Structural variant detection using targeted short read sequencing data from cancer patients

Manuela Cristina Dias Pinheiro
Master in Bioinformatics and Computational Biology
Department of Computer Science | Department of Biology
Faculty of Sciences of the University of Porto
2025

Structural variant detection using targeted short read sequencing data from cancer patients

Manuela Cristina Dias Pinheiro

Dissertation carried out as part of the Master in Bioinformatics and Computational Biology

Department of Computer Science | Department of Biology

2025

Supervisor
Manuel António Rodrigues Teixeira, MD, PhD
Cancer Genetics Group, IPO-Porto Research Center/RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO-Porto)/Porto Comprehensive Cancer Center

Guest Full Professor. School of Medicine and Biomedical Sciences (ICBAS), University of Porto

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

IPOPORTO
INSTITUTO PORTUGUÊS DE ONCOLOGIA DO PORTO FG, EPE

Acknowledgements

I would like to thank my supervisor, Professor Doutor Manuel Teixeira, Director of the Laboratory Genetics Department of IPO-Porto and PI of the Cancer Genetics Group, for the opportunity to develop this work in his research group, and for the knowledge, patience and help during the development of this thesis. I would like to thank Ana Peixoto, for accompanying me closely in every step of my work, for all the effort and all the suggestions that helped me achieve the thesis presented here and especially for the long-lasting friendship. I also would like to thank my friends Ana, Carla, Catarina, Clara Joana e Maria José for the support and friendship given over the years. I would like to thank all my colleagues from the Laboratory Genetics Department of IPO-Porto which contributed to this thesis. And finally, a special thank you to my parents, brothers and remaining family.

Resumo

O cancro é caracterizado por alterações genómicas que podem ativar oncogenes, inativar genes supressores tumorais ou modificar elementos reguladores, contribuindo para a iniciação tumoral, progressão da doença e resistência à terapêutica. Estas alterações genómicas incluem variantes de um único nucleótido, pequenas inserções/deleções e variantes estruturais. Estudos de sequenciação do genoma humano demonstraram que variantes estruturais constituem a classe mais comum de alterações *drivers* em tumores sólidos metastáticos e constituem cerca de 30% das alterações clinicamente acionáveis. Estas alterações consistem em rearranjos genómicos, superiores a 50 pares de bases, e incluem maioritariamente deleções, inserções, duplicações, inversões e translocações. Apesar das inovações tecnológicas, a deteção fiável de variantes estruturais em tumores continua a ser um desafio. Embora os avanços na sequenciação massiva paralela e a melhoria das ferramentas bioinformáticas tenham aumentado as taxas de deteção, nenhuma ferramenta isolada apresenta simultaneamente sensibilidade e especificidade elevadas. Assim, a integração de múltiplas abordagens computacionais é recomendada para uma deteção de elevada confiança. Este projeto teve como objetivo avaliar ferramentas bioinformáticas para a deteção de variantes estruturais em tumores, com especial enfoque em alterações clinicamente acionáveis.

Para tal, o presente estudo comparou o desempenho de diversas ferramentas na deteção de variantes estruturais em amostras de DNA e RNA, extraídos de tumores sólidos incluídos em parafina, após sequenciação de nova geração usando painéis de genes. Ferramentas de deteção de transcritos de fusão após sequenciação do RNA (como Arriba e STAR-Fusion) demonstraram elevada sensibilidade na sua deteção. Ferramentas que analisam DNA (como DELLY e LUMPY) também apresentaram um bom desempenho na deteção de fusões. E neste estudo também foram testadas ferramentas para a deteção da variante fundadora portuguesa *BRCA2* c.156_157insAlu, sendo que o *script* em Python desenvolvido neste estudo apresentou 100% de concordância, evidenciando a eficácia de abordagens direcionadas quando a localização da variante é previamente conhecida.

Em conclusão, os nossos resultados sugerem que nenhuma ferramenta isolada é suficiente para uma deteção abrangente de variantes estruturais, em dados de sequenciação usando painéis de genes. A deteção precisa de variantes estruturais requer uma abordagem combinada, integrando múltiplos algoritmos adaptados ao tipo

de variante, complementada por validação ortogonal sempre que necessário. A utilização de tecnologias de sequenciação de leitura longa, o uso de inteligência artificial e a otimização do design dos painéis poderá melhorar a deteção de variantes estruturais na área da oncologia.

Palavras-chave: Cancro, variantes estruturais, variantes acionáveis, deteção de variantes estruturais, ferramentas computacionais

Abstract

Cancer is characterized by genomic alterations that can activate oncogenes, disrupt tumour suppressor genes, or change regulatory elements, contributing to tumour initiation, progression, and therapy resistance. These genomic alterations can include single nucleotide variants, small insertions/deletions, and structural variants. Whole genome sequencing studies have shown that structural variants are the most common driver alterations in metastatic solid tumours, representing about 30% of all clinical actionable alterations. Structural variants are large-scale rearrangements of more than 50 base pairs of size and include deletions, insertions, duplications, inversions, and translocations. Despite technological innovations, confident structural variant detection in cancer genomes remains challenging. Although advances in massive parallel sequencing and improvements in bioinformatics pipelines have increased detection rates, no individual tool provides comprehensive sensitivity and specificity and integrating multiple computational approaches is recommended to achieve high-confidence detection. This project aims to evaluate bioinformatics tools for the detection of structural variants in cancer patients, with a focus on clinically actionable alterations.

The present study compared the performance of several tools for the detection of structural variants in DNA and RNA samples extracted from formalin-fixed, paraffin-embedded solid tumours following next-generation sequencing using gene panels. RNA sequencing–based fusion transcript detection tools (such as Arriba and STAR-Fusion) demonstrated high sensitivity for fusion detection. DNA-based tools (such as DELLY and LUMPY) also showed good performance in identifying gene fusions. In addition, tools for the detection of the Portuguese founder variant *BRCA2* c.156_157insAlu were evaluated, and the Python script developed in this study achieved 100% concordance, highlighting the effectiveness of targeted approaches when the variant location is known *a priori*.

To conclude, our findings suggest that no single caller is sufficient for comprehensive detection in short-read panel-based settings. Accurate detection of structural variants requires a combined approach, integrating multiple algorithms tailored to variant type, with manual review and orthogonal validation when needed. Future improvements may include the use of long-read technologies, machine learning-based filtering, and enhanced panel designs to improve structural variant detection in routine clinical practice.

Keywords: Cancer, structural variants, actionable variants, structural variant callers, computational tools

Table of Contents

1. Introduction

1.1. Cancer genetics

Cancer is a disease that arises from the progressive accumulation of genetic and epigenetic alterations in the genome. These alterations can lead to the disruption of normal cellular processes such as proliferation, apoptosis, differentiation, and DNA repair, ultimately conferring selective growth advantages on altered cells (Vogelstein et al, 2013). Cancer genomes are characterized by a spectrum of genetic alterations, ranging from single-nucleotide variants (SNVs) to larger structural variants (SVs) such as deletions, duplications, inversions, translocations, and complex rearrangements (Priestley et al, 2019).

Cytogenetic studies were the first to identify the presence of SVs in cancer cells. One of the earliest examples is the Philadelphia chromosome that results from a reciprocal translocation between chromosomes 9 and 22 in chronic myeloid leukaemia. This translocation generates the *BCR*::*ABL1* fusion gene, which encodes a constitutively active tyrosine kinase that drives malignant transformation (Rowley, 1973; Randolph, 2005). This discovery was pivotal in demonstrating that cancer-associated chromosomal changes are not random byproducts of genomic instability but causal events with clear mechanistic and clinical significance. Since then, recurrent SVs have been identified across different tumour types, including the *EML4*::*ALK* fusion in non-small cell lung cancer (Soda et al, 2007; Lin et al, 2018), *TMPRSS2*::*ERG* in prostate cancer (Ahlers et al, 2006), and *ERBB2* amplification in breast cancer (Moasser, 2007).

The advent of massive parallel sequencing, also known as next-generation sequencing (NGS), marked a transformative moment in cancer genomics. Unlike Sanger sequencing, which is limited in throughput and scope, NGS enables the simultaneous sequencing of millions to billions of DNA fragments, thereby allowing comprehensive analysis of entire cancer genomes at base-pair resolution (Shendure et al, 2017). This has enabled researchers to move beyond the detection of single-gene alterations and toward the systematic cataloguing of entire mutational landscapes. Large-scale sequencing projects such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al, 2013), the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium et al, 2010), and more recently the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) have systematically mapped the genomic landscapes of thousands of tumours. These studies revealed that while SNVs

are the most frequent class of somatic alterations, many of the most impactful driver events in cancer are structural in nature, such as chromosomal rearrangements, focal amplifications, and complex events like chromothripsis and chromoplexy (Priestley et al, 2019, Shoshani et al, 2021). Structural alterations were shown to be nearly ubiquitous across tumour types, and in many cancers, they constitute the primary mechanism of oncogenic activation or tumour suppressor loss.

1.2. SVs in cancer

SVs, defined as genomic alterations exceeding 50 base pairs in size, are highly relevant in cancer biology, driving oncogenesis through mechanisms such as oncogene activation, tumour suppressor disruption, and gene fusion formation (Cosenza et al, 2022). Unlike SNVs or small insertions and deletions, which generally affect coding sequences in a more predictable and localized manner, SVs can reshape entire chromosomal landscapes and disrupt multiple genes in a single event. This makes them particularly potent drivers of oncogenesis and highly relevant to both biological understanding and clinical actionability.

These variants encompass a broad spectrum of genomic alterations (Figure 1). Deletions involve the loss of DNA segments, which can remove essential regulatory or coding regions and often result in gene disruption or fusion. Duplications consist of the copying of genomic regions, either in tandem or interspersed, and typically lead to dosage imbalance. Amplifications are characterized by multiple additional copies of a genomic segment, often causing overexpression of genes located within the amplified region. Insertions involve the addition of DNA fragments into new locations, which may disrupt existing sequences or alter regulatory regions; these can include novel sequence insertions or transposable element–mediated insertions, such as Alu elements, which are mobilized through homology-driven or retrotransposition mechanisms. Inversions occur when a segment of DNA is reversed in orientation; depending on breakpoint positions, they may interrupt genes, reconfigure regulatory domains, or generate fusion events. Translocations involve the exchange or transfer of DNA between non-homologous chromosomes, which can disrupt genes or lead to fusions. Finally, complex rearrangements, including chromothripsis and chromoplexy, comprise multiple clustered breakpoints and can reorganize chromosomal architecture in a single catastrophic event (Shoshani et al, 2021).

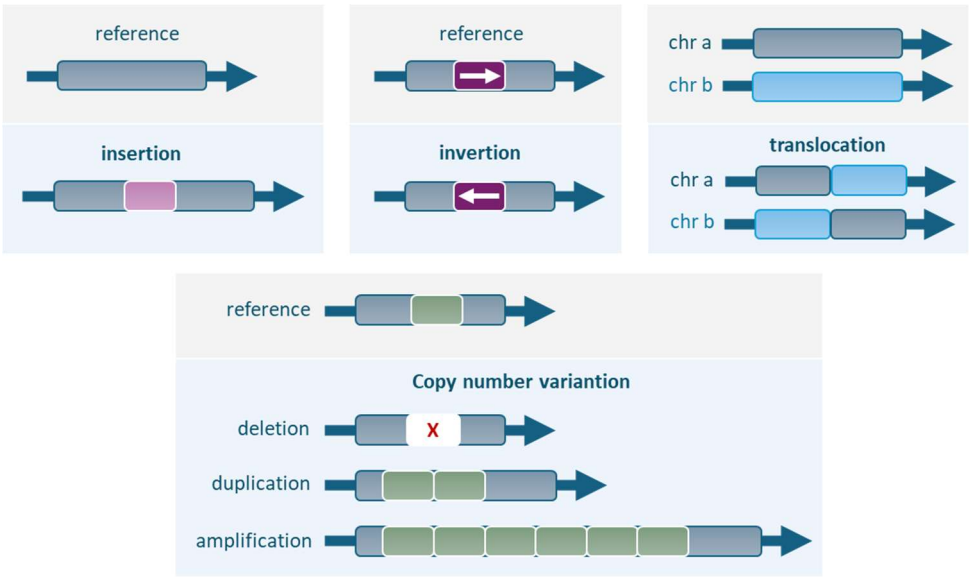

Figure 1 - Most common types of SVs. These alterations include insertions, inversions and translocations besides copy number variations like deletion, duplication (in the example is a tandem duplication but can also be an interspersed duplication) and amplification.

Whole genome sequencing studies in tumours datasets have shown that somatic SVs are the most common driver alterations in metastatic solid tumours, representing about 30% of all clinical actionable alterations (Priestley et al, 2019). Mechanistically, SVs contribute to oncogenesis through multiple ways. They may activate proto-oncogenes by juxtaposing them with strong enhancers or promoters, a phenomenon known as enhancer hijacking, as has been described in medulloblastoma where rearrangements drive *MYC* overexpression (Northcott et al, 2014). Gene amplifications represent another common oncogenic mechanism, exemplified by *ERBB2* amplification in breast cancer, which confer proliferative advantages and predict response to targeted therapy (Moasser, 2007; Dumbrava et al, 2019). Conversely, SVs often inactivate tumour suppressor genes through focal deletions or disruptive rearrangements, leading to inactivation of genes such as *TP53*, *PTEN*, or *CDKN2A* (Beroukhim et al, 2010). Another functional outcome of SVs is the generation of fusion genes, which can encode novel oncogenic proteins or change the regulation of transcriptional pathways. Examples include fusions involving *NTRK1*, *NTRK2*, and *NTRK3* that occur across diverse tumour types and although rare are highly actionable, with response rates often exceeding 70%

when treated with TRK inhibitors such as larotrectinib and entrectinib (Cocco et al, 2018; Wang et al, 2022).

Besides somatic SVs, germline SVs are also reported. These genomic alterations are inherited and present in all cells. While somatic SVs contribute directly to tumorigenesis, germline SVs can predispose individuals to cancer by disrupting tumour suppressors. A hereditary pan-cancer gene panel survey of 376,159 individuals revealed that large genomic rearrangements, a subset of SVs, accounted for 7.2% of all pathogenic variants detected, with the largest proportion identified in *BRCA1* (Pócza et al, 2021). An example is the *BRCA2* c.156_157insAlu insertion, a germline Alu element insertion that disrupts the *BRCA2* coding sequence and predisposes carriers to hereditary breast and ovarian cancer (HBOC) syndrome. This alteration is a Portuguese founder variant that accounts for approximately 25% of all deleterious *BRCA* variants in northern and central Portugal (Peixoto et al, 2009). Chromosomal instability is a hallmark of cancer and refers to an increased rate of chromosomal alterations, including gains, losses, and complex structural rearrangements. This genomic instability contributes not only to tumour progression but can also obscure the detection of underlying germline rearrangements (Vishwakarma et al, 2020). In this context, differentiating somatic structural events from inherited ones remains a major analytical challenge, particularly in clinical settings (Terraf et al, 2022).

1.3. SV detection by massive parallel sequencing

Detection of SVs depends critically on the choice of sequencing technology, each offering distinct advantages and limitations. Short-read sequencing platforms have high per-base accuracy, throughput, and relatively low cost, and they are widely applied in clinical and research settings (Isaic et al, 2025). Although paired-end short read sequencing faces limitations in resolving repetitive sequences, segmental duplications, GC-rich or GC-poor regions, and large insertions or complex rearrangements, that often result in ambiguous alignments or missed breakpoints it has been widely used in:

- Whole genome sequencing (WGS), providing unbiased coverage of coding and noncoding regions and enabling the detection of structural rearrangements, copy number alterations (CNVs), and complex SVs across the entire genome.
- Whole-exome sequencing (WES) that focuses on coding regions, providing higher depth at lower cost but missing intronic, intergenic, and regulatory SVs.

- Targeted panels, either hybrid-capture or amplicon-based, that achieve ultra-deep coverage (>500×), facilitating the detection of low-allele-frequency variants in clinically relevant genes. Hybrid-capture panels allow flexible target design but require more input of DNA/RNA and are susceptible to capture bias, whereas amplicon panels offer rapid turnaround and lower input requirements but are restricted to predefined regions, are more prone to errors and can miss novel or complex rearrangements (Cameron et al, 2019; Cosenza et al, 2022; Satam et al, 2023).

