An Ensemble Method for Calling and Ranking Somatic Structural Variants Using Long and Short Reads

Walter Gallego Gomez
Department of Control and Computer Engineering,
Politecnico di Torino
Italy
walter.gallego@polito.it

Elena Grassi
Department of Oncology, University of Torino
Italy
Candiolo Cancer Institute - FPO IRCCS
Italy
elena.grassi@ircc.it

Andrea Bertotti
Department of Oncology, University of Torino
Italy
Candiolo Cancer Institute - FPO IRCCS
Italy
andrea.bertotti@ircc.it

Gianvito Urgese
Interuniversity Department of Regional and Urban Studies
and Planning, Politecnico di Torino
Italy
gianvito.urgese@polito.it

Abstract

Structural variants (SVs) are a class of genetic alterations that play a crucial role in cancer development. Detecting somatic SVs is challenging, as it requires distinguishing between germline and somatic events and dealing with subclonal variants and the compresence of tumor and normal cells in patient-derived samples. SVs callers based on single-molecule sequencing technologies have emerged as a powerful tool in detecting SVs, thanks to the ability of long reads to span large genomic regions, allowing the detection of more complex rearrangements. However, these tools are still affected by low precision and/or recall, especially in determining somatic SVs. To overcome these limitations, we propose an ensemble method that combines the results of three long-read variant callers with evidence extracted from accompanying short-read alignments. We evaluate our method on a curated truth set provided by the *Espejo Valle-Inclan* benchmark and show that it can leverage the strengths of each tool while mitigating their weaknesses to produce a ranked list of somatic deletions, useful to prioritize downstream analysis and experimental validation. We also provide insights into the performance of the individual tools and discuss future directions for the extension of our method.

CCS Concepts

• **Applied computing → Bioinformatics**.

Keywords

somatic structural variants (SVs), ensemble methods for somatic SVs calling, somatic SVs ranking, long-read somatic SVs callers

ACM Reference Format:
Walter Gallego Gomez, Elena Grassi, Andrea Bertotti, and Gianvito Urgese. 2024. An Ensemble Method for Calling and Ranking Somatic Structural Variants Using Long and Short Reads. In *The 11th International Conference on Bioinformatics Research and Applications (ICBRA 2024), September 13–15, 2024, Milan, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3700666.3700694

ICBRA 2024, September 13–15, 2024, Milan, Italy

ACM ISBN 979-8-4007-1753-6/24/09
https://doi.org/10.1145/3700666.3700694

1 Introduction

Genomic instability, alongside small somatic alterations, is considered one of the hallmarks of cancer [14]. Despite advancements in our knowledge of their abundance and functional role, with estimates of 55% of the driver events being structural variants (SVs) [4], our ability to detect them from sequencing data is still far from the level of confidence reached for small variants. This is due both to biological issues and to technical limitations, mainly driven by the fact that short-read sequencing is often not able to directly capture these large events [18].

The need to distinguish somatic from germline variants, the presence of subclonal variants, and the compresence of tumor cells alongside normal ones in patients-derived samples are all issues that have been faced while developing calling procedures for small alterations but that are more complex to solve for structural variants [23]. This is mainly due to limits in determining breakpoints at single base levels and in the difficulties faced when aligning short reads to regions of the human genome harboring repetitive elements. Such hindrances could be overcome thanks to single molecule sequencing technologies that result in longer reads (23kb [29] vs 150/300), such as Oxford Nanopore [29] and PacBio [22].

While several efforts have been made to develop long-read aligners and callers for SVs [2], we are still far from having definite gold standards such as GATK/Mutect [19] and DRAGEN [13], with known sensitivity and well-defined, portable, reproducible, and scalable pipelines. Although the majority of the proposed approaches are focused on general single-sample calling, some long-read SV callers have been proposed for the specific task of somatic events detection and can be divided into three main categories: callers that only use a tumor sample, such as Sniffles2 [25] and DeBreak [3] that have special modes to account for the low allele-frequency of somatic events, or SHARC [26] that proposes a pipeline to filter out germline SVs; callers like CAMPHOR [12], that compare the SVs obtained independently from the tumor sample with those obtained from the normal one, to determine which events are somatic; and callers that use both the tumor and normal long reads simultaneously to call somatic events, such as *NanomonSV* [24] and *SAVANA* [7]. Methods that compare reads or SV calls from tumor and normal tissue samples can allow for more precise differentiation of somatic SVs from germline SVs without the need for deep sequencing [2].

Another path adopted to increase the quality of the results is the use of ensemble methods, that have been applied to small alterations [11, 28], as well as to general SVs detection using long reads, such as NextSV [10], a meta-caller that integrates three aligners and three SV callers, and combiSV [6], that combines the results from six SV callers into a call set with increased recall and precision.

Following these ideas, we developed an ensemble method specifically designed for somatic structural variants, aiming at combining the strengths and weaknesses of different algorithms for SVs calling using long reads. Our approach also exploits pieces of evidence coming from short reads, to complement those coming from the most recent and still developing long-read technologies, and results in a ranked list of the detected somatic events, from the ones with most support to the least. This should ease manual examinations of the results and experimental validation efforts. In this initial work, we focused on deletions considering that they are the most abundant well-defined SVs in the majority of cancers [17], and that short-read sequencing can be efficiently exploited to complement the rank defined with long reads. Integrative efforts such as this will be instrumental in reaching the maturity level that we have for small alterations for SVs, supporting future single [1, 21] and pan-cancer efforts to define their overall landscape and functional consequences. The scientific community will then be able to fruitfully integrate them into multi-omics studies.

The remainder of the paper is organized as follows. We first describe the pipeline we have designed, the variant callers we used, and the scripts we implemented for the ensembling, validation, and ranking of deletions. We then present the results obtained by running our pipeline with the *Espejo Valle-Inclan* benchmark, comparing our ranked list to the provided truth set. We close with a discussion of our approach, the performance of the variant callers, and the future developments of our pipeline.

