Algorithms for structural variant detection

Lin, J.

Chapter 5

Assessing reproducibility of long-read structural variant detection algorithms

Abstract Recent advances in long-read sequencing and haplotype-aware assemble have enabled phased structural variants (SV) detection and improved SV detection at complex genomic regions. The assembly-based approach for tumor SV detection is further complicated due to heterogeneous cell populations and polyploid tumor genomes. Though a number of alignment-based methods that are more robust to complex tumor genomes have been developed, they lacked systematic evaluation of reproducibility, especially at complex genomic regions, which is critical for promoting long-read application in clinical practices. In this study, we benchmark six alignment-based methods on four real datasets produced by PacBio and Oxford Nanopore sequencers for recall, precision, SV breakpoints and type consistency as well as capability of detecting SVs at repetitive regions. Our results first highlight the important role of aligners in determining SV breakpoint concordance of detection algorithms. Secondly, our analysis based on phased assembly reveals that tandem repeat regions are hotspots for discordant calls of each algorithm detected from different aligners and platforms combinations. In addition, the analysis of tumor-normal paired samples suggest that the number of different SV types varies from tumor unique calls identified from each caller, and integration of tumor unique calls from each caller would substantially improve somatic SV detection. As the importance of SVs are increasingly recognized in disease genomes, our analysis provides important guidelines for selecting dataset, aligner and algorithms for efficient SV detection, and reveals valuable hints for future algorithm development, thereby shedding light on cutting-edge genomic studies and clinical applications.

5.1 Introduction

Structural variants (SVs) comprise different subclasses that consist of unbalanced copy number variants, including deletion, duplication and insertion, as well as balanced rearrangements, such as inversion and translocation [8]. SVs could also have complex internal structures, consisting of multiple combinations of the above-mentioned simple forms of SVs, and this complex form of SV is referred to as complex SV (CSV) [11, 12, 57]. In the past decade, researchers have made great progress in discovering and genotyping SVs in diverse populations and generated phased reference panels of SVs with short-read data. Moreover, researchers found that SVs are enriched for expression quantitative trait loci (eQTLs) up to 50-fold compared with single nucleotide variations, indicating the important role of SVs in regulating gene expression. Remarkably, the widespread application of single-molecule sequencing (SMS) technologies, including Pacific Bioscience (PacBio) and Oxford Nanopore Technology (ONT), greatly improves the sensitivity and precision of detecting SVs comparing with short-read [9, 41]. A study revealed that PacBio long-reads were approximately three times more sensitive than a short-read ensemble achieved, and a large set of SVs, ranging from 50 to 2000bp were unresolvable without long reads [8]. Recently, the haplotype-aware phased assembly facilitated the direct detection of phased SVs [9, 10], enabling systematic analysis of functional impact of SVs as well as SV candidates for adaptive selection within the human population.

Moreover, long-read sequencing also facilitates the analysis and manual curation of CSVs that are usually inaccessible via short-read data. For instance, in 2015, the 1000 Genomes Project (1KGP) published the first previously unexplored CSV classes by integrating both short- and long-read sequencing. Additionally, long-read sequencing revealed SVs in genetic diseases [93, 94, 95] and cancers [45, 90, 96, 97, 98, 99, 100] that are usually undetectable via short-read data. For instance, the ONT data reveals 10,000bp Alzheimer's disease associated ABCA7 Variable Number Tandem Repeats (VNTR) expansion [101] and the PacBio long-read data reveals 10 times more SVs than that of short-read in breast cancer. Additionally, the somatic SVs in tumor are a valuable genetic source to understand tumorigenesis, such as a study showed that long reads could detect two times more somatic SVs than previous short-read study [82].

Detecting SVs from SMS data usually consists of two steps. Firstly, the variant signatures are identified and gathered from two types of aberrant alignments: intra-read and inter-read. Intra-read alignments are derived from reads spanning the entire SV locus, resulting in deletion and insertion

signatures. Inter-read alignments are usually obtained from the supplementary alignments and SV signatures that could be identified from inconsistencies in orientation, location and size during mapping, analogous to read-pair signatures, from which translocation as well as large deletion, duplication and inversion signatures are identified. Secondly, callers typically cluster and merge similar signatures from multiple aberrant alignments, and delineate proximal signatures that support putative SV. Nearly all alignment-based algorithms developed in the past five years, such as Sniffles [18], pbsv, CuteSV [102], SVIM [103], NanoVar [104], NanoSV [105] and Picky [96], detect SVs through combinations of signatures obtained from inter-read and intra-read alignments but differ in their signature clustering heuristics. For example, Sniffles evaluates the signature similarities by examining the signature position and size, and additionally clusters SV supported by the same set of alignments to detect nested SVs. Some methods, such as Phased Assembly Variant (PAV) and SVIM-ASM [103] use the alignment of whole genome assembled contigs as input, referred to as assembly-based approaches, from which aberrant inter-contig and intra-contig alignments are used for SV detection.