A recent pan-cancer study showed that most large clonal SVs (>10 kilobases) are fully resolved by short-read sequencing in 87% of the human genome where copy number can be measured reliably, but many SVs in repetitive or less well-covered regions remain unresolved by short reads (Choo et al, 2023). Long-read sequencing addresses many of the limitations because their reads can span long repetitive sequences, structural breakpoints, and provide better assembly of complex regions (Aydin et al, 2025). However, long reads have higher per-base error rates, higher cost per base, limitations in throughput, and sometimes lower depth, which affects detection of low VAF events (Olivucci et al, 2024). Also, for many clinical specimens (formalin-fixed paraffin embedded, low input DNA, degraded DNA) obtaining long-fragment DNA is technically challenging (Cappello et al, 2022).

The choice between DNA- and RNA-based sequencing depends on the SV type and the intended clinical or biological interpretation. DNA sequencing provides high-resolution mapping of SV breakpoints and structural architecture, but it cannot distinguish expressed from nonfunctional rearrangements (Benayed et al, 2019; Cohen et al, 2020; Yang et al, 2021; Li et al, 2022; Satam et al, 2023). RNA sequencing, conversely, directly identifies expressed fusion transcripts, alternative splicing events, and transcript-level consequences, providing functional insights into gene regulation and potential oncogenic activity (Benayed et al, 2019; Cohen et al, 2020; Yang et al, 2021; Satam et al, 2023). RNA-based detection is, however, prone to false positives arising from transcriptional noise, sequencing artifacts, and aberrant splicing (Benayed et al, 2019; Cohen et al, 2020; Yang et al, 2021; Li et al, 2022). Integration of DNA and RNA sequencing has emerged as a robust strategy for detecting clinically actionable SVs, combining structural resolution with transcriptional confirmation to improve prioritization of variants and reduce false discovery rates (Cohen et al, 2020; Yang et al, 2021).

1.4. Bioinformatic tools for the detection of SVs

Accurate identification of SVs relies on specialized computational tools that interpret the sequencing data. A common approach for detecting SVs using NGS data involves identifying mapping discrepancies during sequence alignment to the reference genome (Figure 2). SV callers are generally categorized into:

- Read depth analysis, which tracks deviations in coverage across the genome or target regions to infer deletions, duplications, and amplifications. Depth signals are strongest for events of sufficient size and in regions with uniform mappability; they are weaker for small-scale SVs, balanced rearrangements, or in low-coverage or low-purity samples.
- Discordant paired-end reads, where read pairs map further apart than the expected insert size, invert orientation, map to different chromosomes, or otherwise deviate from the normal expected pair configuration. These suggest potential breakpoints but require careful filtering (e.g., artifacts, mapping to repeats).
- Split-reads, reads that span a breakpoint, with one part of the read mapping to one locus, the other to another. Split-reads give high precision in localizing breakpoints, especially for SVs of moderate size; but they require enough read length and good mapping quality.
- Local assembly (or de novo assembly) of reads around candidate breakpoint regions: reconstructing contiguous sequence can resolve complex breakpoints or rearrangements in regions hard to map; group reads may span repetitive elements that break mapping for short reads.
- Probabilistic models and machine learning/deep learning techniques: these are increasingly used for SV detection, for merging evidence from diverse signatures, estimating confidence, distinguishing true from false positives, and tuning sensitivity/specificity trade-offs (Pirooznia et al, 2015; Gong et al, 2021; van Belzen et al, 2021).

For DNA-based SV discovery, widely used callers include:

- Manta - hybrid approach combining paired-end, split-read, and local assembly, with strong breakpoint resolution for somatic SVs (Chen et al, 2016).
- LUMPY - probabilistic framework integrating multiple signal types namely paired-end, split-read and read-depth (Layer et al, 2014).

- DELLY - leverages paired-end and split-read evidence, supporting deletions, duplications, inversions, and translocations (Rausch et al, 2012).
- CNVkit - optimized for CNV detection from WES or targeted panels, frequently used in clinical diagnostics (Talevich et al, 2016).
- Savvy - developed specifically for high-sensitivity detection of CNVs and SVs in targeted sequencing assays (Laver et al, 2022).

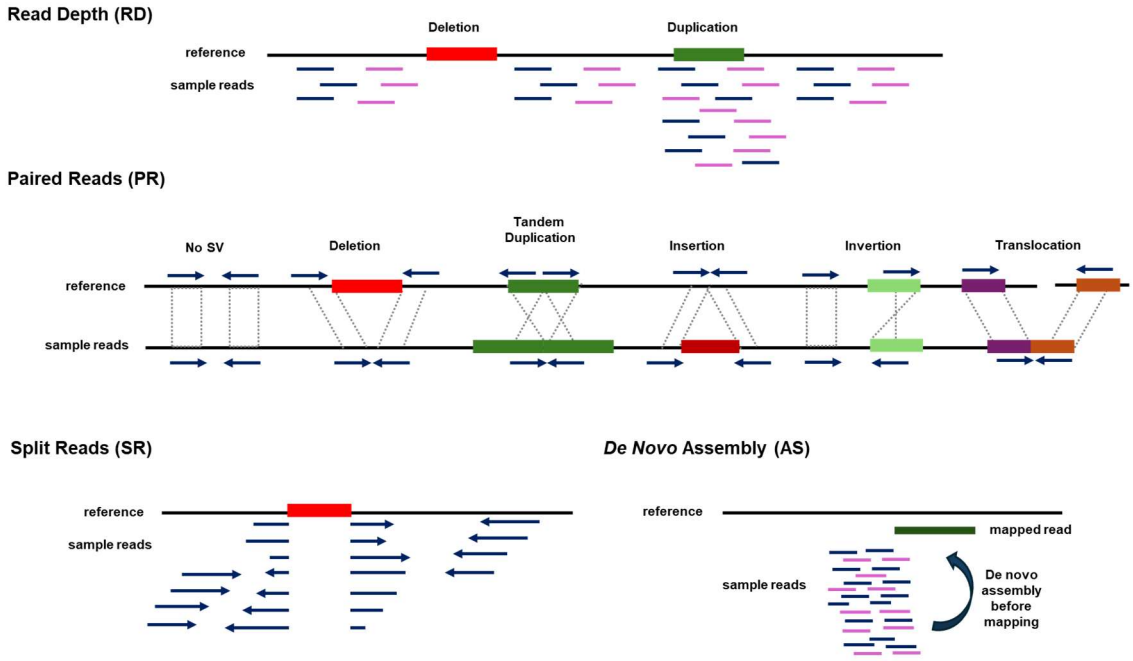

Figure 2 - Four main methods for detecting SVs with sequencing data: read depth (RD), paired-reads (PR), split-read (SR), and De Novo Assembly based (AS) method.

Certain classes of SVs require specialized tools. Retrotransposon insertions, particularly those mediated by Alu or LINE-1 elements, are poorly captured by standard SV callers. Dedicated algorithms such as RetroSeq (Keane et al, 2013) are designed to detect mobile element insertions from short-read data by exploiting discordant read signatures and split-read mappings at insertion breakpoints.

For RNA-based SV discovery, the focus is on detecting expressed fusions and aberrant splicing events and examples of SV callers are:

- STAR-Fusion - based on STAR aligner junction detection, widely used in both research and clinical fusion detection, integrating multiple signal types such as paired-end and split-reads information (Haas et al, 2019).
- Arriba - optimized for sensitivity and specificity in cancer RNA-based sequencing data, with advanced filtering to reduce false positives and considers paired-end and split-reads information (Uhrig et al, 2021).

Benchmarking studies show that choice of aligner and variant caller significantly affects sensitivity, specificity, recall, and precision across variant types and size ranges (Cameron et al, 2019; Zhuang et al, 2024). Furthermore, several studies have highlighted that combining aligners with callers (somatic-aware or not) and integrating multiple evidence types substantially increases recall and precision for SV detection (Cameron et al, 2019; Zhuang et al, 2024). Machine and deep learning frameworks are also emerging to enhance variant discrimination and confidence scoring, particularly in heterogeneous tumour samples or in low-coverage regions. Nonetheless, combining signals tends to improve detection but also complicates error modelling and false positives (spuriously called SVs) and false negatives (missed SVs) arise often (Popic et al, 2023; Xia et al, 2024).

1.5. Challenges in detecting SVs in tumour samples

The detection of SVs in tumour samples presents significant challenges due to the inherent complexity of tumour genomes and the technical limitations of current methodologies (van Belzen et al, 2021; Liu et al, 2015). Accurately identifying low allele frequency, complex and rare rearrangements in heterogeneous tumour samples can be particularly difficult (van Belzen et al, 2021; Liu et al, 2015). Tumour genomes are characterized by unique features such as polyploidy, subclonal genetic diversity, and the presence of clustered breakpoints, all of which complicate the detection of SVs (van Belzen et al, 2021; Liu et al, 2015). Additionally, the biological intricacy of cancer-associated rearrangements, including their diverse mechanisms and functional consequences, further complicates the interpretation and classification of these variants (Tubio, 2015). This complexity underscores the difficulty of identifying biologically relevant and actionable alterations, highlighting the need for computational tools and integrative approaches in SV detection and analysis.

Distinguishing somatic from germline SVs can also be a major problem, especially in settings where matched normal tissue is not available (van Belzen et al, 2021). One

approach is comparing tumour SVs to reference databases of germline variation which can remove common inherited events but fails to account for rare germline SVs, that represent a substantial proportion of human structural diversity (Yi et al, 2023; Li et al, 2024). This creates the risk of misclassifying germline variants as somatic, or vice versa, with important clinical consequences for patient counselling and therapy selection. Paired tumour-normal sequencing should be performed for accurate somatic SV detection, however in the diagnostic setting several constraints such as lack of normal sample and higher costs often preclude its use, making the development of improved computational and statistical models essential (Yi et al, 2023; Li et al, 2024).

Furthermore, even when SVs are robustly detected, interpreting their biological or clinical significance can be difficult. Not all SVs are driver events, a detected gene fusion may not be expressed, or the fused transcripts may not have oncogenic activity (Johansson et al, 2019; Cheong et al, 2024). Similarly, gene amplifications may not result in increased expression or protein activity (Vazquez-Mena et al, 2012). Moreover, depending on tumour type and context, SVs in noncoding regions (e.g., introns, regulatory elements) may have uncertain impact (Lange et al, 2021). Clinical application further requires that the SV be actionable, as only certain alterations have known therapies or clinical trials eligibility (Priestley et al, 2019; van Belzen et al, 2021). Much of the analysis, interpretation and classification of these rearrangements still relies heavily on manual curation, often requiring expertise and training to piece together the data from SV callers, data visualization tools such as Integrative Genomics Viewer (IGV), SVs databases and literature research. Orthogonal validation, including immunohistochemistry, fluorescence in situ hybridization (FISH), or Sanger sequencing, can also be needed to confirm SVs (Sepúlveda-Hermosilla et al, 2021).

Concluding, multi-platform integration, robust bioinformatic pipelines, and curated reference databases enhance diagnostic accuracy. Future directions in clinical SV detection include the development of standardized benchmarking datasets, machine learning–enhanced pipelines, long-read–based clinical assays, and broader multi-omics integration to ensure comprehensive, actionable SV detection for precision oncology.

2. Aim

Given the prevalence and clinical impact of SVs, the biological complexity and detection challenges, the variety of sequencing and computational methods, there is a clear need for improved strategies to detect actionable SVs in targeted sequencing contexts. This project aims to evaluate bioinformatics tools for the accurate detection of SVs in tumour samples, focusing on the detection of actionable alterations. By testing and integrating multiple detection algorithms, we intend to establish a reliable workflow for identifying SVs in several cancer types.

3. Material and methods

3.1. Sequencing data and SV caller selection

This study includes sequencing data obtained from tumour DNA and RNA panel testing of cancer patients that were referred to the Laboratory Genetics Department of the Portuguese Oncology Institute of Porto (IPO-Porto) for predictive, prognostic or diagnostic testing (Table 1). Sequencing was performed with Illumina short read paired-end technology (Illumina, Inc., San Diego, CA, USA), using three different target panels: TruSight RNA Fusion Panel (Illumina) (given the extended number of genes tested with this panel the complete list is indicated in Attachment 1), AVENIO Tumor Tissue panel (Roche, Basel, Switzerland) and the TruSight Hereditary Cancer Sequencing Panel (Illumina) (Table 1). The focus of this analysis was on clinically actionable genomic alterations, particularly SVs with therapeutic or diagnostic relevance.

Specifically, we selected sequencing data (FASTQ or BAM files depending on the SV caller) from 101 solid tumour RNA samples analysed by the TruSight RNA Fusion Panel and the commercial Dragen RNA app (Illumina), with 59 presenting actionable RNA fusion transcripts and 42 reported as negatives (Table 1). We also used sequencing data (BAM files) from 139 solid tumour DNA samples tested by the AVENIO Tumor Tissue panel and analysed by the AVENIO analysis software (Roche), with 44 presenting an actionable DNA gene fusion, 50 presenting CNVs (only *EGFR, ERBB2* and *MET* CNVs are reported by the Avenio software), and 45 reported as negatives for fusions and CNVs (Table 1). Finally, we also tested sequencing data (FASTQ files or BAM files depending on the SV caller) from six DNA tumour samples of patients tested by the TruSight Hereditary Cancer Sequencing Panel and presenting the germline *BRCA2* c.156_157insAlu variant (confirmed in DNA extracted from a blood sample) (Table 1). We also tested for a specific tool, sequencing data (BAM files) from 393 DNA tumour samples of patients that do not present this germline variant (confirmed in DNA extracted from a blood sample).

Selection of SV callers was performed based on a comprehensive literature review considering the type of short read sequencing data (DNA/RNA), type of SV (e.g. fusion, deletion, insertion), and computational requirements. To assess the detection performance and concordance of SV callers, both RNA-based and DNA-based NGS panels were analysed using multiple algorithms. On the RNA side, the analysis was conducted using STAR-Fusion (v1.15.0), Arriba (v2.5.0), and Manta (v1.6.0). For DNA-

based sequencing, SV detection tools such as Manta (v1.6.0), DELLY (v1.3.3), LUMPY (v0.3.1), and CNVkit were used (Table 1 and 2).

Table 1 - Sequencing data information and SV callers selected for this study.

Panel	Number of samples	Type of variant	Genome assembly	Input files	SV callers
TruSight RNA Fusion Panel (RNA tumour samples)	59 positives 42 negatives	Fusions*	hg19	FASTQ /BAM	Manta, Arriba, Star-Fusion
AVENIO Tumor Tissue panel (DNA tumour samples)	44 fusion positives 50 CNV positive 45 negatives	Fusions**: *ALK*, *FGFR2*, *FGFR3*, *NTRK1*, *RET*, *ROS1* CNVs: *EGFR*, *MET* and *ERBB2*	GRCh38	BAM	Manta, DELLY, LUMPY (fusions) CNVkit (CNVs)
TruSight Hereditary Cancer Panel (DNA tumour samples)	6 positive and 393 negatives***	*BRCA2* c.156_157insAlu	hg19	FASTQ /BAM	Manta, RetroSeq, Python-based script

*Complete list of genes tested is indicated in Attachment 1. **Detection of fusions are limited to pre-specified list of positions, referred to as "Loci of Interest" in the AVENIO analysis software. ***The 393 negative cases were only analysed with the Python-based script.

All analyses were executed on a computer with 64 GB RAM and running Ubuntu Linux. This setup ensured sufficient memory and stability for parallelized computation, especially important for the execution of alignment-intensive pipelines and the processing of large FASTQ and BAM files. Each tool was run using its recommended configurations or best practices as specified in official documentation, including filtering parameters, and confidence score thresholds when available.

