiv.org/content/10.1101/2024.03.22.24304756v1.

[53] M. Smolka, L. F. Paulin, C. M. Grochowski, *et al.*, "Detection of mosaic and population-level structural variants with Sniffles2," en, *Nature Biotechnology*, vol. 42, no. 10, pp. 1571–1580, Oct. 2024, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-023-02024-y. [Online]. Available: https://www.nature.com/articles/s41587-023-02024-y.

[54] S. Wang, J. Lin, P. Jia, *et al.*, "De novo and somatic structural variant discovery with SVision-pro," en, *Nature Biotechnology*, pp. 1–5, Mar. 2024, Publisher: Nature Publishing Group, ISSN: 1546-1696. DOI: 10.1038/s41587-024-02190-7. [Online]. Available: https://www.nature.com/articles/s41587-024-02190-7.

[55] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq033. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq033.

[56] K. Cleal, A. Kearsey, and D. M. Baird, *GW: Ultra-fast chromosome-scale visualisation of genomics data*, en, Pages: 2024.07.26.605272 Section: New Results, Sep. 2024. DOI: 10.1101/2024.07.26.605272. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.07.26.605272v5.

[57] G. Perez, G. P. Barber, A. Benet-Pages, *et al.*, "The UCSC Genome Browser database: 2025 update," eng, *Nucleic Acids Research*, vol. 53, no. D1, pp. D1243–D1249, Jan. 2025, ISSN: 1362-4962. DOI: 10.1093/nar/gkae974.

[58] Oxford Nanopore Technologies, *R10.3: The newest nanopore for high accuracy nanopore sequencing*, ja-JP, Section: Technology, Jan. 2020. [Online]. Available: https://

nanoporetech.com/ja/news/news-r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store.

A
Appendix

Table A.1: Software tools used in this work.

Name	Source	Reference
Visual Studio Code v1.93.1	https://github.com/microsoft/vscode	–
Snakemake v8.29.6	https://github.com/snakemake/snakemake	[47]
Miniforge v24.7.1	https://github.com/conda-forge/miniforge	–
VISOR v1.1.2.1	https://github.com/davidebolo1993/VISOR	[48]
minimap2 v2.28	https://github.com/lh3/minimap2	[49]
SAMtools v1.21	https://github.com/samtools/samtools	[50]
SAVANA v1.2.2	https://github.com/cortes-ciriano-lab/savana	[51]
Severus v1.2	https://github.com/KolmogorovLab/Severus	[52]
Sniffles2 v2.4	https://github.com/fritzsedlazeck/Sniffles	[53]
SVision-Pro v2.0	https://github.com/sonGBowang125/SVision-pro	[54]
BEDtools v2.31.1	https://github.com/arq5x/bedtools2	[55]
GW v1.1.1	https://github.com/kcleal/gw	[56]
UCSC Genome Browser v2024	https://genome.ucsc.edu/index.html	[57]
Rstudio v2024.04.2	https://github.com/rstudio/rstudio	–

Table A.2: Size in gigabytes of long-read sequencing files generated using VISOR toolkit at different sequencing depths. The files include BAM, their indexes (BAI), and unaligned reads in FASTQ files.

Coverage	BAM	BAI	FASTQ
30x_normal	112.42	0.0347	173.17
30x	113.20	0.0351	174.02
30x	113.19	0.0350	174.02
30x	113.22	0.0350	174.02
50x_normal	187.01	0.0551	288.58
50x	188.71	0.0556	289.99
50x	188.85	0.0556	289.99
50x	188.70	0.0556	289.99
100x_normal	374.24	0.1060	577.11
100x	377.85	0.1071	579.93
100x	377.86	0.1071	579.92
100x	377.72	0.1071	579.93
200x_normal	748.34	0.2079	1157.12
200x	754.99	0.2099	1157.12
200x	754.89	0.2099	1157.12
200x	755.35	0.2099	1157.12
Total	5726.54	1.6266	8219.22

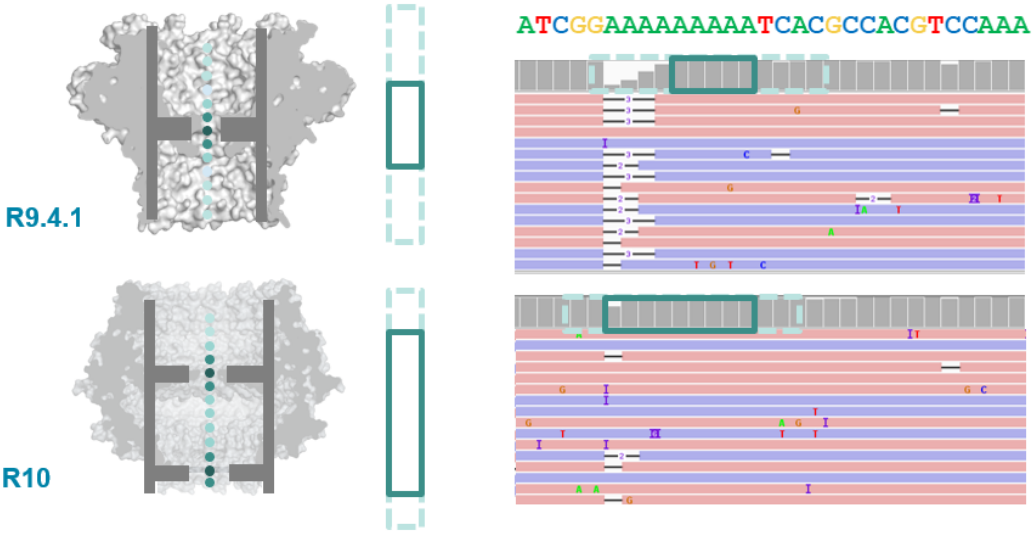

Figure A.1: The R10 nanopore design introduces key improvements over its R9 predecessor: a longer barrel and dual reader head architecture. These structural enhancements enable improved resolution of homopolymeric regions such as VNRTs, significantly increasing the consensus accuracy of nanopore sequencing data. Reproduced from [58].