2 Methods

The general workflow of our approach is shown in Figure 1. It requires a tumor-normal pair, sequenced with both long-read and short-read technologies and aligned to the reference genome. The aligned long-read data is used as the input for three long-read variant callers, that produce a set of VCF files. These are then combined into a single VCF using an ensemble approach, and filtered to keep only the deletions. Next, the resulting VCF file goes through a validation step using the aligned short-read data. Finally, the deletions in the VCF file are ranked based on a set of scores calculated from the calling, ensemble, and validation results.

Two software modules have been created, one containing the source code for the scripts developed in this work and one containing the pipeline that integrates these scripts with the third-party software used. Parameters for the custom scripts and for the third-party tools (such as the minimum length of the SVs to call) are managed by the pipeline and can be customized by the user from a configuration file. The scripts have been developed in *Python*, version 3.10.4. We used *pysam* (version 0.22.0) and *cyvcf2* [20] (version 0.30.22) for the manipulation of BAM, FASTA and VCF files. The pipeline has been written in *Nextflow* [5] version 23.04 and *nf-core* [9] version 2.9, using the templates, modules and guidelines provided by the *nf-core* community.

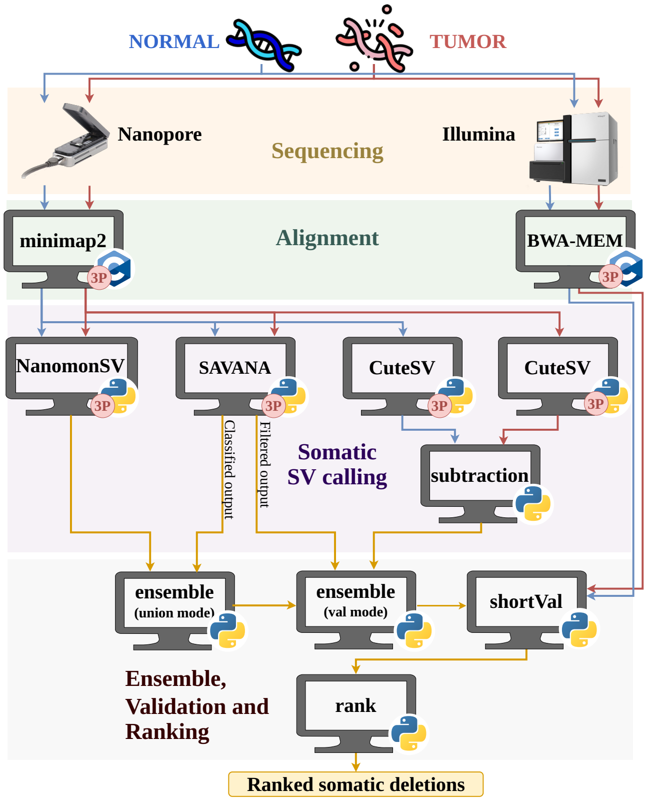

Figure 1: General workflow.
From a tumor-normal pair it generates a set of somatic deletions, ranked according to the evidence found during the calling, ensemble and validation steps. Blue lines represent the flow of normal-sample data, red lines represent the flow of tumor-sample data, and yellow lines that of somatic data. Third-party software is marked with a red circle labeled 3P.

2.1 Alignment

For the alignment of Nanopore long reads to the reference genome we used the epi2me-labs *wf-alignment* workflow, available at GitHub[1]. This workflow uses *minimap2* [16], with the `-x map-ont` option to choose the Nanopore presets. We used version v0.1.3, which ships with version 2.26-r1175 of *minimap2*.

For the alignment of Illumina short reads we used BWA-MEM2 [27] version 2.2.1 with default parameters.

[1] https://github.com/epi2me-labs/wf-alignment

2.2 Somatic variants calling

We used three different tools for the identification of somatic SVs in the tumor-normal pair: *NanomonSV* and *SAVANA* which are specifically designed to use both samples simultaneously, and *CuteSV* which supports only one sample at a time.

NanomonSV is a somatic SV caller for ONT and PacBio data that uses a matched tumor-normal pair. It performs a clustering procedure in which the reads from the tumor sample that presumably cover the same SV are grouped, unless apparent supporting reads in the normal samples are found, in which case the cluster is discarded. The remaining clusters are then refined to improve the accuracy of the breakpoints, and validated to confirm that the putative SV segment sequence is observed in the tumor and not in the normal reads [24]. We installed *NanomonSV* version 0.7.1 and ran the `parse` command in the normal and tumor samples independently. We then ran the `get` command to obtain the somatic SVs. We included the `--use_racon` option to enable the error-correction step, as advised in the documentation, and the `--min_indel_size` option to set the minimum size of the SVs. The rest of the parameters were left as default. Finally, we filtered the results using *bcftools view* to keep only the deletions with `FILTER = "PASS"` set.

SAVANA is a somatic SV caller for long-read data. It takes as input long-read WGS data from a tumor and normal sample pair and scans it to detect split reads and gapped alignments, which are then clustered to define putative SVs. Next, it applies a machine learning-informed set of heuristics to remove false positives arising from mapping errors and sequencing artifacts [7]. We installed *SAVANA* version 1.0.4 and ran it with the `--length` option to set the minimum length of the SVs. The rest of the parameters were left as default. *SAVANA* classifies somatic variants using a random-forest classifier, trained on a range of somatic Oxford Nanopore data labeled with true somatic variants (as determined by supporting Illumina data). During the evaluation of the results, we found that this classifier is too stringent and filters out some true SVs. Therefore we also considered the unclassified output of *SAVANA* and then filtered it using `bcftools view` to keep only the deletions with `TUMOUR_SUPPORT>=3`, `NORMAL_SUPPORT<=1` and `ORIGIN_STARTS_STD_DEV<150`. We end up with two distinct VCF files, that we name *SAVANA-classified* and *SAVANA-filtered*. Their use will be explained in subsection 2.3.