3.2. SVs detection tools applied to RNA sequencing panel data

3.2.1. Manta

SV detection in RNA-based sequencing data was performed using Manta (v1.6.0) in RNA mode (Chen et al, 2016). Although this analysis is still in development and not fully

supported, it can be configured with the “- -rna” flag. This will adjust filtration levels and take other RNA-specific filtration and intron handling steps. Manta uses a two-step procedure to detect SVs: first, it scans the genome to identify regions with potential break-end associations, and second, it analyses these regions to infer, assemble, score, and filter candidate variants. For each sample, the RNA-based sequencing BAM file was provided as input using the “- -normalBam” argument, since RNA mode requires that one sample must be specified as normal input and no matched normal was available. The reference genome assembly used was hg19.

```
#!/bin/bash

# === CONFIGURATION ===
MANTA_DIR="/path/to/manta"           # Path to Manta installation
BAM_DIR="/path/to/rna_bams"         # Folder containing RNA-Seq BAM files
REF_FASTA="/path/to/reference.fa"   # Reference genome fasta (indexed)
OUTPUT_BASE="/path/to/manta_results" # Folder to save Manta output
THREADS=8                            # Number of CPU threads for parallel execution

# Create output directory if it does not exist
mkdir -p "$OUTPUT_BASE"

# === Loop through each RNA-Seq BAM file ===
for BAM in "$BAM_DIR"/*.bam; do
    SAMPLE=$(basename "$BAM" .bam)
    RUN_DIR="${OUTPUT_BASE}/${SAMPLE}_rna"

    echo "Running Manta RNA fusion calling for: $SAMPLE"

    # === Step 1: Configure Manta in RNA mode ===
    "$MANTA_DIR/bin/configManta.py" \
        --rna \
        --normalBam "$BAM" \
        --referenceFasta "$REF_FASTA" \
        --runDir "$RUN_DIR"

    # === Step 2: Adjust sensitivity for candidate detection ===
    if [ -d "$RUN_DIR" ]; then
        sed -i 's/minEdgeObservations = 3/minEdgeObservations = 2/'
"$RUN_DIR/config.ini"
        sed -i 's/minCandidateSpanningCount = 3/minCandidateSpanningCount = 2/'
"$RUN_DIR/config.ini"

        # === Step 3: Run the workflow ===
        cd "$RUN_DIR"
        ./runWorkflow.py -m local -j "$THREADS"

        # === Step 4: Copy fusion VCF output to main output directory ===
        RNA_VCF="$RUN_DIR/results/variants/rnaSV.vcf.gz"
        if [ -f "$RNA_VCF" ]; then
            cp "$RNA_VCF" "${OUTPUT_BASE}/${SAMPLE}_rnaSV.vcf.gz"
            echo "Fusion VCF saved: ${OUTPUT_BASE}/${SAMPLE}_rnaSV.vcf.gz"
        else
            echo "No RNA fusion VCF produced for $SAMPLE"
        fi
    else
        echo "ERROR: Failed to configure Manta for $SAMPLE"
    fi

    echo
done

echo "All RNA-Seq BAMs processed for fusion calling."
```

Figure 3 - Script with Manta workflow in RNA mode.

Table 2 - Characteristics of the SV callers selected for this study.

Tool	Type of variant	Detection strategy	Input	Output	Alignment	Computational requirements	Reference
Arriba	RNA fusions	SR, RP, JR	FASTQ, BAM	TSV	STAR	≥16 GB RAM	Uhrig et al, 2021
STAR-Fusion	RNA fusions	SM, SR, RP, JR	FASTQ, BAM	TSV	STAR	≥32 GB RAM	Haas et al, 2019
Manta (RNA flag)	RNA fusions	RP, SR, RD, assembly	BAM	VCF	STAR, DRAGEN RNA-seq aligner	≥16 GB RAM	Chen et al, 2016
Manta	DNA SVs (DEL, DUP, INV, TRA, INS)	RP, SR, RD, assembly	BAM	VCF	BWA-MEM, Bowtie 2, HISAT2, minimap2	≥16 GB RAM	Chen et al, 2016
DELLY	DNA SVs (DEL, DUP, INV, TRA, INS)	RP, SR, RD	BAM	VCF	BWA-MEM, Bowtie 2, minimap2	≥10 GB RAM	Rausch et al, 2012
LUMPY	DNA SVs (DEL, DUP, INV, TRA)	RP, SR, RD	BAM	VCF	BWA-MEM	≥16 GB RAM	Layer et al, 2014
CNVkit	CNVs	RD	BAM	SEG, VCF	BWA-MEM	≥16 GB RAM	Talevich et al, 2016
RetroSeq	MEI	RP, SR	BAM	VCF	MAQ, BWA	< 8 GB RAM	Keane et al, 2013

Legend: CNVs: Copy number variations; JR: Junction Reads; MEI: Mobile Element Insertions (e.g. Alu, LINE-1, SVA); RP: Read-pairs; SM: Split-mapped reads; SR: Split-reads; RD: Read depth; SVs: structural variants.

The workflow was executed in two main steps (Figure 3). First, Manta was configured using the configManta.py script with the "- -rna" flag to generate a sample-specific run directory containing the workflow configuration and task definitions. Sensitivity for candidate detection was increased by modifying the configuration parameters: "minEdgeObservations" and "minCandidateSpanningCount" were both lowered from 3 to 2, which allows detection of RNA fusions supported by fewer reads. Second, the runWorkflow.py script was executed in local mode with multi-threading to parallelize the analysis across available cores. The workflow produced the primary RNA fusion output in VCF format, named "rnaSV.vcf.gz", which contains information on each predicted fusion, including break-end positions, supporting paired and split reads, confidence intervals, and other relevant annotations.

3.2.2. Arriba

Gene fusion detection in RNA-based sequencing data was also performed using Arriba (v2.5.0), a computational tool that identifies fusion transcripts and other structural rearrangements from RNA sequencing data (Uhrig et al, 2021). Arriba operates downstream of the STAR aligner (v2.7.11b), which performs spliced read alignment in chimeric mode (Dobin et al, 2013). The aligned BAM output from STAR is then processed by Arriba to identify candidate gene fusions, filtered events, and optional annotations including known fusions, blacklisted regions, and protein domain information.

For each RNA-based sequencing sample, paired-end FASTQ files were provided as input. STAR was executed with parameters configured to generate chimeric alignments compatible with Arriba (Figure 4). The resulting BAM output was then streamed directly into Arriba, together with a reference genome FASTA file, a gene annotation GTF file, and optional databases for known fusions, blacklisted regions, and protein domains. Arriba produces two output files per sample: one containing high-confidence fusion calls (*_fusions.tsv) and another containing filtered or discarded events (*_fusions.discarded.tsv). The RNA-based sequencing pipeline was executed in a loop over all samples, using multi-threading to parallelize STAR alignment and Arriba processing. All intermediate and final outputs were stored in a structured output directory for downstream analysis.

```
#!/bin/bash

# === CONFIGURATION ===
STAR_BIN="/path/to/STAR"
ARRIBA_BIN="/path/to/arriba"
FASTQ_DIR="/path/to/fastqs"
STAR_INDEX="/path/to/star_index"
REF_FASTA="/path/to/reference.fa"
REF_GTF="/path/to/annotation.gtf"
DB_DIR="/path/to/arriba_database"
ARRIBA_OUT_DIR="/path/to/output"
THREADS=8

BLACKLIST="${DB_DIR}/blacklist.tsv.gz"
KNOWN="${DB_DIR}/known_fusions.tsv.gz"
PROTEINS="${DB_DIR}/protein_domains.gff3"

# Create output directory
mkdir -p "$ARRIBA_OUT_DIR"

# Loop through paired-end FASTQ files
for R1 in "$FASTQ_DIR"/*_R1*.fastq.gz; do
    SAMPLE=$(basename "$R1" _R1*.fastq.gz)
    R2="${FASTQ_DIR}/${SAMPLE}_R2*.fastq.gz"

    if [[ ! -f "$R2" ]]; then
        echo "Missing R2 for $SAMPLE, skipping..."
        continue
    fi
    echo "Processing $SAMPLE"
    "$STAR_BIN" \
        --runThreadN "$THREADS" \
        --genomeDir "$STAR_INDEX" \
        --readFilesIn "$R1" "$R2" \
        --readFilesCommand zcat \
        --outSAMtype BAM Unsorted \
        --outSAMunmapped Within \
        --chimOutType WithinBAM HardClip \
        --chimSegmentMin 10 \
        --chimJunctionOverhangMin 10 \
        --chimSegmentReadGapMax 3 \
        --chimScoreDropMax 30 \
        --chimScoreSeparation 1 \
        --chimScoreJunctionNonGTAG 0 \
        --chimMultimapNmax 50 \
        --outFilterMultimapNmax 50 \
        --peOverlapNbasesMin 10 \
        --alignSplicedMateMapLminOverLmate 0.5 \
        --alignSJstitchMismatchNmax 5 -1 5 5 \
    | "$ARRIBA_BIN" \
        -x /dev/stdin \
        -a "$REF_FASTA" \
        -g "$REF_GTF" \
        -b "$BLACKLIST" \
        -k "$KNOWN" \
        -t "$KNOWN" \
        -p "$PROTEINS" \
        -o "${ARRIBA_OUT_DIR}/${SAMPLE}_fusions.tsv" \
        -O "${ARRIBA_OUT_DIR}/${SAMPLE}_fusions.discarded.tsv"
done
```

Figure 4 - Script with parameters used for running Arriba tool.

3.2.3. STAR-Fusion

Gene fusion detection in RNA sequencing data was additionally performed with STAR-Fusion (Haas et al, 2019), a method that builds upon the STAR aligner (v2.7.11b) to identify candidate fusion transcripts from chimeric RNA-based sequencing read alignments. The approach consists of two main stages. First, STAR is used to align paired-end reads to the human reference genome (in this study we used the hg19 assembly) while enabling detection of chimeric reads, that is, reads or read pairs aligning to distinct genomic loci. These chimeric alignments provide evidence for potential gene fusion breakpoints and are recorded in the file “Chimeric.out.junction”. This alignment

step was carried out using STAR in two-pass mode with RNA-specific parameters optimized for chimera discovery, including settings for minimum chimeric segment length, overhangs, and mismatch tolerances. Each sample was processed individually, producing both a coordinate-sorted BAM file and the corresponding "Chimeric.out.junction" file.

```
#!/bin/bash

# ====== User-defined variables ======
STAR_BIN="/path/to/STAR"
FASTQ_DIR="/path/to/fastq_trimmed"
STAR_INDEX="/path/to/STAR_index"
REF_FASTA="/path/to/reference.fa"
REF_GTF="/path/to/annotation.gtf"
THREADS=8

# ====== Input: sample FASTQ pair ======
for R1 in "$FASTQ_DIR"/*_R1_trimmed.fastq.gz; do
    SAMPLE=$(basename "$R1" _R1_trimmed.fastq.gz)
    R2="${FASTQ_DIR}/${SAMPLE}_R2_trimmed.fastq.gz"

    if [[ ! -f "$R2" ]]; then
        echo "Missing R2 file for $SAMPLE, skipping..."
        continue
    fi

    echo -e "\n Running STAR for $SAMPLE"

# ====== Output directory ======
OUTDIR="/path/to/star_output/${SAMPLE}"
mkdir -p "$OUTDIR"

# ====== Run STAR alignment ======
"$STAR_BIN" \
  --genomeDir "$STAR_INDEX" \
  --readFilesIn "$R1" "$R2" \
  --readFilesCommand gunzip -c \
  --runThreadN "$THREADS" \
  --twopassMode Basic \
  --outReadsUnmapped None \
  --outSAMstrandField intronMotif \
  --outSAMunmapped Within \
  --outSAMtype BAM SortedByCoordinate \
  --outFileNamePrefix "${OUTDIR}/" \
  --chimSegmentMin 12 \
  --chimJunctionOverhangMin 8 \
  --chimOutJunctionFormat 1 \
  --alignSJDBoverhangMin 10 \
  --alignMatesGapMax 100000 \
  --alignIntronMax 100000 \
  --alignSJstitchMismatchNmax 5 -1 5 5 \
  --outSAMattrRGline ID:${SAMPLE} \
  --chimMultimapScoreRange 3 \
  --chimScoreJunctionNonGTAG -4 \
  --chimMultimapNmax 20 \
  --chimNonchimScoreDropMin 10 \
  --peOverlapNbasesMin 12 \
  --peOverlapMMp 0.1 \
  --alignInsertionFlush Right \
  --alignSplicedMateMapLminOverLmate 0 \
  --alignSplicedMateMapLmin 30 \
  --outTmpDir "/tmp/star_${SAMPLE}_tmp"

echo "STAR alignment complete. Output is in $OUTDIR"

done
```

Figure 5 - Script with parameters for STAR alignment with chimeric detection.

In the second stage, STAR-Fusion processes the “Chimeric.out.junction" file to cluster chimeric junctions and predict candidate fusions. Supporting evidence from split-reads and spanning fragments is aggregated for each candidate, and potential artifacts such as read-through transcription and paralogous gene mappings are filtered out. Fusion breakpoints are annotated with respect to gene models and cross-referenced against the Comprehensive Transcriptome Annotation of Translocations (CTAT) resource library, which integrates GENCODE gene annotations and known fusion databases such as COSMIC and TCGA. The final output consists of a tab-delimited file listing predicted fusions, the number of supporting reads, genomic coordinates of the breakpoints, and annotations regarding known or novel status.

```
#!/bin/bash

CTAT_LIB="/path/to/ctat_genome_lib_build_dir"
STAR_DIR="/path/to/star_output"
OUT_DIR="/path/to/star_fusion_results"

mkdir -p "$OUT_DIR"

for SAMPLE_DIR in "STAR_DIR"/*; do
    SAMPLE=$(basename "$SAMPLE_DIR")
    CHIMERIC="$SAMPLE_DIR/Chimeric.out.junction"

    if [ -f "$CHIMERIC" ]; then
        echo "Processing $SAMPLE"

        /path/to/STAR-Fusion \
          --genome_lib_dir "$CTAT_LIB" \
          --chimeric_junction "$CHIMERIC" \
          --output_dir "$OUT_DIR/$SAMPLE" \
          --CPU 8
    else
        echo "Skipping $SAMPLE (missing Chimeric.out.junction)"
    fi
done
```

Figure 6 - Script with instructions for fusion prediction with STAR-Fusion.

The workflow was implemented in two main steps using custom Bash scripts (Figures 5 and 6). In the first step, paired-end FASTQ files were aligned with STAR to generate chimeric junction files for each sample. In the second step, STAR-Fusion was run on the STAR outputs using the CTAT resource library for hg19 assembly. Results were stored in per-sample output directories containing the main prediction file “star-fusion.fusion_predictions.tsv”. This file provided the primary input for downstream analyses of fusion candidates across the study cohort.

3.3. SVs detection tools applied to DNA sequencing panel data

Sequencing data from tumour DNA samples tested with the AVENIO Tumor Tissue panel were analysed for SVs, using the human reference genome assembly GRCh38. The sequencing data was originally analysed with the AVENIO analysis software, which reports fusions in a restricted set of "loci of interest" (*ALK*, *FGFR2*, *FGFR3*, *NTRK1*, *RET*, *ROS1*) and CNVs in clinically relevant genes (*EGFR*, *MET*, *ERBB2*) (Table 1). For this study, we extended SV detection to additional tools to comprehensively assess performance across different variant classes. Specifically, Manta (v1.6.0), DELLY (v1.3.3) and LUMPY (v0.3.1) were employed for the detection of gene fusions and CNVkit was used for CNV analysis. All tools were applied in tumour-only mode, with BAM files as input.

3.3.1. Manta

SV detection in DNA sequencing data was performed with Manta (v1.6.0). As previously indicated, Manta integrates read-pair, split-read, read depth, and local assembly signals to detect a broad spectrum of variants, including deletions, duplications, inversions, translocations, and insertions (Chen et al, 2016). Each BAM file was analysed independently in tumour-only mode using genome assembly GRCh38. The workflow was configured with configManta.py and executed with runWorkflow.py in local multithreaded mode (Figure 7). Manta outputted VCF files containing SVs breakpoints annotated with break-end coordinates, variant type and supporting read evidence.

```
#!/bin/bash

# CONFIGURATION
MANTA_DIR="/path/to/manta"
BAM_DIR="/path/to/bam_files"
REF_FASTA="/path/to/reference_genome.fa"
OUTPUT_DIR="/path/to/output_directory"
THREADS=8

mkdir -p "$OUTPUT_DIR"

for BAM in "$BAM_DIR"/*.bam; do
    SAMPLE=$(basename "$BAM" .bam)
    RUN_DIR="${OUTPUT_DIR}/${SAMPLE}_analysis"

    echo ">>> Processing sample: $SAMPLE"

    "$MANTA_DIR/bin/configManta.py" \
        --tumorBam "$BAM" \
        --referenceFasta "$REF_FASTA" \
        --runDir "$RUN_DIR"

    if [ -d "$RUN_DIR" ]; then
        cd "$RUN_DIR"
        ./runWorkflow.py -m local -j "$THREADS"
        echo ">>> Completed: $SAMPLE"
    else
        echo "ERROR: Manta config failed for $SAMPLE, skipping."
    fi

    echo
done

echo "All samples processed."
```

Figure 7 - Script with instructions for running Manta tool on tumour-only mode.