Moreover, somatic SVs are driver events for tumorigenesis and they are usually detected by identifying SVs present in tumor but absent from its matched normal sample. For instance, CAMPHOR [82], a computational pipeline, detects somatic SVs by removing SVs present in a 'normal panel'. A similar process can also be completed by SURVIVOR, which identifies putatively somatic SVs that are only present in tumor [90]. However, affected by repetitive sequences and human reference genome defects [87], intensive breakpoint filtering and an external normal reference SV set are required to obtain high-quality somatic SVs [106, 107].

Previous studies have estimated that at least 30% of cancers have a known pathogenic SVs used in diagnosis or treatment [108], and germline variants in cancer predisposition genes underline 5–10% of all cancers [109, 110, 111]. However, the prevalence of SVs in cancer is likely underestimated due to low sensitivity and specificity for short-read based SV discovery at regions of repetitive elements, low sequence complexity and strong GC bias. Recently, long-read assembly approach significantly increased the sensitivity of detecting SVs at complex genomic regions compared to that of short-read data [9, 10], but precise detection of germline SVs and distinguishing tumor unique SVs from germline is further complicated due to tumor heterogeneity and polyploidy. Compared with assembly approaches, alignment-based detection methods are more robust to amplificated tumor genomes that originate from mixed cell populations, while inconsistencies in breakpoints and variant

types confound tumor SV detection, especially somatic SV. Therefore, it is critical to assess the detection consistency of alignment-based algorithms, especially at complex genomic regions, thereby enabling accurate and comprehensive germline and somatic SV detection. In this study, using multiple datasets of two platforms (i.e., HiFi and ONT) mapped by two aligners (i.e., minimap2 and ngmlr), we evaluated the recall, precision, variant breakpoints and type consistency of five alignment-based SV detection algorithms and assess the alignment-based algorithms for tumor SV detection.

In Section 5.2, materials and related methods are described in details. Moreover, results are discussed in Section 5.3 and conclusions are drawn in Section 5.4.

5.2 Materials and methods

In this section, we introduce the datasets and methods used in the evaluation.

5.2.1 Read mapping and SV detection

In this chapter, HiFi and ONT data are obtained for HG002, NA19240, HG00733 and HG00514, while ONT data was used for tumor-normal paired sample COLO829. Then, minimap2 [17] (v2.20) and ngmlr [18] (v0.2.7) were used to map the long-read data of HG002 and COLO829 to hg19 due to the reference version of the benchmark set. The long-read data of NA19240, HG00733 and HG00514 were mapped to reference version GRCh38. For minimap2, parameters '`-a -H -k 19 -O 5,56 -E 4,1 -A 2 -B 5 -z 400,50 -r 2000 -g 5000`' were applied to align HiFi reads, while '`-a -z 600,200 -x map-ont`' were used for ONT reads. For ngmlr, parameters '`-x pacbio`' and '`-x ont`' were used to align HiFi and ONT reads, respectively. For the detection algorithms, SVision (v1.3.6), CuteSV (v1.0.10), pbsv (v2.2.2), SVIM (v1.4.0), Sniffles (v1.0.12) and NanoVar (v1.4.1) were applied to the minimap2 and ngmlr aligned data, respectively. We used default settings for all callers, while at least five supporting reads were required for SV detection in NA19240, HG00733, HG00514 as well as normal-tumor paired COLO829 samples.

5.2.2 Evaluating recall and precision of each algorithm

We first used the evaluation method Truvari (https://github.com/spiralgenetics/truvari) developed by Genome-In-A-Bottle (GIAB) to examine the performance of each algorithm on HG002. The specific steps of SV

calling and processing for SVIM, Sniffles, CuteSV and pbsv were given by CuteSV (https://github.com/tjiangHIT/sv-benchmark). Furthermore, for SVision, SV with 'Covered' filter was considered as passed calls in the algorithm, and we replaced the 'Covered' with 'PASS' for the usage of option '`--passonly`' in Truvari. The raw calls of NanoVar were directly used as input for Truvari evaluation.