CuteSV is a sensitive, fast, and scalable long-read-based SV detection approach that uses tailored methods to collect the signatures of various types of SVs and employs a clustering-and-refinement method to implement sensitive SV detection [15]. We installed *CuteSV* version 2.0.3 and ran it independently for the normal and tumor samples, using as parameters `--genotype`, `--min_size`, `--max_size 1500000` and `--min_support 2`. We then ran a custom script that compares each SV in the tumor sample against SVs of the same type in the normal sample located within the same contig. If no overlapping SV is found in the normal sample, the SV from the tumor sample is considered somatic. Two SVs are considered to be overlapping if at least 75% of their sequence overlap. We call the resulting procedure *CuteSV-sub*.

2.3 Ensemble of long-read variant callers

Ensembling is performed using a custom script that takes as input two VCF files to produce a new one. For each deletion in the first VCF file, we search for overlapping deletions in the second one. If an overlapping deletion is found, we save its characteristics in the first VCF, using the custom information fields described in Table 1. If multiple deletions from the second VCF file overlap with the same deletion from the first one, we consider only the one with the highest overlap. The overlap criteria is the same as the one used in the *CuteSV-sub* tool.

Table 1: Custom information fields for ensemble

Field name	Description
`<CALLER>_ID`	Unique identifier of the variant
`<CALLER>_POS`	Position of the variant
`<CALLER>_LEN`	Length of the variant
`<CALLER>_TUMOR`	No. of supporting reads in tumor sample
`<CALLER>_NORMAL`	No. of supporting reads in normal sample
`<CALLER>_INDEX`	Index indicating the level of overlap

The script gives the possibility to keep only the deletions found by the first tool, with eventual overlaps (*validation mode*), or to keep the deletions found by either one of the two tools (*union mode*).

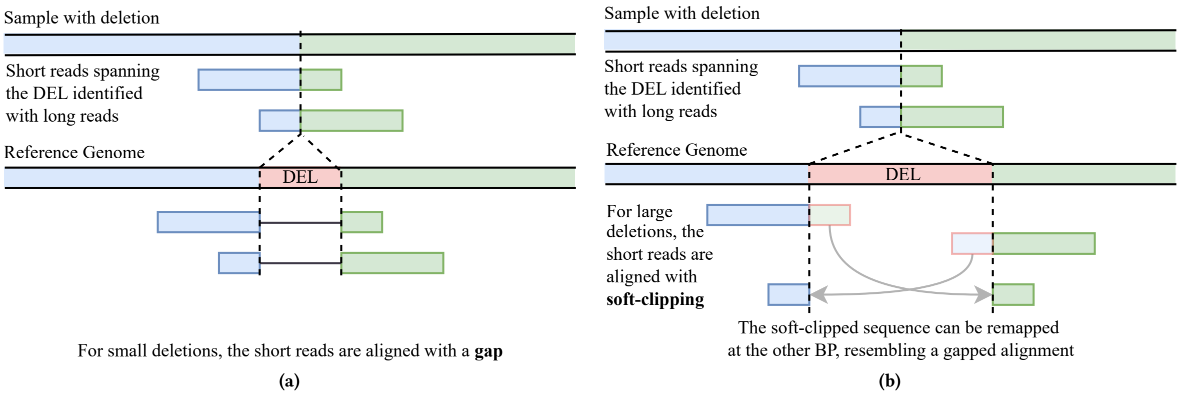
Sample with deletion
Short reads spanning the DEL identified with long reads
Reference Genome
DEL
For small deletions, the short reads are aligned with a **gap**
(a)
Sample with deletion
Short reads spanning the DEL identified with long reads
Reference Genome
DEL
For large deletions, the short reads are aligned with **soft-clipping**
The soft-clipped sequence can be remapped at the other BP, resembling a gapped alignment
(b)

Figure 2: (a) Short deletion resulting in gapped alignment. (b) Long deletion resulting in soft-clipping.

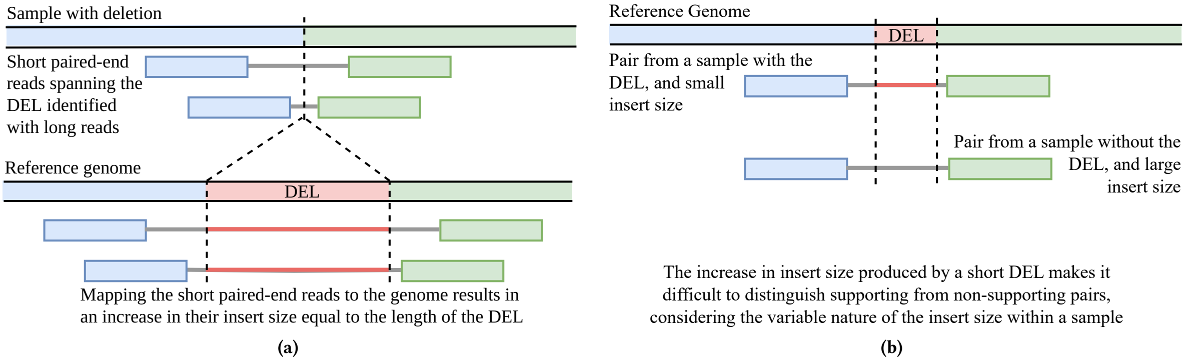
Sample with deletion
Short paired-end reads spanning the DEL identified with long reads
Reference genome
DEL
Mapping the short paired-end reads to the genome results in an increase in their insert size equal to the length of the DEL
(a)
Reference Genome
DEL
Pair from a sample with the DEL, and small insert size
Pair from a sample without the DEL, and large insert size
The increase in insert size produced by a short DEL makes it difficult to distinguish supporting from non-supporting pairs, considering the variable nature of the insert size within a sample
(b)

Figure 3: (a) Increase in insert size. (b) Short deletion results in no significant increase.