3.3.2. LUMPY

SV detection was also tested with LUMPY (v0.3.1), a probabilistic framework combining discordant read-pairs, split-reads, and read depth evidence (Layer et al, 2014). BAM files were pre-processed to extract discordant and split reads, which were then combined with the full alignment as input (Figure 8). Each sample was analysed independently, producing VCF outputs listing SVs, with breakpoint resolution supported by split-read and paired-end evidence when available.

```
#!/bin/bash
# === CONFIGURATION ===
TOOLS_DIR="/path/to/tools"                          # Path to directory containing Picard, samtools.
INPUT_DIR="/path/to/bam_files"                      # Directory containing input BAM files
RG_DIR="${INPUT_DIR}/with_RG"                       # Output directory for BAMs with added Read Groups
OUTPUT_DIR="/path/to/lumpy_results"                 # Directory to store LUMPY output VCFs
REFERENCE="/path/to/reference_genome.fa"            # Reference genome FASTA file
PICARD_JAR="${TOOLS_DIR}/picard/picard.jar"         # Path to Picard .jar file

# Create necessary directories
mkdir -p "$RG_DIR" "$OUTPUT_DIR"

# === MAIN LOOP ===
for BAM in "${INPUT_DIR}"/*.bam; do
    BASENAME=$(basename "$BAM" .bam)
    RG_BAM="${RG_DIR}/${BASENAME}_RG.bam"
    VCF_OUT="${OUTPUT_DIR}/${BASENAME}.lumpy.vcf"

    echo ">>> Processing sample: $BASENAME"

    # Step 1: Add Read Groups (required for LUMPY)
    java -jar "$PICARD_JAR" AddOrReplaceReadGroups \
        I="$BAM" \
        O="$RG_BAM" \
        RGID="$BASENAME" \
        RGLB="lib1" \
        RGPL="illumina" \
        RGPU="unit1" \
        RGSM="$BASENAME"

    # Step 2: Index the new BAM file
    samtools index "$RG_BAM"

    # Step 3: Run LUMPY via lumpyexpress
    lumpyexpress \
        -B "$RG_BAM" \
        -R "$REFERENCE" \
        -o "$VCF_OUT"

    echo "✓ Finished processing: $BASENAME"
    echo
done

echo "All samples processed successfully."
```

Figure 8 - Script with instructions for running LUMPY.

3.3.3. DELLY

DELLY (v1.3.3) was applied for the detection of SVs including deletions, duplications, inversions, and translocations (Rausch et al, 2012). DELLY integrates paired-end and split-read information and provides genotype likelihoods for each variant. Analyses were performed in tumour-only mode using GRCh38 assembly as reference (Figure 9). Each BAM file produced a VCF reporting SVs, including variant type, break-end position, genotype likelihoods, and supporting read counts.

```
#!/bin/bash

# === CONFIGURATION ===
BAM_DIR="/path/to/bam_files"                        # Directory containing input BAM files
REF="/path/to/reference_genome.fa"                  # Reference genome in FASTA format
EXCL="/path/to/exclusion_regions.tsv"               # Regions to exclude (e.g., centromeres, telomeres)
OUT_DIR="/path/to/output_directory"                 # Directory to store DELLY output VCFs
DELLY_EXEC="/path/to/delly"                         # Path to the compiled DELLY binary

# Create output directory if it doesn't exist
mkdir -p "$OUT_DIR"

# === MAIN LOOP ===
for BAM in "$BAM_DIR"/*.bam; do
    SAMPLE=$(basename "$BAM" .bam)
    echo ">>> Processing sample: $SAMPLE"

    "$DELLY_EXEC" call \
        -g "$REF" \
        -x "$EXCL" \
        "$BAM" > "$OUT_DIR/${SAMPLE}.vcf"

    echo "✓ Finished: $SAMPLE"
    echo
done

echo "All samples processed."
```

Figure 9 - Script with instructions for running DELLY in tumour-only mode.

3.3.4. CNVkit

The detection of CNVs was performed with CNVkit (Talevich et al, 2016), which estimates copy number ratios from targeted sequencing data by analysing read depth across target and off-target regions. Each tumour BAM was processed with the panel BED file and GRCh38 assembly as reference (Figure 10). CNVkit applies multiple bias corrections (GC content, target size, coverage) before segmenting the genome into regions of gain or loss. Output included per-sample copy ratio files, segmented CNV profiles (SEG), and optional CNV calls in VCF format.

```
#!/bin/bash
set -euo pipefail

# Input data
BAM_DIR="bam"
REF_DIR="reference"
RESULTS_DIR="results"
THREADS=8

# Reference files
TARGET_BED="${REF_DIR}/targets.bed"
ACCESS_BED="${REF_DIR}/access.bed"
REF_FASTA="${REF_DIR}/genome.fa"

# 1. Optimize target and antitarget bins
cnvkit.py autobin \
  ${BAM_DIR}/*.bam \
  -t ${TARGET_BED} \
  -g ${ACCESS_BED}

# 2. Build CNV reference
cnvkit.py reference \
  -t *.target.bed \
  -a *.antitarget.bed \
  -f ${REF_FASTA} \
  -o ${REF_DIR}/cnv_reference.cnn

# 3. CNV analysis and visualization
cnvkit.py batch \
  ${BAM_DIR}/*.bam \
  -r ${REF_DIR}/cnv_reference.cnn \
  -m hybrid \
  -p ${THREADS} \
  -d ${RESULTS_DIR} \
  --scatter
```

Figure 10 - Script with instructions for running CNVkit.

3.4. *BRCA2* c.156_157insAlu variant detection tools applied to DNA sequencing panel data

3.4.1. Python-based script

To detect the germline *BRCA2* c.156_157insAlu Portuguese founder variant in FFPE tumour DNA samples, we developed a Python-based script that automates the analysis of NGS data stored in BAM format. This script focuses on a defined genomic region on chromosome 13 (chr13:32893278-32893320), where the Alu element was inserted. Using embedded pattern-matching logic, the script searches for reads containing the insertion junction and wild-type sequences, employs Samtools software (Danecek et al, 2021) to extract the relevant region from each BAM file and counts sequence matches in the resulting SAM file. Results for each sample are saved in a tab-separated file, enabling downstream aggregation and review. The script also features a graphical user interface for file selection and production of logs when processing errors occur.

This method was employed on the BAM files of the six tumour DNA samples of the patients with the germline *BRCA2* c.156_157insAlu variant and on the 393 tumour DNA samples of patients that do not present this germline variant.

3.4.2. Retroseq

RetroSeq is a computational method designed for the discovery and genotyping of transposable elements (TE), also called mobile element insertions (MEIs), from NGS reads aligned to a reference genome in BAM format (Keane et al, 2013). This tool was initially developed to detect non-reference TE insertions from Illumina paired-end whole-genome sequencing data (Keane et al, 2013). Its main objective is to identify TE insertions that are absent from the reference genome but present in the sequenced sample. This tool operates in two main phases. In the discovery phase, RetroSeq scans aligned sequencing reads in BAM format to detect discordant mate pairs, in which one read aligns uniquely to the reference genome while its mate aligns to a known TE sequence. Candidate TE-supporting reads are then assigned to a class of elements (e.g., Alu, LINE, SINE) using either annotated TE positions in the reference genome or alignment against a supplied library of consensus TE sequences. In the subsequent calling phase, RetroSeq clusters the supporting reads to define candidate insertion sites, applies quality thresholds (e.g., minimum read support, mapping quality, sequence identity), and filters out calls overlapping annotated reference elements. The results are outputted in VCF format, with supporting read counts and confidence levels annotated. Calls with higher “FL” scores indicate better-resolved breakpoints and stronger evidence.

This method was applied to the BAM files obtained from sequencing six tumour DNA samples of patients carrying the germline *BRCA2* c.156_157insAlu variant. The BAM files originally available in our department were not fully compatible with RetroSeq. To address this, raw sequencing data (FASTQ files) were re-aligned with BWA-MEM against the human reference genome assembly hg19, generating RetroSeq-compatible BAM files. Subsequently, RetroSeq was executed in its standard two-phase workflow, discovery and calling, to identify candidate MEIs. The script used for execution is shown in Figure 11.

```
#!/bin/bash
# RetroSeq pipeline for detection of mobile element insertions (MEIs)

# ==============================
# Configuration
# ==============================

# Define base directories
RETROSEQ_DIR="/path/to/RetroSeq"
BAM_DIR="/path/to/bam_files"
REF_FASTA="/path/to/reference_genome.fa"
EREF="/path/to/eref.tab"   # tab-delimited file linking TE types to
consensus sequences
RETROSEQ="$RETROSEQ_DIR/bin/retroseq.pl"
RESULTS_DIR="/path/to/results"

# Use RetroSeq-compatible samtools (v0.1.19 recommended)
export PATH="/path/to/samtools-0.1.19:$PATH"

# Create results directory if it does not exist
mkdir -p "$RESULTS_DIR"

# ==============================
# Run RetroSeq
# ==============================

for BAM in "$BAM_DIR"/*.bam; do
  SAMPLE=$(basename "$BAM" .bam)
  echo "Processing $SAMPLE..."

  # Step 1: Discovery
  "$RETROSEQ" -discover \
    -bam "$BAM" \
    -eref "$EREF" \
    -align \
    -output "$RESULTS_DIR/${SAMPLE}_discovery.txt"

  # Step 2: Call
  "$RETROSEQ" -call \
    -bam "$BAM" \
    -input "$RESULTS_DIR/${SAMPLE}_discovery.txt" \
    -ref "$REF_FASTA" \
    -output "$RESULTS_DIR/${SAMPLE}_calls.vcf" \
    -reads 2 \
    -depth 1000

  echo "Finished: $SAMPLE"
done
```

Figure 11 - Script with instructions used to run RetroSeq tool.

As indicated previously RetroSeq analysis required alignment of raw FASTQ data with BWA-MEM to produce suitable BAM files (Li et al, 2009). Due to storage constraints, alignment of negative control samples, requiring several hundred gigabytes, was not feasible, and therefore RetroSeq analysis was restricted to positive cases.

3.4.3. Manta

SV detection was performed using Manta (v1.6.0) in tumour-only mode, as no matched normal samples were available and as previously described (only the reference genome

was changed to hg19). An automated workflow was implemented using the Bash script indicated previously (Figure 7), which iterates over all BAM files in a designated directory, configures Manta for each sample, and executes the workflow locally with parallel threads. The primary output of this analysis was the tumorSV.vcf.gz file, which contains all candidate SVs detected. Each variant record includes paired-read and split-read support counts, as well as an "IMPRECISE" flag for variants where assembly could not fully resolve the breakpoints. Additional unscored SV candidates were captured in "candidateSV.vcf.gz" files and small indels (<50 bp) in "candidateSmallIndels.vcf.gz" files. Filtering of high-confidence variants relied primarily on the paired and split reads support counts, with caution applied to imprecise variants.

3.5. SVs detection tools performance evaluation

Evaluation metrics included sensitivity, specificity, precision, and F1-score, and it was calculated by comparing the SV calls generated by each tool against the reference calls previously detected with the current bioinformatic tools used at the Genetics Department. For RNA sequencing data, results from Manta, Arriba, and STAR-Fusion were benchmarked against DRAGEN RNA app (Illumina), a commercially available fusion detection software. For DNA-based analyses, SVs identified by Manta, DELLY, LUMPY, and CNVkit were compared to AVENIO software, which served as the reference platform. All the fusions were manually inspected in IGV.

Metrics were computed in RStudio (version 2024.09.0+375, R Foundation for Statistical Computing, Vienna, Austria) using custom scripts. Briefly, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were defined for each variant type per sample. Sensitivity was calculated as TP/(TP+FN), specificity as TN/(TN+FP), precision as TP/(TP+FP), and recall as TP/(TP+FN). The F1-score, representing the harmonic mean of precision and recall, was calculated as 2×(precision× recall)/(precision+recall). Additionally, summary statistics were visualized using ggplot2, allowing comparison of performance across SV callers and variant types (Wickham et al, 2016).

4. Results

4.1. SV callers' performance using RNA sequencing data

The performance of SV callers using RNA sequencing data was assessed using the Illumina® DRAGEN RNA app (Illumina) results as the reference standard. Metrics included true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity, specificity, precision, and F1-score (Table 3). Among the evaluated tools, Arriba and STAR-Fusion achieved the highest sensitivity (1.00), while Manta demonstrated slightly lower sensitivity (0.86), missing a total of 8 fusion events detected by DRAGEN RNA app (Illumina). In terms of specificity, Manta (0.93) slightly outperformed STAR-Fusion (0.88) and Arriba (0.88) reflecting fewer false positives.

Table 3 - Performance metrics of the tools used to detect SVs in RNA sequencing data.

Tool	TP	TN	FP	FN	Sensitivity	Specificity	Precision	F1-score
Manta	51	39	3	8	0.86	0.93	0.94	0.90
Arriba	59	38	4	0	1.00	0.90	0.94	0.97
STAR-Fusion	59	37	5	0	1.00	0.88	0.92	0.96

Legend: TP: true positives, TN: true negatives; FP: false positives; FN: false negatives.

Figure 12 shows a grouped bar plot of sensitivity, specificity, precision, and F1-score for each tool, highlighting their comparative performance.

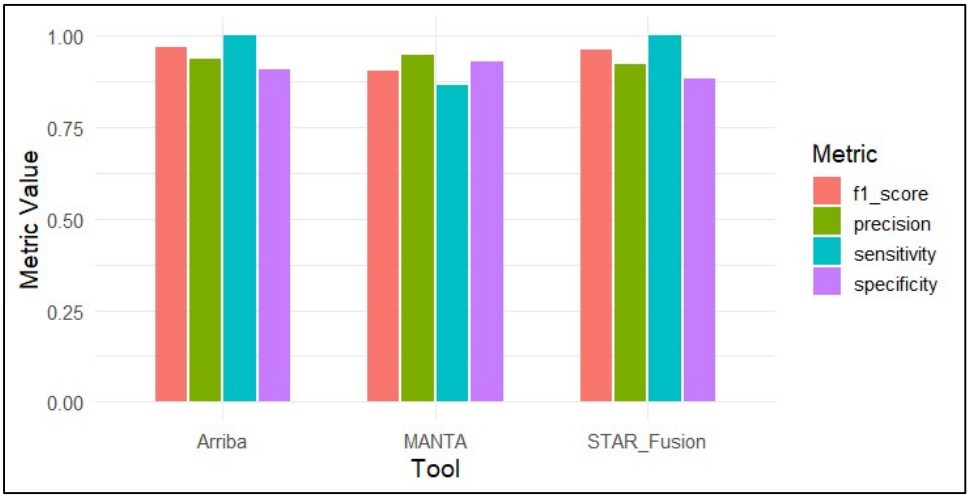

Figure 12 - Performance metrics of the tools used to detect SVs in RNA sequencing data.

The distribution of SV types, that gave origin to transcript fusions, across all RNA sequencing data revealed that translocations were the most frequent type (n = 35), followed by inversions (n = 21), deletions (n = 7), and duplications (n = 1) (Figure 13).