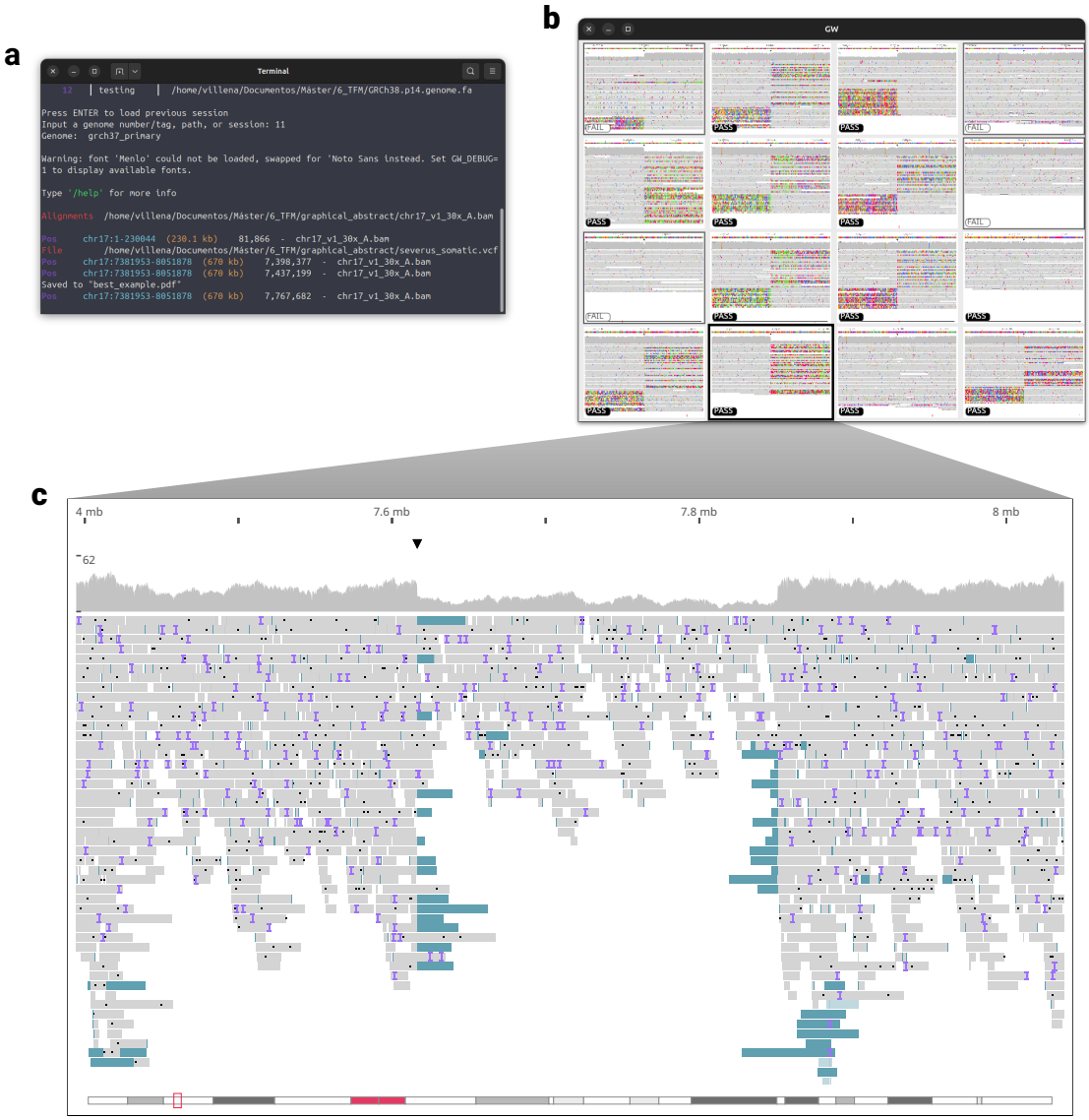

Figure A.2: Alignment-based SV Validation using GW. The tool is launched through the terminal, initially prompting for reference genome selection (**a**) and opening an empty graphical window. Loading BAM and SV-containing VCF files by dragging them into the graphical window generates a grid of thumbnail images corresponding to VCF-indicated breakpoints (**b**). Hovering over an image displays breakpoint coordinates in the terminal, while clicking opens a detailed alignment view with chromosome-wide navigation capabilities (**c**). Users can return to the image grid and toggle True/False buttons in the lower-left corners to track SV validation status, which can be exported as a list for further analysis.