We start with the output of *NanomonSV* and *SAVANA-classified* as these are the tools specifically designed for somatic detection, using the tool in *union mode*. We then add the output of *CuteSV-sub* to the ensemble, using the tool in *validation mode*. This is because *CuteSV* is not a somatic caller, so deletions found only with the custom *subtraction* procedure are not as reliable as those found by other tools that use the tumor-normal pair simultaneously [24]. Similarly, we add the output of *SAVANA-filtered* to the ensemble, again in *validation mode*. The filtered method is not as stringent as the classified one, so we consider the deletions found only by *SAVANA-filtered* as less reliable. At the end of this procedure, we obtain a VCF file with the deletions found either by *NanomonSV*, *SAVANA-classified*, or both, validated with the deletions found by *CuteSV-sub* and *SAVANA-filtered*.

2.4 Validation using short reads

For samples that have been sequenced using short-read technologies, like Illumina, we can use the *.bam* alignment files to extract evidence that supports the deletions found by the long-read variant callers. Supporting evidence in the tumor sample increases the confidence that the deletion is real and somatic while supporting evidence in the normal sample decreases it.

We devised two methods, one based on *gaps and soft-clipped bases* and another on the *insert size* of paired-end reads. We applied them independently to each deletion in the VCF file resulting from the ensemble step, to obtain a new VCF with additional custom information fields, the most relevant described in Table 2. The two methods are described in the following subsections.

Table 2: Custom information fields for short-read validation

Field name	Description	Acronym
NUM_GAP_READS	No. of reads with gap	NGR
AVG_GAP_SIM	Avg. gap similarity	AGS
NUM_SC_READS_<BP>	No. of reads with S.C.	NSR<BP>
AVG_AL_SCORE_<BP>	Avg. alignment score of S.C. bases	AAS<BP>
AVG_GAP_SIM_SC_<BP>	Avg. reconstructed gap similarity	AGS<BP>
NUM_PE_PASS_READS	No. of pairs with valid insert size	NPR

<BP> is either LEFT (L) or RIGHT (R), indicating the breakpoint.
S.C. stands for soft-clipping or soft-clipped.

2.4.1 *Evidence from gaps and soft-clipped bases.*

When mapping short reads coming from a sample that contains a deletion, some of the short reads spanning the deletion will be mapped to the reference genome with a *gap* if the deletion is short enough, or at one of the two breakpoints, with a *soft-clipped* part, for longer deletions (Figure 2).

We extract (among other metrics) the number of reads with a gap in the same location as the deletion and the average similarity of such gaps with the deletion's position and length.

For the soft-clipped reads, we observe that the soft-clipped bases should match the sequence on the other side of the other breakpoint, and as can be seen in Figure 2b, remapping these bases results in an alignment resembling that with a gap. We extract the number of reads with soft-clipped bases, the average score of realigning the soft-clipped bases (obtained with a standard pairwise aligner), and the average similarity of the resulting gaps with the deletion's position and length.

2.4.2 *Evidence from insert size variation.*

This method leverages the expected insert size and is applicable for cases where the short reads are paired-end. When aligning short reads from a sample that contains a deletion, some pairs will map around the deletion, resulting in a larger-than-usual insert size. The increase in insert size should correspond to the length of the deletion (Figure 3a).

As the method relies on the expected insert size, which is not the same for all pairs, we take into consideration its distribution within the sample. We use the `samtools stat` command to obtain the insert size average and standard deviation values. Only deletions that are longer than the average insert size plus three times the standard deviation are considered for this method, as shorter deletions would not result in a significant increase in insert size to be distinguishable (Figure 3b). For each deletion passing the length criteria, we get the pairs that map around it and adjust their insert size by subtracting the length of the deletion. We consider supporting pairs those whose adjusted insert size falls within 2 standard deviations of the average expected value.

2.5 Ranking of deletions

We use the VCF file augmented with the information from the ensemble and short-read validation steps to calculate a set of score values for each deletion as indicated in the following list, and obtain a final score by tallying the evidence for and against the deletion (plus and minus signs respectively):

+ The support value from each of the SV callers that found the deletion. For *NanomonSV* and *SAVANA* we calculate it by subtracting the number of long reads supporting the deletion in the tumor sample from those supporting it in the normal sample. For *CuteSV-sub*, we use the support in the tumor sample.
+ Overlapping index calculated during the ensemble step, which is a measure of how much two variant callers agree on the length and location of each deletion.
+ Unified gap and soft-clipping evidence from the short-read validation step in the tumor sample. We calculate a single gap score for each deletion and two soft-clipping scores, one for each breakpoint, that are added together, as described in Equation (1).
- Unified gap and soft-clipping evidence in the normal sample (calculated as described for the tumor sample).
+ Insert-size evidence from the short-read validation step in the tumor sample. We use the number of read pairs that have passed the insert-size validation check described in the previous section.
- Insert-size evidence in the normal sample (calculated as described for the tumor sample).

$$\begin{aligned} unified\_score &= gap\_score + sc\_score\_l + sc\_score\_r \\ gap\_score &= NGR \times AGS \\ sc\_score\_l &= NSRL \times AASL \times AGSL \\ sc\_score\_r &= \underbrace{NSRR \times AASR \times AGSR}_{\text{Refer to Table 2 for the meaning of these acronyms}} \end{aligned} \tag{1}$$

Finally, we normalize each of the calculated score values, removing outliers and scaling them to a range between 0 and 1. We then add these normalized values to obtain the final score of each deletion. The deletions are then sorted, with the highest-ranking deletions being the most likely somatic events.

3 Results

To evaluate the efficacy of our pipeline we used the *Espejo Valle-Inclan* benchmark and compared the results with the truth set provided by the authors, consisting of a curated set of somatic SVs obtained from a paired melanoma and normal lymphoblastoid COLO829 cell line, using four different sequencing technologies, four tools for SV calling and three experimental validation methods [8].

Besides the truth set, the benchmark provides .bam alignment files for the Nanopore and Illumina samples obtained with *NGMLR* and *BWA-MEM*, respectively. However the *NanomonSV* and *SAVANA* callers prefer the output of *minimap2*, so we have rerun the alignment step for the ONT reads. The Illumina alignments were used as is. During the analysis of results, we have also used intermediate *raw* VCF files provided by the benchmark, with the deletions detected by individual variant callers. We configured the pipeline to run with a minimum SV length of 30 BP, and to use GRCh37 as reference genome, to match the choices made in the *Espejo Valle-Inclan* benchmark.