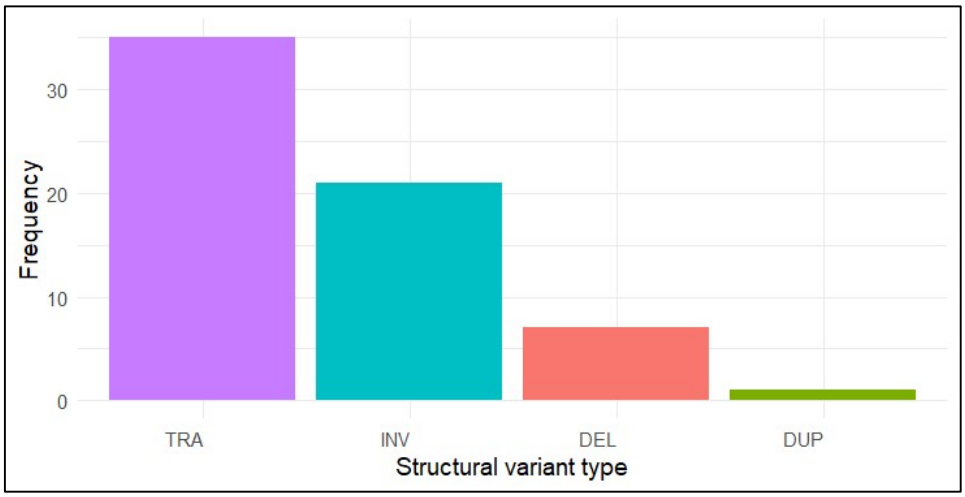

Figure 13 - Bar plot showing the distribution of SV types detected in RNA sequencing data across all SVs detection tools. The SV categories include translocations (TRA), inversions (INV), deletions (DEL), and duplications (DUP), with translocations being the most frequent and duplications less common.

Concordance analysis using UpSet plot (Figure 14) indicated substantial overlap among the callers. However, the total number of detected SVs differed among tools, with STAR-Fusion reporting the highest number (n = 64), followed by Arriba (n = 63), DRAGEN RNA app (n = 59), and Manta (n = 54). The caller that missed more events was Manta tool (n = 10). Nonetheless, the tool used in the clinical diagnostics of our Department (Dragen RNA app) missed five fusion transcripts, namely one *MEAF6*::*PHF1* (Figure 15A), one *YWHAE*::*NUTM2B* (Figure 15B), one *CIC*::*DUX4 and* two *RPSAP52*::*HMGA2* (Figure 15C) fusion transcripts. Although these fusion transcripts present low quality reads and a low frequency of split-reads, the fusions where detected at least by two tools with high confidence, are in-frame (as indicated by the Arriba tool and by manual inspection of the BAM files using IGV) and the breakpoints are in accordance with the reported in the literature for the same type of tumours.

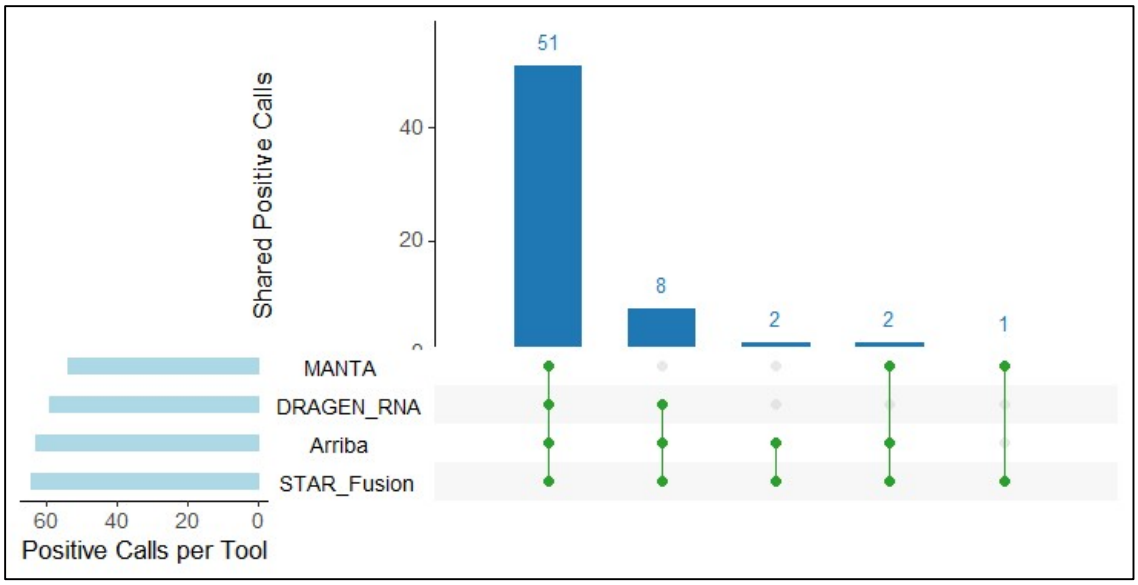

Figure 14 - Upset plot illustrating the overlap in SV detection among four RNA sequencing-based tools (DRAGEN RNA app, Manta, Arriba, and STAR-Fusion). The vertical bars represent the number of shared fusion events detected by the combination of tools indicated by the filled dots below each bar. Horizontal bars on the left show the total number of positive calls detected by each tool. The plot highlights both concordant and tool-specific fusion calls, emphasizing the variability in fusion detection across algorithms.

4.2. SV callers' performance using DNA sequencing data

Regarding fusions genes, the performance evaluation of the SV callers used Avenio software (Roche) as the reference. All the tools demonstrated strong performance with DELLY and LUMPY achieving perfect sensitivity (1.00), identifying all true positives (n = 43) with no false negatives. Manta also performed well, with a sensitivity of 0.95 and only two missed fusions (Table 4, Figure 16).

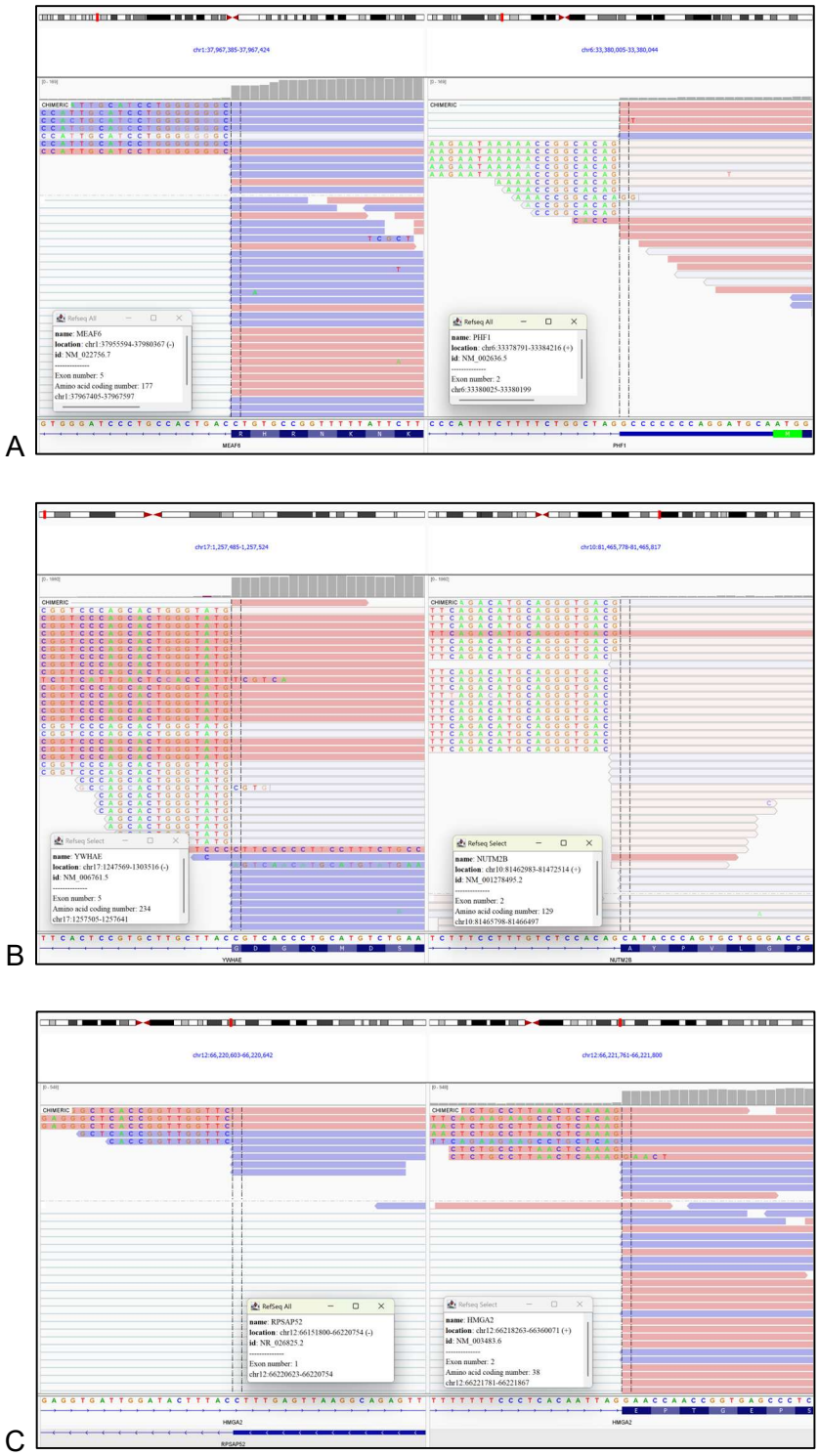

Figure 15 - IGV visualization of three fusion transcripts not detected by the DRAGEN RNA app (Illumina). A) *MEAF6*::*PHF1* in-frame fusion transcript visualized with split-reads supporting the junction site across exons 5 and 2 of *MEAF6* and *PHF1* genes, respectively. B) *YWHAE*::*NUTM2B* in-frame fusion, showing soft-clipping at the breakpoint in exons 5 and 2 of the *YWHAE* and *NUTM2B* genes, respectively. C) *RPSAP52*::*HMGA2* in-frame fusion transcript, with split-reads confirming the fusion junction at exon 1 and 2 of the *RPSAP52* and *HMGA2* genes, respectively.

Table 4 - Performance metrics of the tools used to detect SVs in DNA sequencing data.

Tool	TP	TN	FP	FN	Sensitivity	Specificity	Precision	F1-score
Manta	41	44	2	2	0.95	0.95	0.95	0.95
DELLY	43	44	2	0	1	0.96	0.96	0.98
LUMPY	43	44	2	0	1	0.96	0.98	0.98

Legend: TP: true positives, TN: true negatives; FP: false positives; FN: false negatives.

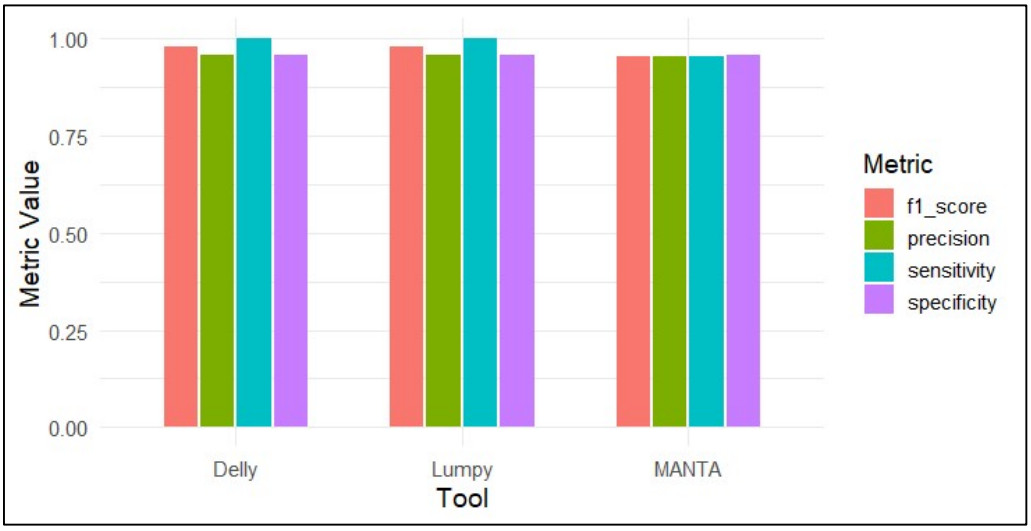

Figure 16 - Performance metrics of the tools used for the detection of fusion genes in DNA sequencing data.

In our series, fusion gene analysis revealed that the majority resulted from inversion mechanisms (n = 35), followed by deletion (n = 6) and translocation (n = 4) (Figure 17).

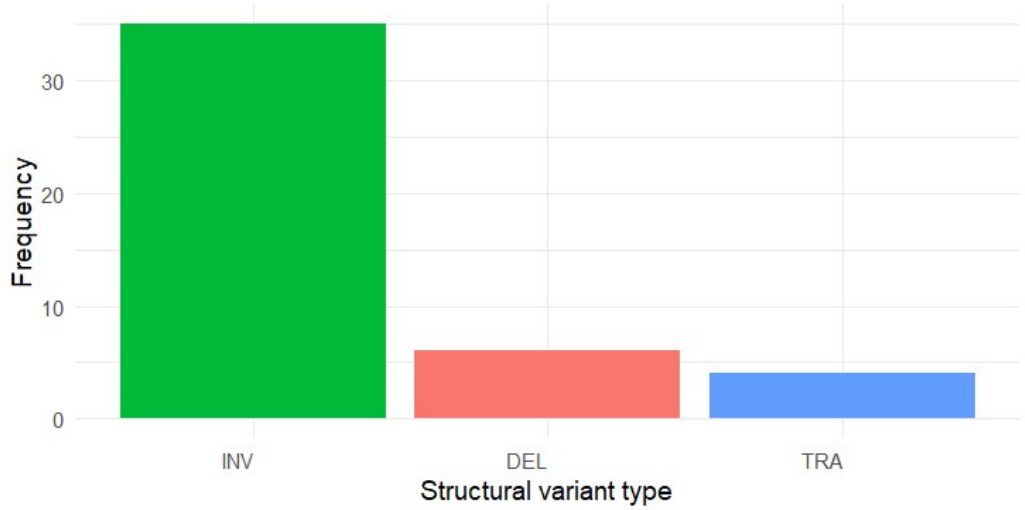

Figure 17 - Bar plot showing the distribution of the different SV types detected in DNA sequencing that gave rise to fusion events. The SV categories include translocations (TRA), inversions (INV) and deletions (DEL), with inversions being the most frequent in our series.

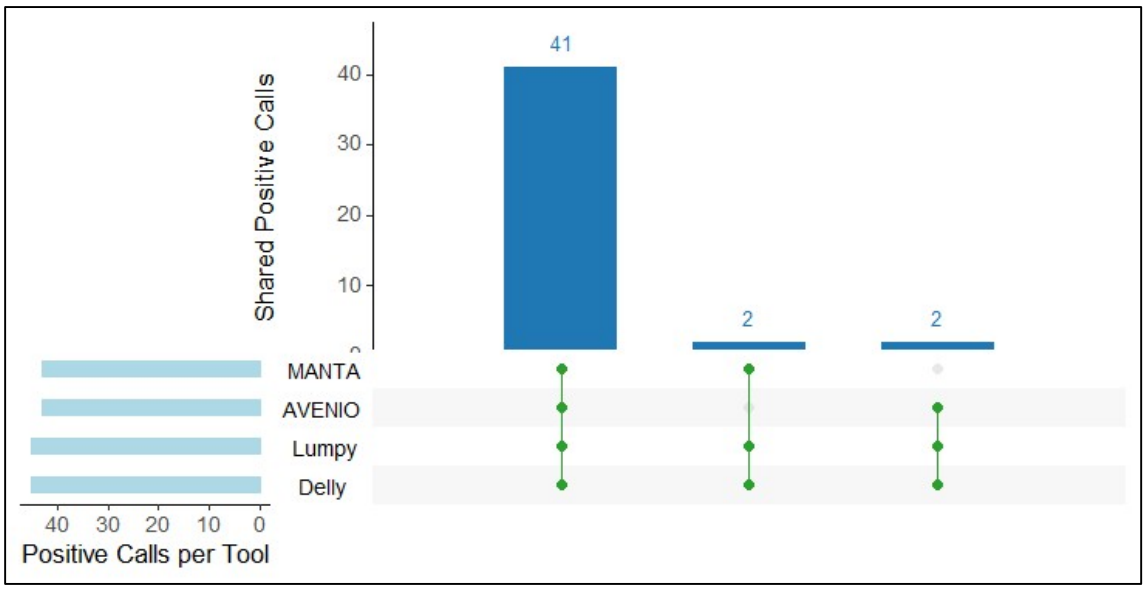

Figure 18 - UpSet plot showing the high concordance of fusion gene detection in DNA sequencing data among four tools (AVENIO software, Manta, DELLY, LUMPY). The intersection bars depict the number of shared fusion calls between tools, while the matrix below indicates which tools contributed to each intersection.