Moreover, the PAV call sets of NA19240, HG00733 and HG00514 were used to evaluate each algorithm. Note that the breakends, such as translocations, were first excluded from the raw detections and SVs ranging from 50bp to 100kbp were included in the analysis. BEDtools [85] (v2.30.0) was used to find the correct detections via the 50% reciprocal overlap test, while those failing the overlap test were considered as false detections. Specifically, we used command '`bedtools intersect -c -a pav.bed -b algorithm.bed -f 0.5 -r`' to count the unique number of matched ground truth calls. Given the number of ground truth calls (N), number of detections (D) and number of correct detections (D), the Recall, Precision and F-score were calculated as follows:

$$\text{Precision} = C/D$$

$$\text{Recall} = C/N$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2.3 Identification and classification of PAV calls missed by each algorithm

Using command '`bedtools intersect -c -a pav.bed -b algorithm.bed -f 0.5 -r`', the missed PAV calls of each algorithm were labeled as zero matches in the last column of the output. Then, the simple repeats and Repeat Masker files obtained from UCSC Genome Browser were used to label the repeat element and calculate the percentage of repeat overlap. For simple repeats, the VNTR was assigned if the repeat unit length was longer than 7bp, otherwise, it was considered as STR. In this study, we only used repeat element LINE, SINE, LTR, VNTR and STR, while other repeat elements were classified as Others.

Additionally, we developed a pipeline to classify missed PAV calls according to the read mapping signatures. Firstly, the missed PAV calls were classified to three types of regions according to the average read mapping quality (avg_{mapq}), including i) no read mapping region (No_reads), ii) low

mapping quality regions (Low_mapq, $avg_{mapq} < 20$) and high confident mapping regions (High_mapq, $avg_{mapq} \geq 20$). The average mapping quality threshold was set according to the default minimum read quality used for SV detection algorithms. Secondly, we extracted the potential SV signature reads that span the PAV calls in the high confident mapping quality regions. In general, the 'I' and 'D' tags in the CIGAR string, and the primary reads and their supplementary alignments were collected and used to identify deletion (DEL), insertion (INS), inversion (INV) and duplication (DUP) signatures. The total number of SV signature reads spanning PAV calls was referred to as signature count. Afterwards, we applied the same implementation as Truvari to match PAV calls and detected SV signature reads. Specifically, for a given SV signature read with start and end position, we calculated the minimum distance between this signature and PAV call as well as their size similarity. If the minimum distance and the size similarity of a signature read was smaller than 500bp and larger than 0.5, respectively, it was considered as the nearest signature.

5.2.4 Evaluating breakpoint accuracy

To evaluate the breakpoint accuracy of each caller, the correct detection, compared with the benchmarks (i.e., PAV calls and short-read calls) was considered as the nearest one with similar size, where the distance and size similarity threshold were 500bp and 0.5, respectively. Note that for short-read benchmark calls, we used Manta with default settings to detected SVs from Illumina reads and evaluate the minimum breakpoint shift of overlapped detections as described above. We calculated the minimum breakpoint shift of the concordant detections to evaluate the breakpoint accuracy of each caller. For the breakpoint assessment of recurrent SVs, SURVIVOR [112] was used to identify the recurrent SVs among three samples for each caller with command '`SURVIVOR 500 3 0 0 0 50`', while translocations were excluded in breakpoint accuracy assessment. For other SV types, the breakpoint accuracy was evaluated by calculating the standard deviation of variant start and end position in the merged VCF file. If the standard deviation of both start and end position was smaller than 50bp, the corresponding recurrent SV was considered as accurate detection.

5.2.5 Examine call set overlaps between platforms and aligners

For each caller, the overlapped and unique calls of different platforms and aligners were identified with SURVIVOR, running command '`SURVIVOR 500 1 0 0 0 50`'. In particular, we only examined whether an SV was detected at a specific region of different aligners or platforms, while the SV type was not considered. For example, the ngmlr and minimap2 unique and overlapped calls detected by SVision on HiFi reads was obtained from the '`SUPP_VEC`' value of SURVIVOR merged output. Specifically, '`SUPP_VEC=11`' indicates overlapped calls, while '`SUPP_VEC=10`' or '`SUPP_VEC=01`' represents aligner unique detections. This comparison between aligners of identical platform was termed as fixed-platform, and the same process was applied to compare the detections between different platforms mapped with identical aligner, referring as fixed-aligner. Afterwards, the same repeat annotation procedure was applied to annotate the unique calls from fixed-platform and fixed-aligner. This process was also applied to identify tumor unique calls, which were obtained from variant of '`SUPP_VEC=10`'.

5.2.6 Data availability

Both the HiFi and ONT data for HG002 are obtained from ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son, and the benchmark [92] for HG002 used in this chapter is from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/. The HiFi data for NA19240, HG00733 and HG00514 are obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/, and the ONT data [9] for these samples are available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20181210_ONT_rebasecalled/. The Phased Assembly Variant (PAV, v1.1.2) [10] for NA19240, HG00733 and HG00514 are downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20210806_PAV_VCF/. The normal ONT data for COLO829 is obtained from Sequence Read Archive (SRA) with ERR2752451, and the tumor ONT data is downloaded with ERR2752452. The somatic SV truth set of COLO829 is obtained from https://github.com/UMCUGenetics/COLO829_somaticSV .