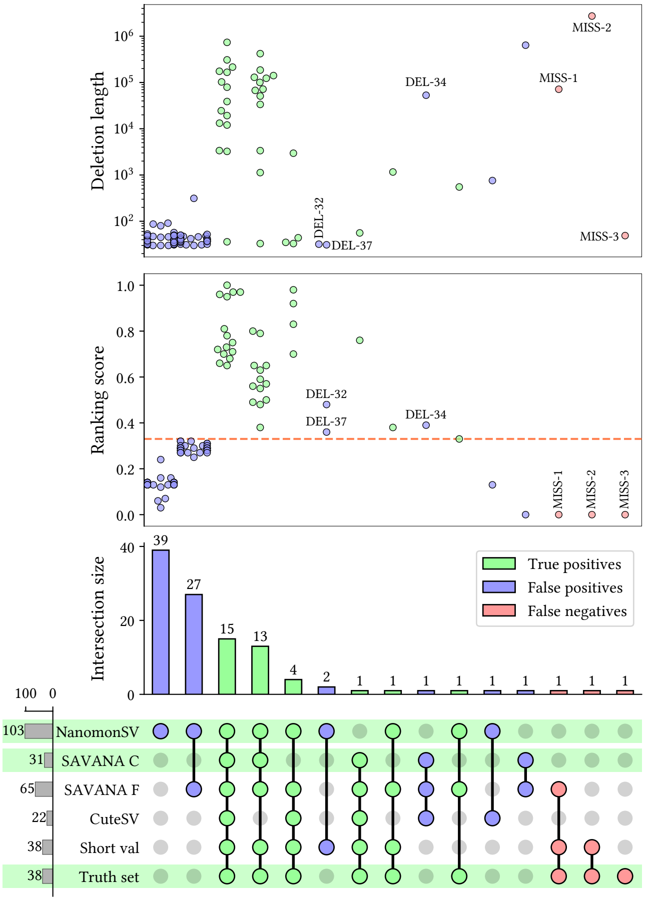

Figure 4: Results for the *Espejo Valle-Inclan* benchmark. From bottom to top: Upset plot with intersections between the tools and the truth set. Swarm plot with the ranking score of each deletion, by intersection subset. Swarm plot with the length of each deletion, by intersection subset. The three plots are colored to represent: true positives (green), false positives (blue), and false negatives (red). False positives with a score higher than the lowest ranking true positive are labeled, as well as the false negatives.

3.1 Obtained deletions

Figure 4 summarizes the results of our pipeline for the *Espejo Valle-Inclan* benchmark. The Upset plot shows the intersections between the tools and the truth set. We highlight *NanomonSV* and *SAVANA* as these are the tools that are used in *union mode*. The two swarm plots show the length and ranking score for each of the deletions in the data, organized by intersection subset. We use green to represent true positives (deletions in the truth set found by the pipeline), blue for false positives (deletions not in the truth set found by the pipeline) and red for false negatives (deletions in the truth set not found by the pipeline).

True positives. Out of the 38 deletions in the truth set, the pipeline found 35. 15 of them are common to all methods, 13 to all except *CuteSV-sub*, 4 to all except *SAVANA*, 1 to all except *NanomonSV*, and 2 were missing in two or more of the tools. The ranking score of these deletions falls in the range of 0.33 to 1, and their length ranges from 30 BP (minimum set in the configuration) to 73 KBP.

False positives. There is a total of 71 deletions found by the pipeline that are not in the truth set. However, 39 of these deletions are found only by *NanomonSV*, 27 by *NanomonSV* and *SAVANA-filtered*, 1 by *NanomonSV* and *CuteSV*, and 1 by *SAVANA*, leading to a ranking score lower than the threshold set by the lowest ranking true positive, of 0.33. Interestingly, the deletions in these four subsets are short, most of them being under 100 BP.

We focus our attention on the three remaining false positives, that have a score higher than 0.33. *DEL-32* and *DEL-37* are 32 and 31 BP long, respectively, and are found by *NanomonSV* with very good support. Additionally, they are supported by the *gap/soft-clipping* method. A manual visual inspection using IGV shows that there is support for the deletions in both long and short reads. Finally, from verification of raw VCF files from the benchmark, we have found that the deletions are present, with a length of 29, which points to the deletions being filtered out by some of the tools and from the final truth set due to their length.

DEL-34 on the other hand is 53,090 BP long and was found by *SAVANA* and *CuteSV-sub* with medium read support. Visual inspection with IGV shows that in the tumor sample, there are no reads mapped in the region, for both short and long reads, and in the long reads there is evidence of soft-clipping around the breakpoints. In the normal sample, there is a decrease in coverage in the region and there is evidence of soft-clipping around the breakpoints. By examining the raw VCF files from the benchmark, we have found that a nearby deletion is detected on both the tumor and normal samples and therefore it is probably filtered out as a germline event.

False negatives. Three of the deletions in the truth set were not found by our pipeline, as they were not reported by either *NanomonSV* or *SAVANA-classified*. We have run the short validation step for these deletions using their description obtained from the benchmark and these results are also reported in Figure 4.

MISS-1 (*truthset_13* in the benchmark) is 71,459 BP long. The benchmark reports that it was found only in the Illumina reads and validated by PCR and Capture probes. From our results, it is detected only by *SAVANA-filtered* and the *gap/soft-clipping* method, with low support. Visual inspection does not reveal much as there is no obvious evidence of a deletion in the region.

MISS-2 (*truthset_41*) is 2,732,608 BP long and in the benchmark it was found and validated by the same methods as *MISS-1*. From our results, only the *gap/soft-clipping* method detects it, with very low support. *SAVANA* reports a BND event at the same position (`C]10:33386465]`), that has a BND mate at another chromosome (`C]1:87337010]`), indicating a more complex rearrangement.