Concordance analysis using UpSet plot (Figure 18) also indicated substantial overlap among the callers. LUMPY and DELLY reported the highest number (n = 45), followed by Avenio software (Roche) and Manta (n = 43). The two fusion events missed by the Avenio software (Roche) were two *EML4*::*ALK* fusions present in NSCLC samples with breakpoint sites well established in the literature. These fusions were confirmed by manual inspection in IGV and one of them was previously detected by other methodology.

Regarding amplification events, we only evaluated *EGFR*, *ERBB2* and *MET* genes by CNVkit tool and considered amplification when copy number gain ≥ 6 (Kowalewski et al, 2025). CNVkit revealed poor sensitivity (0.24) as indicated in Table 5. In Figure 19 is displayed a copy number ratio plot illustrating *EGFR* amplification in a tumour sample, as detected by CNVkit.

Table 5 - Performance metrics of the CNVkit tool used to detect CNV in DNA sequencing data.

Tool	TP	TN	FP	FN	Sensitivity	Specificity	Precision	F1-score
CNVkit	12	45	0	38	0.24	1	1	0.39

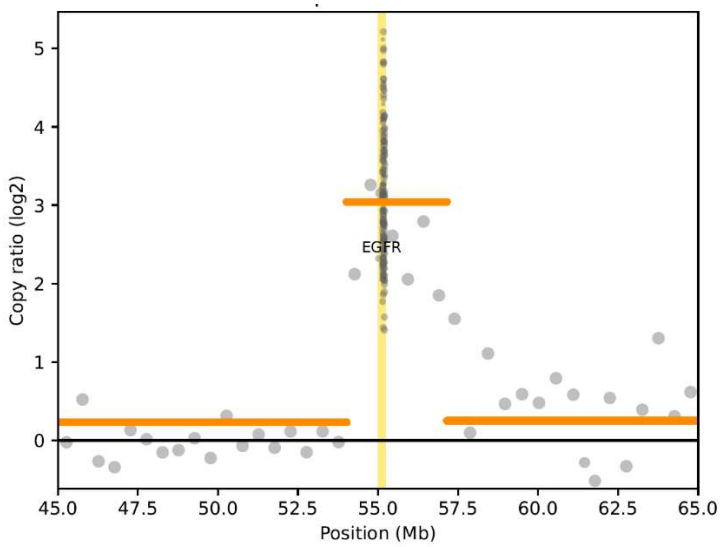

Figure 19 - Copy number ratio plot illustrating *EGFR* amplification in a tumour sample, as detected by CNVkit. The x-axis represents genomic position on chromosome 7 (Mb), and the y-axis shows the $\log_2$ copy ratio. A prominent high-level amplification is observed at approximately 55 Mb, corresponding to the *EGFR* locus (highlighted in yellow).

4.3. *BRCA2* c.156_157insAlu detection tools performance on DNA sequencing data

To assess the ability of SV callers to detect MEIs, we analysed six DNA tumour samples from patients with a confirmed germline *BRCA2* c.156_157insAlu variant (Figure 20). We tested the BAM files of these six tumour DNA samples using our python-based program and all the cases were detected. Furthermore, we tested 393 tumour DNA samples of patients that do not present this germline variant (confirmed in DNA from a blood sample), and all the cases were negative using this methodology, revealing a concordance of 100%. RetroSeq successfully identified the Alu insertion in four out of the six tumour samples, while Manta failed to detect the insertion in any of the samples with high confidence.

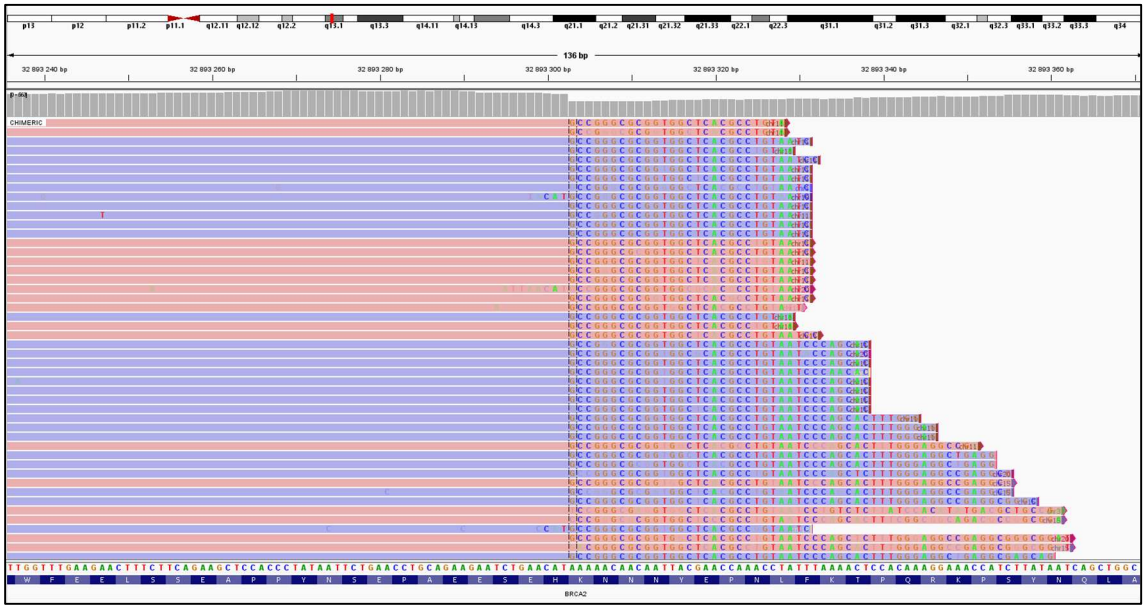

Figure 20 - IGV visualization of the *BRCA2* c.156_157insAlu variant. The insertion is indicated by the presence of split-reads at the breakpoint region in exon 3 of the *BRCA2* gene. A characteristic drop in coverage at the insertion site further supports the presence of this Alu element.

5. Discussion

In this study, we systematically compared multiple bioinformatic tools for detecting SVs in tumour samples analysed through targeted DNA and RNA sequencing panels.

5.1. SV callers' performance using RNA sequencing data

Among the RNA-based sequencing fusion callers, Arriba and STAR-Fusion achieved perfect sensitivity against the DRAGEN RNA app (Illumina), whereas Manta (in RNA mode) failed to detect eight out of 59 validated fusion events. However, Manta displayed higher specificity relative to STAR-Fusion and Arriba by producing fewer false positives. This result is consistent with previous benchmarking studies, which have placed STAR-Fusion and Arriba among the best-performing fusion detectors in terms of speed and sensitivity (Haas et al, 2019; Uhrig et al, 2021; Creason et al, 2021).

Manta's lower sensitivity could be because this tool is primarily designed for DNA-based SVs and is not fully validated for RNA analysis. Its more conservative filtering and breakpoint assembly requirements are likely to reduce false positives but at the expense of missing lower-evidence fusions (Shrestha et al, 2023). In contrast, RNA-centred callers like Arriba and STAR-Fusion explicitly optimize transcriptomic evidence, chimeric read detection, and splicing junction constraints (Haas et al, 2019; Uhrig et al, 2021).

The discordance in specificity also warrants attention. STAR-Fusion's lower specificity may reflect its more aggressive fusion calling strategy, potentially generating more false positives in low-expression or noisy regions. In diagnostic settings, false positives are a problem, as they may trigger unnecessary validation or clinical decisions. However, our results show that STAR-Fusion does not lack specificity, but that DRAGEN RNA app, a commercial clinical-grade pipeline used in our laboratory, missed five fusions that were robustly detected by at least two other callers. Among these were fusions such as *MEAF6*::*PHF1*, *YWHAE*::*NUTM2B*, *CIC*::*DUX4*, and *RPSAP52*::*HMGA2*. Though they had low read support or low-quality alignments, they were in-frame and concordant with known breakpoints; manual inspection and literature evidence suggest they may be bona fide events. This underscores that even validated clinical pipelines can miss relevant fusions, especially in challenging or low-evidence contexts. For example, the detection of *CIC*::*DUX4* fusion transcript is technically challenging due to *DUX4* gene location in the highly repetitive sub-telomeric genomic regions (4q35 or 10q26), which complicates alignment and breakpoint resolution (Jeck et al, 2021).

Taken together, these observations reinforce that no single RNA SV caller is enough. In clinical practice, combining tools or applying consensus rules can mitigate false negatives while controlling false positives.

5.2. SV callers' performance using DNA sequencing data

In the DNA panel dataset, DELLY and LUMPY achieved 100 % sensitivity in calling fusion events (n = 43), correctly identifying all known positives. Manta tool missed two, yielding a lower sensitivity. This strong performance is encouraging and aligns with expectations that SV callers optimized for DNA can capture rearrangements with accuracy (Joe et al, 2024; Zhuang et al, 2024). Nonetheless, Avenio software missed two *EML4*::*ALK* transcript fusions. The missed calls by Avenio software could be linked to clustered breakpoint regions, where multiple rearrangements occur in close genomic proximity. Such "clustered breakpoints" introduce complexity in the rearrangement landscape that can hinder accurate fusion calling, particularly in pipelines with conservative filtering thresholds (van Belzen et al, 2021; Lin et al, 2022). From a computational perspective, these regions pose significant challenges for variant callers, especially those with rigid heuristics or simplistic breakpoint models (Onishi-Seebacher et al, 2011). In the case of fusion genes, clustered breakpoints can lead to multiple potential junctions between the same gene partners, complicating read alignment and junction resolution; ambiguous or conflicting evidence (e.g., split-reads supporting multiple alternative breakpoints) and reduced mapping confidence due to repetitive or low-complexity sequence near the junction (Onishi-Seebacher et al, 2011; Nardone et al, 2025). Tools like LUMPY, which utilize discordant paired-end reads in combination with split-read evidence and probabilistic clustering, are inherently more tolerant of such complexity (Layer et al, 2014; Sarwal et al, 2022). Furthermore, the potential for false-negative results in DNA-based sequencing can be higher when fusion breakpoints reside within extensive intron regions containing repetitive elements or low complexity regions (Davies et al, 2019). Additionally, as the Avenio software commercial pipeline is closed we cannot exclude that it can use stricter thresholds for reporting, including minimum supporting read counts, confidence scores, and internal blacklists for problematic regions. This may contribute to missed fusions, even when signal exists in the raw data, as previously reported by Maansson and collaborators (2023) that compared an open-source tool (DNAfusion) with Avenio software and observed that the sensitivity for detecting *EML4*::*ALK* fusions was significantly lower for the latter (Maansson et al, 2023).

When focusing on CNV detection, particularly gene amplifications, the performance of the CNVkit tool was evaluated at the *EGFR*, *MET*, and *ERBB2* loci, using a copy number threshold of ≥6 (Kowalewski et al, 2025). The observed sensitivity was low, indicating that a substantial fraction of clinically relevant amplifications was not detected. Low sensitivity in somatic CNV detection from sequencing data has been widely reported and can be attributed to several factors, including low tumour purity, subclonal copy number events, limited sequencing depth, and the absence of matched normal samples, all of which reduce the signal-to-noise ratio of read-depth–based methods. Large-scale whole-genome and whole-exome sequencing studies have demonstrated that CNV calling algorithms exhibit markedly variable performance depending on tumour purity, ploidy, and coverage, with sensitivity declining sharply in low-purity or heterogeneous samples (Rieber et al, 2017; Wang et al, 2025). In addition, although CNVkit is specifically designed for targeted sequencing data and leverages both on- and off-target reads with bias-correction strategies, showing good concordance with array CGH and competitive performance even in tumour-only analyses, its performance in hybrid-capture panels remains constrained by residual coverage variability and limited resolution; these limitations may be further accentuated in the present analysis by the absence of matched or pooled normal reference samples, potentially reducing sensitivity for detecting focal or high-level amplifications (Talevich et al, 2016; Kato et al, 2023). Taken together, the low sensitivity observed in this analysis is therefore not unexpected and is consistent with the known methodological limitations of panel-based CNV detection. These findings argue against relying solely on targeted sequencing–derived CNV calls in a diagnostic setting and orthogonal validation methods, such as fluorescence in situ hybridization (FISH) or array-based copy number profiling, should be used when required.

5.3. *BRCA2* c.156_157insAlu detection tools performance on DNA sequencing data

Detection of MEIs remains a known blind spot in short-read SV calling pipelines. Most general-purpose SV callers are optimized for canonical large-scale rearrangements (e.g., deletions, duplications, translocations) and are ill-suited to capture small insertions (~300 bp) of repetitive elements like Alu. MEIs frequently occur in low-complexity regions, lack unique flanking sequences, and often fail standard filtering thresholds due to ambiguous alignment, imprecise breakpoints, or low mapping quality (Chu et al, 2021; Hayashi et al, 2025).

To address this and given the frequency of the *BRCA2* c.156_157insAlu Portuguese founder variant (Peixoto et al, 2009) and lack of methods to detect this germline alteration in tumour DNA, we implemented a custom Python-based script targeting the Alu insertion in *BRCA2* exon 3. This variant-specific approach achieved 100% sensitivity and specificity, correctly identifying all six germline-positive tumour samples and returning true negatives across 393 germline-negative samples. Its performance underscores the utility of locus-specific tools for known insertions, particularly in clinical contexts where founder or recurrent variants are well-characterized (Bujakowska et al, 2015; Won et al, 2021). By narrowing analysis to a predefined region, the method can detect insertion-supporting reads even in degraded or low-input FFPE-derived DNA.

However, this method specificity is a major limitation as it cannot detect novel or differently positioned Alu insertions and depends on exact breakpoint knowledge. It is thus valuable for targeted screening, not for discovery. To enable genome-wide MEI detection, we applied RetroSeq, a tool designed to identify non-reference MEIs using discordant read pairs and split-read signals aligned to known transposable elements (Keane et al, 2013). Using this approach, RetroSeq detected the *BRCA2* insertion in four out of six positive samples. While more flexible than the custom breakpoint-driven script, RetroSeq was originally developed and optimized for WGS data and assumes sufficient and uniform coverage to capture MEI-supporting reads (Keane et al, 2013). Its application to targeted sequencing panels is therefore inherently challenging, as off-target coverage is limited and uneven. In the present dataset, MEI detection is further constrained by the use of FFPE tumour samples, which are known to exhibit DNA fragmentation and sequencing artefacts. The reduced sensitivity observed for RetroSeq likely reflects a combination of insufficient coverage, sample degradation, and alignment ambiguity in repetitive genomic region, limitations that have been widely reported for short-read MEI detection in targeted or degraded samples (Wildschutte et al, 2015).

Manta failed to detect the *BRCA2* c.156_157insAlu germline variant in any of the tumour samples. This supports previous observations that this SV caller struggles with MEI detection (Demidov et al, 2021; Wijngaard et al, 2024). While Manta integrates paired-end, split-read, and local assembly evidence to call SVs, its algorithms rely on mappable, unique breakpoints, which are typically absent in MEIs (Demidov et al, 2021). Furthermore, its detection of insertions <1 kb depends on adequate coverage and clean breakpoint junctions, conditions not always met in panel sequencing of FFPE tumours. Even if insertion-supporting reads are present, Manta may flag them as ambiguous or imprecise, leading to filtering under default quality thresholds (Demidov et al, 2021;

Wijngaard et al, 2024). Nonetheless, Manta remains effective for larger deletions, duplications, and translocations, particularly in well-covered, unique genomic regions (Chen et al, 2015).

Taking together, our results indicate that generalist SV callers often miss highly specific variant types unless they are supplemented with dedicated tools. For MEIs, variant-specific scripts or MEI-focused tools such as RetroSeq could be necessary complements, especially in clinical sequencing panels where insertions may have diagnostic or therapeutic implications. These results also emphasize a fundamental limitation of short-read SV detection pipelines in capturing repetitive insertions, even with high read depths (Balachandran et al, 2022). Looking forward, long-read sequencing technologies offer a promising solution. Their ability to span repetitive regions and directly sequence through MEIs enables more accurate detection of insertion events that are inaccessible to short-read methods (Jain et al, 2018; Fernández-Suárez et al, 2024). However, their integration into routine diagnostics remains limited by cost, throughput, and computational and bioinformatics challenges (Moustakli et al, 2025). Until these technologies become standard, a hybrid approach, combining targeted detection scripts, and MEI-specific callers represents the most efficient strategy for comprehensive MEI identification in cancer genomics.