5.3 Results

In this section, we first assess the impact of aligners and platforms on SV detection consistency of each alignment-based detection methods. Then, we examine the recall and precision of each method affecting by aligners and platforms. Moreover, we systematically compare SVs detected by alignment-based approach and assembly approach, especially their breakpoint consistency. Finally, using tumor-normal paired sample, we assess the impact of aligners on detecting germline and somatic SVs.

5.3.1 Evaluating the impact of aligners and platforms on detection algorithms

Platform and aligner independency is one of the important features for detection algorithm in clinical usage. The detection consistency was thus assessed with three well-characterized samples (i.e., NA19240, HG00733 and HG00514) sequenced by HiFi and ONT technologies. As a result, more SVs were detected from minimap2 aligned data than that of ngmlr, and such difference was even significant for ONT data (Figure 5.1A). Though the percentage of detected deletions and insertions per genome varied across platform and aligner combinations, 20% more insertions and deletions were detected from minimap2 alignments than that of ngmlr. Notably, approximately 98% of SVIM discoveries were insertions or deletions from minimap2 aligned HiFi data, which was 15% and 38% more than pbsv and NanoVar detected, respectively (Figure 5.2).

Further analysis showed that a large number of duplications (around ≈7,000 without aligner or platform bias) detected by NanoVar was the major factor leading to a lower proportion of detected insertions and deletions (Figure 5.1C). We also noticed that the large number of duplications detected from ngmlr aligned data contributed to 20% difference of detected insertions and deletions between aligners for each caller (Figure 5.1C). Though pbsv, CuteSV, Sniffles and NanoVar could distinguish duplications from insertions, SVIM was the first algorithm that was capable of detecting tandem duplications (DUP:TANDEM) and dispersed duplications (DUP:INT), where around 10 dispersed duplications and 100 tandem duplications per genome were identified. Note that SVision and Sniffles were capable of identifying CSVs, where SVision reported ≈100 CSVs per sample and Sniffles identified three types of CSV (i.e., DEL/INV, DUP/INS and INVDUP) (Figure 5.1C).

We then examined the impacts of aligners on SV detection from different platforms, termed fixed-platform evaluation. The overlapping calls between

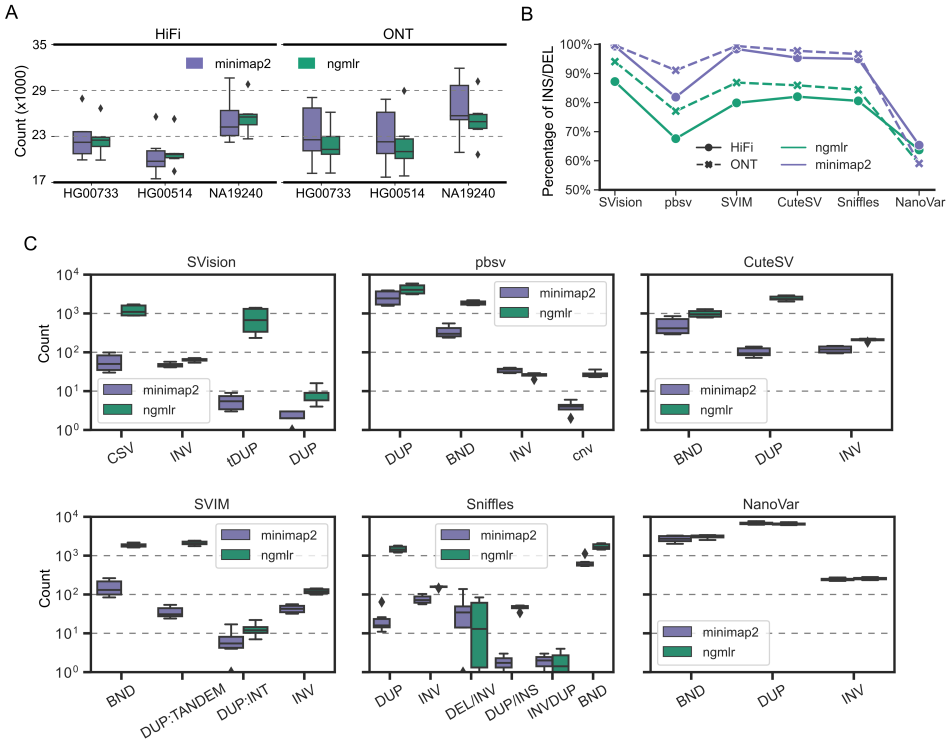

Figure 5.1: Overview of structural variants detected by six callers from three samples. (A) Number of structural variants of three samples detected from data generated by different aligners and platforms. (B) Percentage of deletions and insertions detected by each caller. (C) Number of detected structural variants of different types, excluding insertions and deletions.

two aligners were around 80% for both ONT and HiFi reads (Figure 5.2A), and breakpoint difference of most aligner concordant calls was less than 20bp (Figure 5.2B). Notably, breakpoint difference of pbsv calls was closer to 0bp on both platforms compared with other callers, indicating SV breakpoints reported by pbsv were less affected by aligners. Further analysis of aligner discordant calls revealed that all callers identified more duplications from ngmlr aligned HiFi and ONT data (Figure 5.2C), which was consistent with our previous observation on overall discoveries (Figure 5.1C), suggesting SV types reported by callers were depend on aligners. We reasoned that this limitation was largely due to the model-based SV detection approach, so that more duplications were detected from duplication like abnormal alignments observed in ngmlr aligned data.