MISS-3 (*truthset_62*) is 49 BP long. The benchmark reports that it was found in the ONT and PacBio reads and validated only by PCR. From our results, it is not detected by any of the tools, but *CuteSV-sub* reports a deletion of 36 BP in the proximity of *MISS-3* (70 BP apart) with very good support. From visual inspection with IGV, there is very good evidence in the tumor long reads of the deletion found by *CuteSV-sub* (reads with gapped alignment). In the normal sample, there is evidence of a smaller deletion in the region, which may have caused the other tools to not call the larger deletion as somatic. Indeed, in the benchmark the deletion is annotated as *clear depth change, nested with a germline deletion.*

3.2 Performance of individual tools

Figure 5 shows the ranking score against the support given by each tool, colorized to distinguish between true and false positives, and sized to represent the length of the deletion.

NanomonSV is the tool that produces more false positives (69 in total), but it is also the tool that supports more true positives, missing only one. Most of the false positives are small deletions, with less than 10 supporting reads. The true positives vary in length, most of them having more than 10 supporting reads. *DEL-13* (*truthset_50* in the benchmark), the true positive that *NanomonSV* misses is 56 BP long and has good support from the other tools (over 40 reads). From visual inspection with IGV, there is very clear evidence of the deletion being somatic and present in both long and short reads. Interestingly, this deletion together with *truthset_37* (which *SAVANA-classified* misses) are the only ones reported as *NOT VALIDATED* in the truth set. These two deletions were not validated by any targeted assay but were supported by multiple technologies and manually verified from raw sequencing data [8].

To visualize the behavior of the two *SAVANA* outputs, the plot utilizes two different markers and color shades. *SAVANA-classified* is more stringent in its output, and although it only reports two false positives, it misses six true positives. Adding the output from *SAVANA-filtered* to the ensemble step (in *validation mode*) recovers the six true positives, but also adds 27 false positives that intersect with those by *NanomonSV* to form a cluster of small deletions with low support and ranking score close to the threshold. Regarding the six deletions that *SAVANA-classified* misses (*DEL-<02, 07, 08, 19, 35, 38>*, *truthset_<17, 37, 68, 47, 32, 42>* in the benchmark), all of them are filtered by the classifier as *likely noise*, including the first four in the list which have good support (32, 27, 35, 15 reads respectively). Visual inspection with IGV reveals that at least for these four deletions, there is very good evidence of them being somatic and present in both long and short reads.

CuteSV-sub produces only two false positives (when used in *validation mode*), but it is the tool that misses more true positives found by all other tools. This suggests that the *subtraction* process is too stringent, but in any case useful to increase the confidence in the deletions that it does report.

For the short validation step, we have used two markers to differentiate between the two methods used. The *insert size* approach reports no false positives, partially because the method is not applied for small deletions, and most of the false positives (that originate from *NanomonSV*) are small events. It reports support for all true positives that are large enough for the method to be applicable, with very good support, thanks to the high sequencing depth of the short reads in the tumor sample (mean depth of 97). The *gap/soft-clipping* method reports only two false positives with considerable support, and as discussed before these two events are likely somatic deletions filtered due to their size (*DEL-32, DEL-37*). It misses four true positives, but for three of these, the *insert-size* method reports support. This results in an overall short-validation process that leverages the strengths of both methods and provides a very low false positive rate with very good support for true positives.

In table 3 we report a summary of the described results for the somatic SV callers, *NanomonSV* and *SAVANA-classified*, and the ensemble pipeline. Besides the threshold-based approach, that

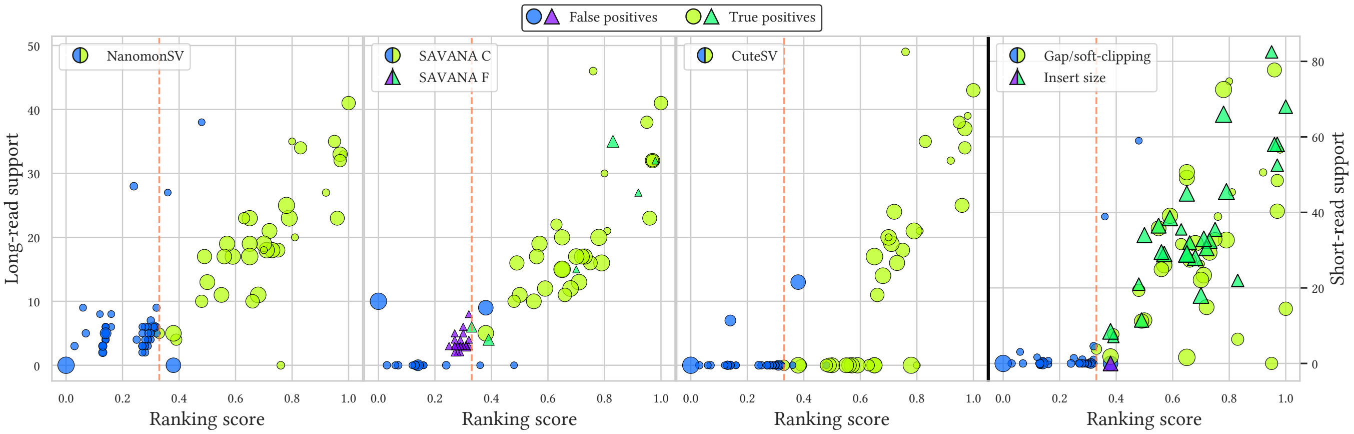

Figure 5: Individual tools results for the *Espejo Valle-Inclan* benchmark.

Ranking score vs supporting reads, for each tool. In green tones, the true positives; in blue tones, the false positives. A dashed line indicates the threshold of 0.33. The size of the markers is proportional to the log of the deletion length. For *SAVANA*, two marker styles are used to differentiate between deletions found by *SAVANA-classified* and *SAVANA-filtered*. Similarly for the short-validation plot, to differentiate between the *gap/soft-clipping* and *Insert size* methods.

has the drawback of working only for evaluation w.r.t. a truth set, we also present results for a more general rule-based approach that considers a deletion as somatic only if it is supported by at least three of the tools, and at least one of them is *NanomonSV* or *SAVANA-classified*. This rule-based approach results in an increased precision at the cost of a decreased recall when compared to the threshold-based method.