5.4. Limitations of this study

While our benchmarking study provides valuable insights into the performance of various SV detection tools, several limitations must be acknowledged when interpreting these results:

- Reliance on commercial pipelines as reference standards: we used proprietary tools such as DRAGEN RNA app (Illumina) and AVENIO software (Roche) as reference standards. These are closed-source platforms, which do not disclose internal parameters and cannot be modified by the user. And as they themselves can produce false positives or false negatives, our performance metrics should be interpreted as relative to these reference tools, rather than as absolute measures of accuracy.
- Restricted genomic scope of panel-based sequencing: our analyses were conducted using targeted panel sequencing. SVs located outside of the targeted regions are inherently undetectable in this context. This constrained coverage

may lead to an overestimation of tool performance compared to whole-genome sequencing, which provides more comprehensive SV detection.

- Absence of matched normal samples: the use of only tumour-only algorithms precludes confident differentiation between somatic and germline variants. This limitation increases the likelihood of misclassifying inherited variants or sequencing artifacts as somatic events. While this reflects common practice in clinical settings, it underscores the importance of integrating germline variant databases for effective filtering.
- Potential for subjectivity in manual review: variant calls, particularly discordant or borderline cases, were manually reviewed in IGV. Although this step enhances interpretative accuracy, it introduces an element of subjectivity and reduces reproducibility. Moreover, manual inspection limits the scalability and automation of the analysis pipeline.
- Limited sample size for certain variant classes: some variant categories, such as MEIs, were represented by a small number of positive cases (e.g., six *BRCA2* c.156_157insAlu variants). This restricts statistical power and limits the robustness of performance assessments for these variant types.

6. Conclusion

Our comparative benchmarking study of SV detection tools in targeted short read sequencing data reveals important disparities in performance across algorithms, particularly when addressing clinically actionable variants. While RNA-sequencing based callers such as Arriba and STAR-Fusion demonstrate high sensitivity in detecting transcript-level rearrangements, they vary in specificity and often produce discordant calls requiring careful validation. DNA-based tools such as DELLY, LUMPY, and Manta perform well for canonical SVs like deletions and fusions but show limited effectiveness for CNVs and fail to reliably detect specific variants like MEIs. In such cases, custom variant-specific scripts can provide high sensitivity when the insertion site is known *a priori*.

These findings indicate that no single bioinformatic tool or strategy is sufficient to capture the full spectrum of SV with both high sensitivity and specificity. Short-read sequencing data imposes significant technical limitations, particularly in repetitive or poorly mappable regions, necessitating the integration of multiple detection approaches tailored to specific variant classes.

To address these limitations and improve diagnostic robustness, we recommend the development and deployment of hybrid SV detection pipelines that incorporate complementary callers. For example, tools optimized for read-pair and split-read analysis may be integrated with RNA-specific fusion callers and MEI-focused scripts. Manual review and cross-validation with orthogonal techniques, such as FISH, RT-PCR, or expression analysis, remain essential, particularly for variants with potential clinical impact. Variant classification should be guided by curated knowledgebases (e.g., OncoKB, CIViC, ClinVar) to prioritize actionable findings in the clinical context.

The careful selection, testing, and integration of SV detection tools, combined with targeted validation strategies and ongoing performance benchmarking, will enable more accurate and clinically meaningful SV discovery in cancer genomics. This is especially critical in precision oncology, where missed or misclassified events can directly impact therapeutic decisions and patient outcomes.

References

Ahlers CM, Figg WD. ETS-TMPRSS2 fusion gene products in prostate cancer. Cancer Biol Ther. 2006 Mar;5(3):254-5.

Aydin SK, Yilmaz KC, Acar A. Benchmarking long-read structural variant calling tools and combinations for detecting somatic variants in cancer genomes. Sci Rep. 2025 Mar 13;15(1):8707.

Balachandran P, Walawalkar IA, Flores JI, Dayton JN, Audano PA, Beck CR. Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun. 2022 Nov 19;13(1):7115.

Benayed R, Offin M, Mullaney K, Sukhadia P, Rios K, Desmeules P, Ptashkin R, Won H, Chang J, Halpenny D, Schram AM, Rudin CM, Hyman DM, Arcila ME, Berger MF, Zehir A, Kris MG, Drilon A, Ladanyi M. High Yield of RNA Sequencing for Targetable Kinase Fusions in Lung Adenocarcinomas with No Mitogenic Driver Alteration Detected by DNA Sequencing and Low Tumor Mutation Burden. Clin Cancer Res. 2019 Aug 1;25(15):4712-4722.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

Bujakowska KM, White J, Place E, Consugar M, Comander J. Efficient In Silico Identification of a Common Insertion in the MAK Gene which Causes Retinitis Pigmentosa. PLoS One. 2015 Nov 11;10(11):e0142614.

Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019 Jul 19;10(1):3240.

Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113-20.

Cappello F, Angerilli V, Munari G, Ceccon C, Sabbadin M, Pagni F, Fusco N, Malapelle U, Fassan M. FFPE-Based NGS Approaches into Clinical Practice: The Limits of Glory from a Pathologist Viewpoint. J Pers Med. 2022 May 5;12(5):750.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220-2.

Cheong TC, Jang A, Wang Q, Leonardi GC, Ricciuti B, Alessi JV, Di Federico A, Awad MM, Lehtinen MK, Harris MH, Chiarle R. Mechanistic patterns and clinical implications of oncogenic tyrosine kinase fusions in human cancers. Nat Commun. 2024 Jun 14;15(1):5110.

Choo ZN, Behr JM, Deshpande A, Hadi K, Yao X, Tian H, Takai K, Zakusilo G, Rosiene J, Da Cruz Paula A, Weigelt B, Setton J, Riaz N, Powell SN, Busam K, Shoushtari AN, Ariyan C, Reis-Filho J, de Lange T, Imieliński M. Most large structural variants in cancer genomes can be detected without long reads. Nat Genet. 2023 Dec;55(12):2139-2148.

Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, Park PJ. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat Commun. 2021 Jun 22;12(1):3836.

Cocco E, Scaltriti M, Drilon A. NTRK fusion-positive cancers and TRK inhibitor therapy. Nat Rev Clin Oncol. 2018 Dec;15(12):731-747.

Cohen D, Hondelink LM, Solleveld-Westerink N, Uljee SM, Ruano D, Cleton-Jansen AM, von der Thüsen JH, Ramai SRS, Postmus PE, Graadt van Roggen JF, Hoppe BPC, Clahsen PC, Maas KW, Ahsmann EJM, Ten Heuvel A, Smedts F, van Rossem RN, van Wezel T. Optimizing Mutation and Fusion Detection in NSCLC by Sequential DNA and RNA Sequencing. J Thorac Oncol. 2020 Jun;15(6):1000-1014.

Cosenza MR, Rodriguez-Martin B, Korbel JO. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. Annu Rev Genomics Hum Genet. 2022 Aug 31;23:123-152.

Creason A, Haan D, Dang K, Chiotti KE, Inkman M, Lamb A, Yu T, Hu Y, Norman TC, Buchanan A, van Baren MJ, Spangler R, Rollins MR, Spellman PT, Rozanov D, Zhang

J, Maher CA, Caloian C, Watson JD, Uhrig S, Haas BJ, Jain M, Akeson M, Ahsen ME; SMC-RNA Challenge Participants; Stolovitzky G, Guinney J, Boutros PC, Stuart JM, Ellrott K. A community challenge to evaluate RNA-seq, fusion detection, and isoform quantification methods for cancer discovery. Cell Syst. 2021 Aug 18;12(8):827-838.e5.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008.

Davies KD, Aisner DL. Wake Up and Smell the Fusions: Single-Modality Molecular Testing Misses Drivers. Clin Cancer Res. 2019 Aug 1;25(15):4586-4588.

Demidov G, Park J, Armeanu-Ebinger S, Roggia C, Faust U, Cordts I, Blandfort M, Haack TB, Schroeder C, Ossowski S. Detection of mobile elements insertions for routine clinical diagnostics in targeted sequencing data. Mol Genet Genomic Med. 2021 Dec;9(12):e1807.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21.

Dumbrava EEI, Balaji K, Raghav K, Hess K, Javle M, Blum-Murphy M, Ajani J, Kopetz S, Broaddus R, Routbort M, Demirhan M, Zheng X, Pant S, Tsimberidou AM, Subbiah V, Hong DS, Rodon J, Shaw KM, Piha-Paul SA, Meric-Bernstam F. Targeting *ERBB2* (*HER2*) Amplification Identified by Next-Generation Sequencing in Patients With Advanced or Metastatic Solid Tumors Beyond Conventional Indications. JCO Precis Oncol. 2019 Oct 21;3:PO.18.00345.

Fernández-Suárez E, González-Del Pozo M, Méndez-Vidal C, Martín-Sánchez M, Mena M, de la Morena-Barrio B, Corral J, Borrego S, Antiñolo G. Long-read sequencing improves the genetic diagnosis of retinitis pigmentosa by identifying an Alu retrotransposon insertion in the EYS gene. Mob DNA. 2024 May 4;15(1):9.

Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. Brief Bioinform. 2021 May 20;22(3):bbaa056.

Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019 Oct 21;20(1):213.

Hayashi T, Mizobuchi K, Higa N, Maki H, Kuniyoshi K, Ikeda Y, Kondo M, Saitsu H. Limitations of short-read NGS in detecting *RP1 Alu* insertions: a case emphasizing Sanger confirmation. Ophthalmic Genet. 2025 Nov 25:1-5.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82-93.

International Cancer Genome Consortium; Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigó R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, López-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigó R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, López-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolás P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G,

Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhortchouk E, Banks RE, Uhlén M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporté I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clément B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifenberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, López-Otín C, Estivill X, Guigó R, de Sanjosé S, Piris MA, Montserrat E, González-Díaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H. International network of cancer genome projects. Nature. 2010 Apr 15;464(7291):993-8.

Isaic A, Motofelea N, Hoinoiu T, Motofelea AC, Leancu IC, Stan E, Gheorghe SR, Dutu AG, Crintea A. Next-Generation Sequencing: A Review of Its Transformative Impact on Cancer Diagnosis, Treatment, and Resistance Management. Diagnostics (Basel). 2025 Sep 23;15(19):2425.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A,

Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr;36(4):338-345.

Jeck WR, Iafrate AJ, Nardi V. Nanopore Flongle Sequencing as a Rapid, Single-Specimen Clinical Test for Fusion Detection. J Mol Diagn. 2021 May;23(5):630-636.

Joe S, Park JL, Kim J, Kim S, Park JH, Yeo MK, Lee D, Yang JO, Kim SY. Comparison of structural variant callers for massive whole-genome sequence data. BMC Genomics. 2024 Mar 28;25(1):318.

Johansson B, Mertens F, Schyman T, Björk J, Mandahl N, Mitelman F. Most gene fusions in cancer are stochastic events. Genes Chromosomes Cancer. 2019 Sep;58(9):607-611.

Lange M, Begolli R, Giakountis A. Non-Coding Variants in Cancer: Mechanistic Insights and Clinical Potential for Personalized Medicine. Noncoding RNA. 2021 Aug 2;7(3):47.

Laver TW, De Franco E, Johnson MB, Patel KA, Ellard S, Weedon MN, Flanagan SE, Wakeling MN. SavvyCNV: Genome-wide CNV calling from off-target reads. PLoS Comput Biol. 2022 Mar 16;18(3):e1009940.

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60.

Li H, Meng L, Wang H, Cui L, Sheng H, Zhao P, Hong S, Du X, Yan S, Xing Y, Feng S, Zhang Y, Fang H, Bai J, Liu Y, Lan S, Liu T, Guan Y, Xia X, Yi X, Cheng Y. Precise identification of somatic and germline variants in the absence of matched normal samples. Brief Bioinform. 2024 Nov 22;26(1):bbae677.

Li W, Wan R, Guo L, Chang G, Jiang D, Meng L, Ying J. Reliability analysis of exonic-breakpoint fusions identified by DNA sequencing for predicting the efficacy of targeted therapy in non-small cell lung cancer. BMC Med. 2022 May 10;20(1):160.

Lin J, Yang X, Kosters W, Xu T, Jia Y, Wang S, Zhu Q, Ryan M, Guo L, Zhang C; Human Genome Structural Variation Consortium; Lee C, Devine SE, Eichler EE, Ye K. Mako: A Graph-based Pattern Growth Approach to Detect Complex Structural Variants. Genomics Proteomics Bioinformatics. 2022 Feb;20(1):205-218.

Lin JJ, Zhu VW, Yoda S, Yeap BY, Schrock AB, Dagogo-Jack I, Jessop NA, Jiang GY, Le LP, Gowen K, Stephens PJ, Ross JS, Ali SM, Miller VA, Johnson ML, Lovly CM, Hata

AN, Gainor JF, Iafrate AJ, Shaw AT, Ou SI. Impact of EML4-ALK Variant on Resistance Mechanisms and Clinical Outcomes in ALK-Positive Lung Cancer. J Clin Oncol. 2018 Apr 20;36(12):1199-1206.

Liu B, Conroy JM, Morrison CD, Odunsi AO, Qin M, Wei L, Trump DL, Johnson CS, Liu S, Wang J. Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget. 2015 Mar 20;6(8):5477-89.

Maansson CT, Andersen ER, Ulhoi MP, Meldgaard P, Sorensen BS. DNAfusion: an R/Bioconductor package for increased sensitivity of detecting gene fusions in liquid biopsies. BMC Bioinformatics. 2023 Apr 4;24(1):131.

Moasser MM. Targeting the function of the HER2 oncogene in human cancer therapeutics. Oncogene. 2007 Oct 11;26(46):6577-92.

Moustakli E, Christopoulos P, Potiris A, Zikopoulos A, Mavrogianni D, Karampas G, Kathopoulis N, Anagnostaki I, Domali E, Tzallas AT, Drakakis P, Stavros S. Long-Read Sequencing and Structural Variant Detection: Unlocking the Hidden Genome in Rare Genetic Disorders. Diagnostics (Basel). 2025 Jul 17;15(14):1803.

Nardone GG, Andrioletti V, Santin A, Morgan A, Spedicati B, Concas MP, Gasparini P, Girotto G, Limongelli I. A Hitchhiker Guide to Structural Variant Calling: A Comprehensive Benchmark Through Different Sequencing Technologies. Biomedicines. 2025 Aug 9;13(8):1949.

Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, Shih DJ, Hovestadt V, Zapatka M, Sturm D, Jones DT, Kool M, Remke M, Cavalli FM, Zuyderduyn S, Bader GD, VandenBerg S, Esparza LA, Ryzhova M, Wang W, Wittmann A, Stark S, Sieber L, Seker-Cin H, Linke L, Kratochwil F, Jäger N, Buchhalter I, Imbusch CD, Zipprich G, Raeder B, Schmidt S, Diessl N, Wolf S, Wiemann S, Brors B, Lawerenz C, Eils J, Warnatz HJ, Risch T, Yaspo ML, Weber UD, Bartholomae CC, von Kalle C, Turányi E, Hauser P, Sanden E, Darabi A, Siesjö P, Sterba J, Zitterbart K, Sumerauer D, van Sluis P, Versteeg R, Volckmann R, Koster J, Schuhmann MU, Ebinger M, Grimes HL, Robinson GW, Gajjar A, Mynarek M, von Hoff K, Rutkowski S, Pietsch T, Scheurlen W, Felsberg J, Reifenberger G, Kulozik AE, von Deimling A, Witt O, Eils R, Gilbertson RJ, Korshunov A, Taylor MD, Lichter P, Korbel JO, Wechsler-Reya RJ, Pfister SM. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014 Jul 24;511(7510):428-34.

Olivucci G, Iovino E, Innella G, Turchetti D, Pippucci T, Magini P. Long read sequencing on its way to the routine diagnostics of genetic diseases. Front Genet. 2024 Mar 6;15:1374860.