In addition, we evaluated the platform influences, referred to as fixed-aligner evaluation, where the percentage of platform concordant calls ranged from 70% to 90% for different callers (Figure 5.2D). Though the platform concordant call took 90% of SVIM HiFi discoveries, three times more ONT unique calls were observed than HiFi unique calls (Figure 5.2D). Moreover, consistent with fixed-platform evaluation, pbsv produced concordant SV breakpoints of platform concordant calls (Figure 5.2E), suggesting pbsv was able to report consistent SV breakpoints that are less affected by aligners or platforms. Altogether, our results suggested that aligners played an important role in producing consistent SV breakpoints and types across platforms for each caller.

5.3.2 Evaluation recall and precision of detection algorithms using different benchmarks

Furthermore, it was critical to understand the sensitivity and specificity of detection algorithms for clinical applications. Therefore, we first benchmarked SVision, pbsv, CuteSV, Sniffles, NanoVar and SVIM with ground truth SVs of sample HG002. The ground truth set was an integration of multiple platforms and released by Genome-In-A-Bottle (GIAB), containing high-confident deletion and insertion calls, which had been widely used to evaluate the performance of SV detection algorithms [92]. The callers were applied to 30X HiFi data and 47X ONT data aligned with minimap2 and ngmlr, respectively. The results showed that SVision, pbsv, SVIM, CuteSV and Sniffles outperformed NanoVar across platforms and aligners. In addition, we noticed that all callers achieved the best performance on minimap2 aligned HiFi and ONT reads, and CuteSV achieved the highest F-score, followed by SVision, Sniffles and pbsv (Figure 5.3A). Though callers produced fewer