Table 3: Performance metrics for the somatic SV tools and the ensemble pipeline

Method	Total	TP	FP	FN	Precis.	Recall
NanomonSV	103	34	69	4	0.33	0.89
SAVANA-classified	31	29	2	9	0.93	0.76
Ensemble (thold)	38	35	3	3	0.92	0.92
Ensemble (rule)	34	34	0	4	1.00	0.89

4 Discussion and Future Works

We have presented our ensemble approach to detect and rank somatic deletions from long reads, supported by evidence in short reads. We have shown that our method can detect and rank somatic deletions with high precision and recall, in agreement with the *Espejo Valle-Inclan* benchmark results, even if using only two technologies, ONT and Illumina. We have also shown that the ensemble approach can compensate for the weaknesses of the individual tools and that the ranking order can be used to prioritize the events and distinguish between true and false positives.

Besides the functional pipeline, our work has also discussed the strengths and weaknesses of the individual tools. *NanomonSV* is the tool that reported most of the true positives, but it also has a high number of false positives, that could nonetheless be real somatic events that have been filtered by other tools and the truth set due to their small size. *SAVANA* has demonstrated some weakness with its classifier mode, which misses real events; this issue can be mitigated with the careful usage of its legacy mode, filtering the raw breakpoints output. *CuteSV*, even though is not a somatic variant caller, can be used with the *subtraction* approach as a reinforcement method for the outputs of the somatic variant callers. We have also shown that our *short-validation* methods, *gap/soft-clipping* and *insert-size*, compensate each other to support deletions of different sizes, resulting in a reliable approach to validate the events detected by the long-read variant callers.

Even though our pipeline is functional and our approach has shown promising results, there are still some improvements that we can implement. We should perform more testing with different benchmarks, to improve the characterization and tuning of the individual variant callers and the ensemble, validation and ranking methods. This however is not trivial, as benchmarks for somatic structural variants with tumor-normal pairs and availability of long and short reads are scarce.

Separating true from false positives is a crucial step in the analysis of SVs. In our approach, we have used the rank to prioritize the events, and we have set a threshold value based on the truth set. This approach needs to be improved to support the analysis of real data, where the truth set is not available. A possible strategy lies in the identification of the intersection of tools that provides a high confidence in precision and recall, taking into account the performance of each tool. For now we have briefly introduced a simple rule-based approach, that needs to be refined with the results obtained from more extensive benchmarking.

Currently, our approach is focused on deletions, but to transform it into a complete tool for somatic SVs ensemble and ranking, we need to extend it to support other types of SVs, such as insertions. Unfortunately, insertions are harder to characterize than deletions as there is less agreement between SVs callers, and they are harder to validate using short-read approaches [18]. Additionally, there is less benchmarking information for insertions; for instance, the *Espejo Valle-Inclan* truth set only contains three of them.

We have shown that our ensemble approach can produce a high-quality ranked list of somatic deletions, thanks to the exploitation of long and short-read sequencing technologies and the use of multiple variant callers. The ranking is useful for downstream analysis and the experimental validation of events, allowing the user to focus on the most promising candidates. The comparison of our ranked list with the output of other tools should help in determining the strengths and weaknesses of each of them, giving insights into how to proceed with further refinements of the calling algorithms. Ideally, the same procedure could be carried out for PacBio sequencing data, to determine if also different technologies matter in the quality of the resulting calls and pointing at the calling algorithms best suited for each one.

References

[1] Sergey Aganezov, Sara Goodwin, Rachel M Sherman, Fritz J Sedlazeck, Gayatri Arun, Sonam Bhatia, Isac Lee, Melanie Kirsche, Robert Wappel, Melissa Kramer, Karen Kostroff, David L Spector, Winston Timp, W Richard McCombie, and Michael C Schatz. 2020. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Research* 30, 9 (4 sep 2020), 1258–1273. https://doi.org/10.1101/gr.260497.119

[2] Mian Umair Ahsan, Qian Liu, Jonathan Elliot Perdomo, Li Fang, and Kai Wang. 2023. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nature Methods* 20, 8 (June 2023), 1143–1158. https://doi.org/10.1038/s41592-023-01932-w

[3] Yu Chen, Amy Y. Wang, Courtney A. Barkley, Yixin Zhang, Xinyang Zhao, Min Gao, Mick D. Edmonds, and Zechen Chong. 2023. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nature Communications* 14, 1 (Jan. 2023). https://doi.org/10.1038/s41467-023-35996-1

[4] Marco Raffaele Cosenza, Bernardo Rodriguez-Martin, and Jan O Korbel. 2022. Structural variation in cancer: role, prevalence, and mechanisms. *Annual Review of Genomics and Human Genetics* 23 (31 aug 2022), 123–152. https://doi.org/10.1146/annurev-genom-120121-101149

[5] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* 35, 4 (April 2017), 316–319. https://doi.org/10.1038/nbt.3820

[6] Nicolas Dierckxsens, Tong Li, Joris R. Vermeesch, and Zhi Xie. 2021. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biology* 22, 1 (Dec. 2021). https://doi.org/10.1186/s13059-021-02551-4

[7] Hillary Elrick, Jose Espejo Valle-Inclan, Katherine E. Trevers, Francesc Muyas, Rita Cascão, Angela Afonso, Cláudia C. Faria, Adrienne M. Flanagan, and Isidro Cortés-Ciriano. 2023. Abstract LB080: SAVANA: a computational method to characterize structural variation in human cancer genomes using nanopore sequencing. *Cancer Research* 83, 8_Supplement (April 2023), LB080–LB080. https://doi.org/10.1158/1538-7445.am2023-lb080