Onishi-Seebacher M, Korbel JO. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. Bioessays. 2011 Nov;33(11):840-50.

Peixoto A, Santos C, Rocha P, Pinheiro M, Príncipe S, Pereira D, Rodrigues H, Castro F, Abreu J, Gusmão L, Amorim A, Teixeira MR. The c.156_157insAlu BRCA2 rearrangement accounts for more than one-fourth of deleterious BRCA mutations in northern/central Portugal. Breast Cancer Res Treat. 2009 Mar;114(1):31-8.

Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet. 2015 Apr 13;6:138.

Popic V, Rohlicek C, Cunial F, Hajirasouliha I, Meleshko D, Garimella K, Maheshwari A. Cue: a deep-learning framework for structural variant discovery and genotyping. Nat Methods. 2023 Apr;20(4):559-568.

Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W, Roepman P, Voda M, Bloemendal HJ, Tjan-Heijnen VCG, van Herpen CML, Labots M, Witteveen PO, Smit EF, Sleijfer S, Voest EE, Cuppen E. Pan-cancer whole-genome analyses of metastatic solid tumors. Nature. 2019 Nov;575(7781):210-216.

Pócza T, Grolmusz VK, Papp J, Butz H, Patócs A, Bozsik A. Germline Structural Variations in Cancer Predisposition Genes. Front Genet. 2021 Apr 14;12:634217.

Kato K, Okami J, Nakamura H, Honma K, Sato Y, Nakamura S, Kukita Y, Nakatsuka SI, Higashiyama M. Analytical Performance of a Highly Sensitive System to Detect Gene Variants Using Next-Generation Sequencing for Lung Cancer Companion Diagnostics. Diagnostics (Basel). 2023 Apr 19;13(8):1476.

Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013 Feb 1;29(3):389-90.

Kowalewski A, Siemanowski-Hrach J, Stehle T, Rehker J, Siebolts U, Merkelbach-Bruse S, Heydt C. NGS coverage accurately predicts *MET* and *HER2* (*ERBB2*) gene amplifications in a real-world non-small cell lung cancer cohort. Front Oncol. 2025 Jul 29;15:1618509.

Randolph TR. Chronic myelocytic leukemia--Part I: History, clinical presentation, and molecular biology. Clin Lab Sci. 2005 Winter;18(1):38-48. Erratum in: Clin Lab Sci. 2005 Summer;18(3):149.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333-i339.

Rieber N, Bohnert R, Ziehm U, Jansen G. Reliability of algorithmic somatic copy number alteration detection from targeted capture data. Bioinformatics. 2017 Sep 15;33(18):2791-2798.

Rowley JD. Chromosomal patterns in myelocytic leukemia. N Engl J Med. 1973 Jul 26;289(4):220-1.

Sarwal V, Niehus S, Ayyala R, Kim M, Sarkar A, Chang S, Lu A, Rajkumar N, Darfci-Maher N, Littman R, Chhugani K, Soylev A, Comarova Z, Wesel E, Castellanos J, Chikka R, Distler MG, Eskin E, Flint J, Mangul S. A comprehensive benchmarking of WGS-based deletion structural variant callers. Brief Bioinform. 2022 Jul 18;23(4):bbac221.

Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023 Jul 13;12(7):997.

Sepúlveda-Hermosilla G, Freire M, Blanco A, Cáceres J, Lizana R, Ramos L, Assar Cuevas R, Ampuero D, Aren O, Chernilo S, Spencer ML, Bernal G, Flores J, Rasse G, Sánchez C, Marcelain K, Rivas S, Branco GP, Galli de Amorim M, Nunes DN, Dias-Neto E, Freitas HC, Fernández C; NIRVANA team; Pérez P, Armisén R. Concordance Analysis of ALK Gene Fusion Detection Methods in Patients with Non-Small-Cell Lung Cancer from Chile, Brazil, and Peru. J Mol Diagn. 2021 Sep;23(9):1127-1137.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. DNA sequencing at 40: past, present and future. Nature. 2017 Oct 19;550(7676):345-353.

Shoshani O, Brunner SF, Yaeger R, Ly P, Nechemia-Arbely Y, Kim DH, Fang R, Castillon GA, Yu M, Li JSZ, Sun Y, Ellisman MH, Ren B, Campbell PJ, Cleveland DW. Chromothripsis drives the evolution of gene amplification in cancer. Nature. 2021 Mar;591(7848):137-141.

Shrestha M, Blay S, Liang S, Swanson D, Lerner-Ellis J, Dickson B, Wong A, Charames GS. Improving RNA Fusion Call Confidence and Reliability in Molecular Diagnostic Testing. J Mol Diagn. 2023 Jun;25(6):320-330.

Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007 Aug 2;448(7153):561-6.

Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016 Apr 21;12(4):e1004873.

Terraf P, Pareja F, Brown DN, Ceyhan-Birsoy O, Misyura M, Rana S, O'Reilly E, Carlo MI, Aghajanian C, Liu Y, Derakhshan F, Jayakumaran G, Weigelt B, Walsh M, Stadler Z, Offit K, Ladanyi M, Robson M, Zehir A, Reis-Filho JS, Mandelker D. Comprehensive assessment of germline pathogenic variant detection in tumor-only sequencing. Ann Oncol. 2022 Apr;33(4):426-433.

Tubio JM. Somatic structural variation and cancer. Brief Funct Genomics. 2015 Sep;14(5):339-51.

Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, Toprak UH, Neumann O, Stenzinger A, Scholl C, Fröhling S, Brors B. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021 Mar;31(3):448-460.

van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. NPJ Precis Oncol. 2021 Mar 2;5(1):15.

Vazquez-Mena O, Medina-Martinez I, Juárez-Torres E, Barrón V, Espinosa A, Villegas-Sepulveda N, Gómez-Laguna L, Nieto-Martínez K, Orozco L, Roman-Basaure E, Muñoz Cortez S, Borges Ibañez M, Venegas-Vega C, Guardado-Estrada M, Rangel-López A, Kofman S, Berumen J. Amplified genes may be overexpressed, unchanged, or downregulated in cervical cancer cell lines. PLoS One. 2012;7(3):e32667.

Vishwakarma R, McManus KJ. Chromosome Instability; Implications in Cancer Development, Progression, and Clinical Outcomes. Cancers (Basel). 2020 Mar 29;12(4):824.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546-58.

Wang H, Li ZW, Ou Q, Wu X, Nagasaka M, Shao Y, Ou SI, Yang Y. NTRK fusion positive colorectal cancer is a unique subset of CRC with high TMB and microsatellite instability. Cancer Med. 2022 Jul;11(13):2541-2549.

Wang N, Tao ZY, Wu T, Wang J, Wang W, Shi H, Liu XS. Benchmarking copy number variation detection with low-coverage whole-genome sequencing. Brief Bioinform. 2025 Aug 31;26(5):bbaf514.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wijngaard R, Demidov G, O'Gorman L, Corominas-Galbany J, Yaldiz B, Steyaert W, de Boer E, Vissers LELM, Kamsteeg EJ, Pfundt R, Swinkels H, den Ouden A, Te Paske IBAW, de Voer RM, Faivre L, Denommé-Pichon AS, Duffourd Y, Vitobello A, Chevarin M, Straub V, Töpf A, van der Kooi AJ, Magrinelli F, Rocca C, Hanna MG, Vandrovcova J; Solve-RD consortium; Ossowski S, Laurie S, Gilissen C. Mobile element insertions in rare diseases: a comparative benchmark and reanalysis of 60,000 exome samples. Eur J Hum Genet. 2024 Feb;32(2):200-208.

Wildschutte JH, Baron A, Diroff NM, Kidd JM. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res. 2015 Dec 2;43(21):10292-307.

Won D, Hwang JY, Shim Y, Byeon SH, Lee J, Lee CS, Kim M, Lim HT, Choi JR, Lee ST, Han J. In Silico identification of a common mobile element insertion in exon 4 of RP1. Sci Rep. 2021 Jun 28;11(1):13381.

Xia Z, Xiang W, Wang Q, Li X, Li Y, Gao J, Tang T, Yang C, Cui Y. CSV-Filter: a deep learning-based comprehensive structural variant filtering method for both short and long reads. Bioinformatics. 2024 Sep 2;40(9):btae539.

Yang SR, Aypar U, Rosen EY, Mata DA, Benayed R, Mullaney K, Jayakumaran G, Zhang Y, Frosina D, Drilon A, Ladanyi M, Jungbluth AA, Rekhtman N, Hechtman JF. A Performance Comparison of Commonly Used Assays to Detect RET Fusions. Clin Cancer Res. 2021 Mar 1;27(5):1316-1328.

Yi D, Nam JW, Jeong H. Toward the functional interpretation of somatic structural variations: bulk- and single-cell approaches. Brief Bioinform. 2023 Sep 20;24(5):bbad297.

Zhuang G, Zhang X, Du W, Xu L, Ma J, Luo H, Tang H, Wang W, Wang P, Li M, Yang X, Wu D, Fang S. A benchmarking framework for the accurate and cost-effective

detection of clinically-relevant structural variants for cancer target identification and diagnosis. J Transl Med. 2024 Jan 16;22(1):65.

Attachments

Attachment 1: TruSight RNA Fusion Panel Gene List

Gene	Gene	Gene	Gene	Gene	Gene	Gene	Gene	Gene	Gene
ABI1	BRWD3	COL1A2	ERLIN2	HHEX	LNP1	NFIB	POU5F1	SLC34A2	TGFBR3
ABL1	BTBD18	COL6A3	ESR1	HIP1	LPP	NGF	PPAP2B	SLC45A3	THADA
ABL2	BTG1	COX6C	ETS1	HIPK1	LPXN	NGFR	PPARG	SLCO1B3	THRAP3
ACACA	C11orf1	CPSF6	ETV1	HIST1H4I	LRMP	NIN	PPARGC1A	SMAP1	TIRAP
ACE	C11orf95	CRADD	ETV4	HLF	LRRC37B	NIPBL	PPFIBP1	SMARCA5	TLX1
ACER1	C2CD2L	CREB1	ETV5	HMGA2	LTBP1	NKX2-5	PPP2R1B	SMARCB1	TLX3
ACKR3	C3orf27	CREB3L1	ETV6	HNF1A	LYL1	NONO	PRCC	SNHG5	TMPRSS2
ACSL6	CAMTA1	CREB3L2	EWSR1	HOXA10	MACROD1	NOTCH1	PRDM16	SORBS2	TNFRSF17
ADD3	CAPRIN1	CREBBP	EZR	HOXA11	MAF	NPM1	PRKACA	SORT1	TOP1
AFF1	CARS	CRLF2	FAM19A2	HOXA13	MAFB	NR4A3	PRKAR1A	SP3	TOP2B
AFF3	CASC5	CRTC1	FCGR2B	HOXA9	MALT1	NR6A1	PRKG2	SPECC1	TP53BP1
AFF4	CASP7	CSF1	FCRL4	HOXC11	MAML2	NSD1	PRRX2	SPTBN1	TPM3
AGR3	CBFA2T3	CSF1R	FEN1	HOXC13	MAPRE1	NT5C2	PSIP1	SQSTM1	TPM4
AHI1	CBFB	CTDSP2	FEV	HOXD11	MBNL1	NTF3	PSMD2	SRF	TRHDE
AHRR	CBL	CTNNB1	FGF8	HOXD13	MBTD1	NTF4	PTPRR	SRSF3	TRIM24
ALK	CCAR2	CUX1	FGFR1	HSP90AA1	MDS2	NTRK1	PVT1	SS18	TRIP11
ANKRD28	CCDC28A	DAB2IP	FGFR1OP	ID4	MEAF6	NTRK2	RABEP1	SS18L1	TRPS1
AR	CCDC6	DACH1	FGFR1OP2	IKZF1	MECOM	NTRK3	RAD51B	SSBP2	USP16
ARHGAP20	CCDC88C	DACH2	FGFR2	IL2	MGEA5	NUMA1	RAF1	SSX1	USP42
ARHGAP26	CCNB1IP1	DDIT3	FGFR3	IL21R	MKL1	NUP107	RANBP2	SSX2	USP6
ARNT	CCNB3	DDX10	FGFR4	IL3	MKL2	NUP214	RAP1GDS1	SSX4	VGLL3
ASPSCR1	CCND1	DDX20	FHIT	INPP5D	MLF1	NUP98	RARA	ST6GAL1	WASF2
ASTN2	CCND2	DEK	FIP1L1	IQCG	MLLT1	NUTM1	RBM15	STAT5B	WDR18
ATF1	CCND3	DMRT1	FLI1	IRF2BP2	MLLT10	NUTM2A	RBM6	STAT6	WDR70
ATIC	CD74	DNAJB1	FLNA	IRF4	MLLT11	NUTM2B	RCOR1	STRN	WHSC1
ATP1B4	CDH11	DPM1	FLT3	IRS4	MLLT3	OFD1	RCSD1	SUGP2	WHSC1L1
AUTS2	CDK5RAP2	DUSP22	FLT3LG	ITK	MLLT4	OLIG2	RET	SUZ12	WSB1
BACH2	CDK6	DUX4	FNBP1	JAK1	MLLT6	OLR1	RHOH	SYK	WT1
BAG4	CDX1	EBF1	FOSB	JAK2	MN1	OMD	RNF213	TACC1	WWTR1
BAIAP2L1	CDX2	EEFSEC	FOSL1	JAZF1	MNX1	P2RY8	ROS1	TACC2	XIAP
BAZ2A	CEBPA	EGFR	FOXO1	KANK1	MSI2	PAPPA	RPL22	TACC3	YAP1
BCAS3	CEBPB	EGR1	FOXO4	KAT6A	MSN	PATZ1	RPN1	TAF15	YTHDF2
BCAS4	CEBPD	EGR2	FOXP1	KAT6B	MUC1	PAX3	RREB1	TAL1	YWHAE
BCL10	CEBPE	EGR3	FRK	KDM5A	MUTYH	PAX5	RRM1	TAL2	ZBTB16
BCL11A	CEP170B	EGR4	FRYL	KIAA1524	MYB	PAX7	RTEL1	TAOK1	ZC3H7A
BCL11B	CEP85L	EIF4A2	FUS	KIF5B	MYBL1	PAX8	RUNX1	TBX15	ZC3H7B
BCL2	CHD6	ELF4	GAS5	KMT2A	MYC	PBX1	RUNX1T1	TCF12	ZFP64
BCL2L1	CHIC2	ELK4	GAS7	KPNB1	MYH11	PCM1	SARNP	TCF3	ZFPM2
BCL3	CHMP2B	ELL	GATA1	KSR1	MYH9	PDE4DIP	SEC31A	TCL1A	ZFYVE19
BCL6	CHST11	ELN	GIT2	LASP1	MYO18A	PDGFB	SEPT2	TCTA	ZMIZ1
BCL9	CIC	EML1	GLI1	LCK	MYO1F	PDGFRA	SEPT5	TEAD1	ZMYM2
BCOR	CIITA	EML4	GOSR1	LCP1	NAB2	PDGFRB	SEPT6	TEAD2	ZMYND11
BCR	CLP1	EP300	GOT1	LGR5	NAPA	PER1	SEPT9	TEAD3	ZNF207
BDNF	CLTC	EP400	GPR128	LHFP	NBEAP1	PHF1	SERPINE1	TEAD4	ZNF384
BICC1	CLTCL1	EPC1	GPR34	LHX2	NBR1	PHF23	SERPINF1	TEC	ZNF444
BIRC3	CMKLR1	EPOR	GRHPR	LHX4	NCOA1	PICALM	SET	TENM1	ZNF521
BIRC6	CNBP	EPS15	GRID1	LINC00598	NCOA2	PIM1	SETBP1	TET1	ZNF585B
BRAF	CNOT2	ERBB3	GTF2I	LINC00982	NCOA3	PLAG1	SFPQ	TFE3	ZNF687
BRD1	CNTRL	ERC1	H2AFX	LMBRD1	NDE1	PML	SH3D19	TFG	
BRD3	COG5	ERCC1	HAS2	LMO1	NF1	POM121	SH3GL1	TFPT	
BRD4	COL1A1	ERG	HEY1	LMO2	NFATC2	POU2AF1	SIK3	TFRC	