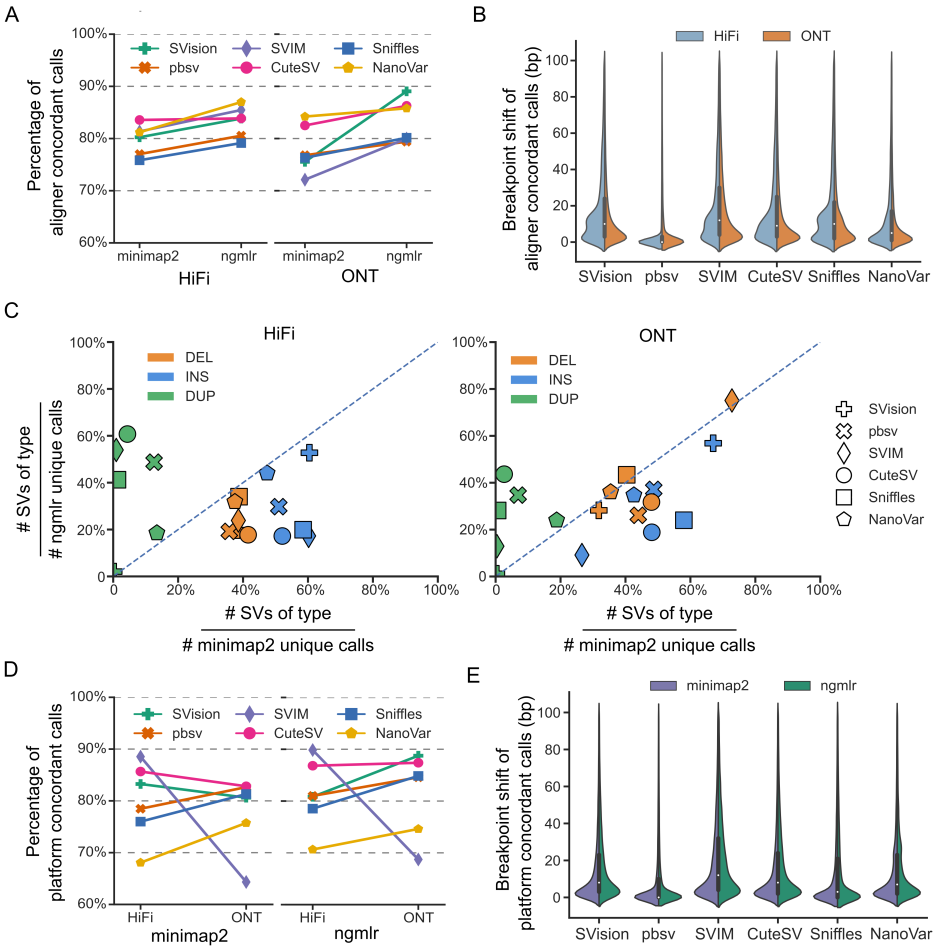

Figure 5.2: Effects of aligners and platforms on structural variants detection. (A-C) Fixed-platform evaluation of each caller. (A) Percentage of aligner concordant calls among all discoveries detected from ngmlr (vertical axis) and minimap2 (horizontal axis) alignments. (B) Breakpoint difference of aligner concordant calls. (C) Percentage of structural variant (SV) types among aligner discordant calls, i.e., minimap2 (horizontal axis) and ngmlr (vertical axis). (D-E) Fixed-aligner evaluation of each caller. (D) Percentage of platform concordant calls detected among all SVs detected from ONT (vertical axis) or HiFi (horizontal axis) reads. (E) Breakpoint difference of platform concordant calls.

correct detections on ngmlr aligned data, the precision of the six callers was comparable to minimap2 or even higher on ONT reads. For example, the precision of SVision detections on the minimap2 aligned ONT data was 80.5%, which increased to 89.9% on the ngmlr aligned ONT data (Figure 5.3A).

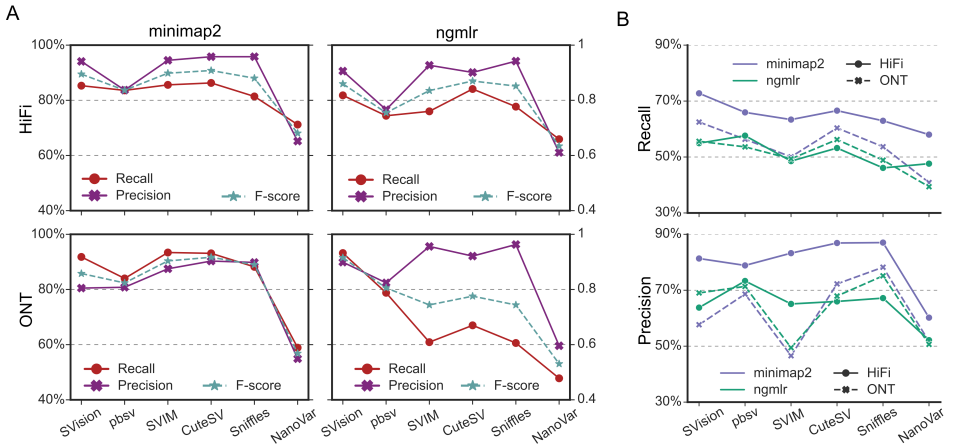

Figure 5.3: Evaluating recall and precision of six callers using different benchmarks. (A) Performance evaluated on sample HG002 HiFi and ONT data. (B) Average recall and precision evaluated on HG00514, HG00733 and NA19240.

In addition, PAV callsets of HG00514, HG00733 and NA19240 were used as ground truth to assess recall and precision of each caller. The PAV calls were detected from the highly contiguous haplotype assemblies released by HGSVC [10], which significantly improved the SV discoveries at repetitive regions compared with the HG002 truth set. Thus, the PAV callset was able to evaluate SV detection algorithms at both simple and complex genomic regions. Briefly, the SVs detected from mapped reads (i.e., HiFi and ONT aligned with minimap2 and ngmlr) of each caller were compared with the PAV calls by examining the reciprocal overlaps. Since translocation (BND) was not included in PAV calls, the BNDs from the raw calls from each caller were excluded and SVs ranging from 50bp to 100kbp were used for the performance assessment. As a result, all algorithms achieved their own best performance on minimap2 aligned HiFi reads, where SVision and pbsv ranked first on minimap2 and ngmlr aligned HiFi reads across samples, respectively (Figure 5.4B). We reasoned that this biased performance was largely due to the method of detecting PAV calls, i.e., detecting from the minimap2

aligned HiFi assemblies with extra alignment trimming. Though SV detection performance on ONT reads was not comparable with HiFi reads, the F-score of each caller based on different aligners were approximately equal, indicating less impact from aligners. Altogether, our results indicated that aligners affect more than platforms on recall and precision, where Sniffles, SVision, pbsv and CuteSV showed similar performance and consistently outperformed NanoVar across different platforms and aligners.