[8] Jose Espejo Valle-Inclan, Nicolle J.M. Besselink, Ewart de Bruijn, Daniel L. Cameron, Jana Ebler, Joachim Kutzera, Stef van Lieshout, Tobias Marschall, Marcel Nelen, Peter Priestley, Ivo Renkens, Margaretha G.M. Roemer, Markus J. van Roosmalen, Aaron M. Wenger, Bauke Ylstra, Remond J.A. Fijneman, Wigard P. Kloosterman, and Edwin Cuppen. 2022. A multi-platform reference for somatic structural variation detection. *Cell Genomics* 2, 6 (June 2022), 100139. https://doi.org/10.1016/j.xgen.2022.100139

[9] Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* 38, 3 (Feb. 2020), 276–278. https://doi.org/10.1038/s41587-020-0439-x

[10] Li Fang, Jiang Hu, Depeng Wang, and Kai Wang. 2018. NextSV: a meta-caller for structural variants from low-coverage long-read sequencing data. *BMC Bioinformatics* 19, 1 (May 2018). https://doi.org/10.1186/s12859-018-2207-1

[11] Li Tai Fang, Pegah Tootoonchi Afshar, Aparna Chhibber, Marghoob Mohiyuddin, Yu Fan, John C Mu, Greg Gibeling, Sharon Barr, Narges Bani Asadi, Mark B Gerstein, Daniel C Koboldt, Wenyi Wang, Wing H Wong, and Hugo Y K Lam. 2015. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biology* 16, 1 (17 sep 2015), 197. https://doi.org/10.1186/s13059-015-0758-2

[12] Akihiro Fujimoto, Jing Hao Wong, Yukiko Yoshii, Shintaro Akiyama, Azusa Tanaka, Hitomi Yagi, Daichi Shigemizu, Hidewaki Nakagawa, Masashi Mizokami, and Mihoko Shimada. 2021. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Medicine* 13, 1 (April 2021). https://doi.org/10.1186/s13073-021-00883-1

[13] Amit Goyal, Hyuk Jung Kwon, Kichan Lee, Reena Garg, Seon Young Yun, Yoon Hee Kim, Sunghoon Lee, and Min Seob Lee. 2017. Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGENTM Bio-IT Processor for Precision Medicine. *Open journal of genetics* 07, 01 (2017), 9–19. https://doi.org/10.4236/ojgen.2017.71002

[14] Douglas Hanahan. 2022. Hallmarks of cancer: New Dimensions. *Cancer discovery* 12, 1 (jan 2022), 31–46. https://doi.org/10.1158/2159-8290.{CD}-21-1059

[15] Tao Jiang, Yongzhuang Liu, Yue Jiang, Junyi Li, Yan Gao, Zhe Cui, Yadong Liu, Bo Liu, and Yadong Wang. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* 21, 1 (Aug. 2020). https://doi.org/10.1186/s13059-020-02107-y

[16] Heng Li. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 37, 23 (Oct. 2021), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

[17] Yilong Li, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O. Korbel, James E. Haber, Marcin Imielinski, Joachim Weischenfeldt, Rameen Beroukhim, and Peter J. Campbell. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 7793 (feb 2020), 112–121. https://doi.org/10.1038/s41586-019-1913-9

[18] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. 2019. Structural variant calling: the long and the short of it. *Genome Biology* 20, 1 (20 nov 2019), 246. https://doi.org/10.1186/s13059-019-1828-7

[19] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 9 (sep 2010), 1297–1303. https://doi.org/10.1101/gr.107524.110

[20] Brent S Pedersen and Aaron R Quinlan. 2017. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* 33, 12 (Feb. 2017), 1867–1869. https://doi.org/10.1093/bioinformatics/btx057

[21] Tobias Rausch, Rene Snajder, Adrien Leger, Milena Simovic, Mădălina Giurgiu, Laura Villacorta, Anton G Henssen, Stefan Fröhling, Oliver Stegle, Ewan Birney, Marc Jan Bonder, Aurelie Ernst, and Jan O Korbel. 2023. Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. *Cell Genomics* 3, 4 (12 apr 2023), 100281. https://doi.org/10.1016/j.xgen.2023.100281

[22] Anthony Rhoads and Kin Fai Au. 2015. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute* 13, 5 (oct 2015), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

[23] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 15, 6 (jun 2018), 461–468. https://doi.org/10.1038/s41592-018-0001-7

[24] Yuichi Shiraishi, Junji Koya, Kenichi Chiba, Ai Okada, Yasuhito Arai, Yuki Saito, Tatsuhiro Shibata, and Keisuke Kataoka. 2023. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Research* 51, 14 (June 2023), e74–e74. https://doi.org/10.1093/nar/gkad526

[25] Moritz Smolka, Luis F. Paulin, Christopher M. Grochowski, Dominic W. Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, Mira Gandhi, Karl Hong, Davut Pehlivan, Sonja W. Scholz, Claudia M.B. Carvalho, Christos Proukakis, and Fritz J Sedlazeck. 2022. Comprehensive Structural Variant Detection: From Mosaic to Population-Level. (April 2022). https://doi.org/10.1101/2022.04.04.487055

[26] Jose Espejo Valle-Inclan, Christina Stangl, Anouk C. de Jong, Lisanne F. van Dessel, Markus J. van Roosmalen, Jean C. A. Helmijr, Ivo Renkens, Roel Janssen, Sam de Blank, Chris J. de Witte, John W. M. Martens, Maurice P. H. M. Jansen, Martijn P. Lolkema, and Wigard P. Kloosterman. 2021. Optimizing Nanopore sequencing-based detection of structural variants enables individualized circulating tumor DNA-based disease monitoring in cancer patients. *Genome Medicine* 13, 1 (May 2021). https://doi.org/10.1186/s13073-021-00899-7

[27] Md. Vasimuddin, Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

[28] Mingyi Wang, Wen Luo, Kristine Jones, Xiaopeng Bian, Russell Williams, Herbert Higson, Dongjing Wu, Belynda Hicks, Meredith Yeager, and Bin Zhu. 2020. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Scientific Reports* 10, 1 (30 jul 2020), 12898. https://doi.org/10.1038/s41598-020-69772-8

[29] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* 39, 11 (8 nov 2021), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x