5.3.3 Features of PAV calls missed by detection algorithms

We then examined PAV calls missed by each caller on three samples (i.e., NA19240, HG00733 and HG00514), aiming to understand limitations of alignment-based SV detection algorithms. The missed PAV calls were considered those without matched detections via the reciprocal overlap test, and the best recall of detecting PAV calls was around 70% (Figure 5.3B). Among missed PAV calls, 70% and 28% of missed PAV calls were insertions and deletions, respectively (Figure 5.4A). Moreover, 80%, 70% and 60% of NanoVar, pbsv and CuteSV uniquely missed PAV calls were insertion, respectively, whereas more than 60% of SVIM and Sniffles missed PAV calls were deletions (Figure 5.4B). Further repeat annotation revealed that a large majority (≈70%) of missed SVs overlapped with VNTR regions, followed by STR regions (≈10%) (Figure 5.4C). These results suggested that an assembly-based approach significantly increased the sensitivity of detecting insertions and SVs in tandem repeat regions (i.e., VNTR and STR) compared with alignment-based detection. The above results were consistent with the conclusion drawn by HGSVC, where the predominant increase of PAV was among small SVs (<250bp) localized to simple repeat sequences.

Though the assembly-approach achieved remarkable results on SV detection, it was difficult to generalize for tumor genomes because of heterogeneity and aneuploidy. Therefore, we investigated whether the missed PAVs were detectable from alignment-based approaches. Firstly, we noticed that 80% of missed PAVs were located at high mapping quality regions (Figure 5.4D), providing confident alignments for SV signature reads identification. Afterwards, for missed PAV calls at high mapping quality regions, variant spanning reads were extracted and analyzed to find SV signatures. The results showed that the percentage of missed PAV calls with SV signature reads was independent of aligners for both HiFi and ONT reads, where NanoVar failed to report SVs from 88% and 77% of the genomic regions with SV signatures (Figure 5.4E).

Furthermore, we examined the nearest SV signatures, providing the direct evidence of detecting missed PAV calls. In principle, missed PAVs were not

able to be discovered from read mapping if we cannot identify the nearest SV signatures. On average, approximately 55% of missed PAVs contained nearest signatures for HiFi reads aligned with both aligners, whereas ONT reads were likely to produce more nearest signatures when aligned with minimap2 (Figure 5.4E). This indicated that half of missed PAV calls in high mapping quality regions could be recovered, while they were missed by routine SV callers due to the inaccurate breakpoints in repeat regions. Specifically, the nearest signatures could be identified from 90% of the missed PAV regions contained signatures, and the highest average PAV recall rate (≈70%) was achieved by minimap2 aligned HiFi reads, and we thus reasoned 17% more PAVs in high mapping quality regions could be detected based on signatures. Our analysis indicated that most of the PAV missed calls at simple repeat regions contain SV signature reads, and these PAV calls could be detected with proper breakpoint fine mapping.

5.3.4 Examining the effects of platforms and aligners on breakpoint accuracy

In addition, accurate breakpoints are critical to the downstream SV functional annotation such as gene annotation and known pathogenetic variant annotation, and we thus investigated the breakpoint accuracy of each caller by comparing with two independent call sets generated via orthogonal approaches, i.e., phased assembly and short-read. For phased assembly evaluation, using PAV calls, the breakpoint difference of ≈80% concordant calls were smaller than 50bp for minimap2 and ngmlr across different callers (Figure 5.5A). Moreover, consistent with the fixed-platform (Figure 5.2B) and fixed-aligner (Figure 5.2E) evaluation, pbsv achieved the most accurate breakpoints (breakpoint difference smaller than 10bp) without aligner and platform bias (Figure 5.5A). We next divided the concordant calls into two groups: i) accurate detections (breakpoint difference smaller than 50bp, Figure 5.5B) and ii) inaccurate detections (breakpoint difference larger than 50bp), and found that a significant number of inaccurate detections were located at VNTR regions (78%) (Figure 5.5C). This suggested that the breakpoints of SVs detected from read and assembly were largely different at simple repeat regions, especially in VNTR. Due to the aligner bias of PAV calls, breakpoint accuracy was further evaluated with short-read data, of which pbsv also showed the most accurate breakpoints and it was independent of aligners and platforms (Figure 5.6A).

On the contrary, the breakpoint accuracy of other callers was dependent on aligners, where we found the percentage of breakpoint shift smaller than

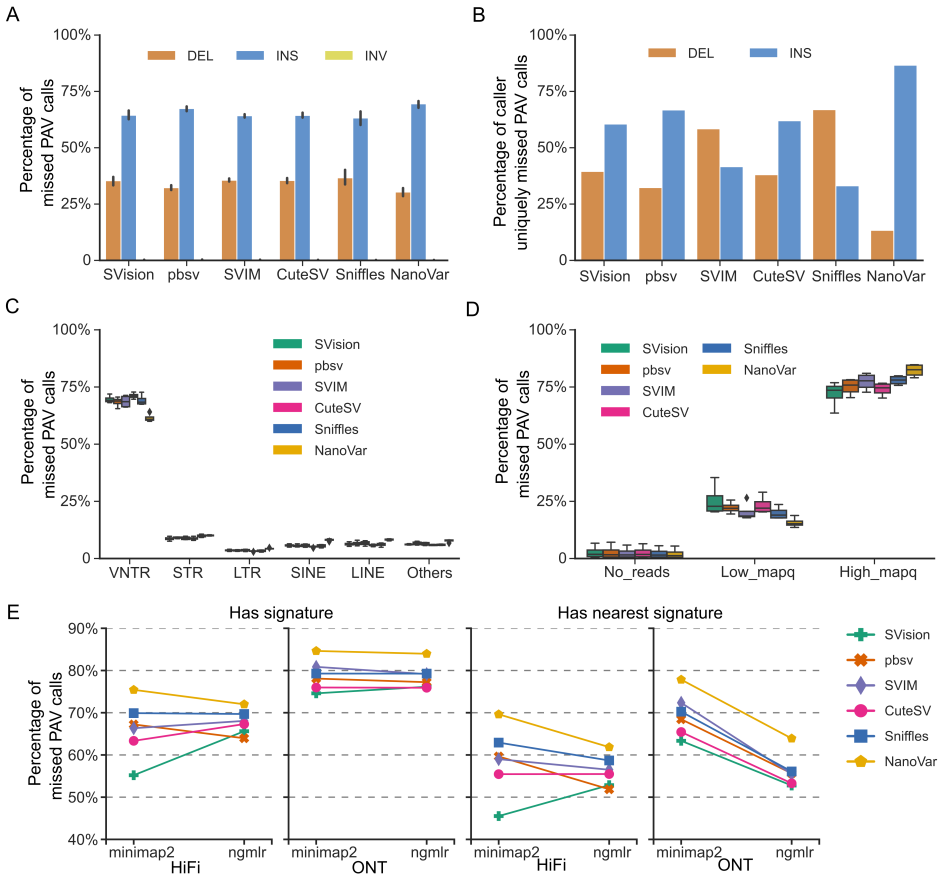

Figure 5.4: Features of missed Phased Assembly Variant by six callers. (A) Distribution of missed Phased Assembly Variants (PAVs) detected from different aligners and platforms. (B) Types of caller uniquely missed PAV calls. (C) Repeat annotation of missed PAVs. (D) Mapping quality of the missed PAV loci, including no read mapping (No_reads), low mapping quality (Low_mapq) and high mapping quality (High_mapq). (E) Missed PAV loci that had signatures and nearest signature identified from long reads aligned with minimap2 and ngmlr.

10bp increased 30% on minimap2 aligned reads (Figure 5.6A). Both PAV and short-read data revealed that HiFi data paired with minimap2 would produce the most accurate breakpoints for all callers. To avoid potential aligner bias of the benchmarks, we assessed the breakpoint accuracy of different callers by comparing the breakpoints of recurrent SVs among different samples. As a result, SVs detected by callers except SVision were likely to have consistent breakpoints on minimap2 aligned HiFi or ONT data, where Sniffles outperformed other callers among different platforms and aligners (Figure 5.6B). Our results suggested that the selection of aligner was critical to get consistent breakpoints for routine SV detection algorithms, while tandem repeat regions (i.e., VNTR) required extra breakpoint refinement if the caller was applied to repeat expansion related diseases, such as Huntington disease.

5.3.5 Effects of aligners on tumor SV detection

The above results suggested that aligners play an important role for consistent SV detection. We then evaluated the impact of aligner for tumor genome analysis, especially the performance of detecting somatic SVs from tumor unique calls. Briefly, each routine SV caller was used to detect SVs from tumor (ONT, ≈60X coverage) and normal (ONT, ≈40X coverage) data of COLO829 separately, and the filtering-based approach was applied to identify tumor unique calls, which are also called putatively somatic SVs. As a result, the total number of SVs detected by NanoVar from tumor and normal tissues was independent of aligners, whereas Sniffles, CuteSV and SVIM detected more SVs from minimap2 alignments comparing to ngmlr, thereby leading to 5% more minimap2 unique detections than that of ngmlr (Figure 5.7A). Furthermore, we investigated the impact of aligners on identifying tumor unique calls, which is one of the critical steps to obtain somatic SVs. The results showed that the percentage of tumor unique calls obtained from NanoVar and Sniffles was less affected by aligners (Figure 5.7B), and NanoVar had the largest number of tumor unique calls, i.e., 7,626 and 7,676 from minimap2 and ngmlr alignments, respectively.

On average, 50% of the tumor unique calls were inside the repetitive regions, of which the majority of them were annotated as SINE or LINE. As for the SV types of tumor unique calls, ≈4,500 putatively somatic deletions were identified from SVIM calls detected based on minimap2 alignments, which was four times more than detected insertions (≈1,000 events) (Figure 5.7C). Comparably, approximately 3,300 of the tumor unique calls identified from NanoVar was translocations, attributing to 44% of the tumor unique calls,

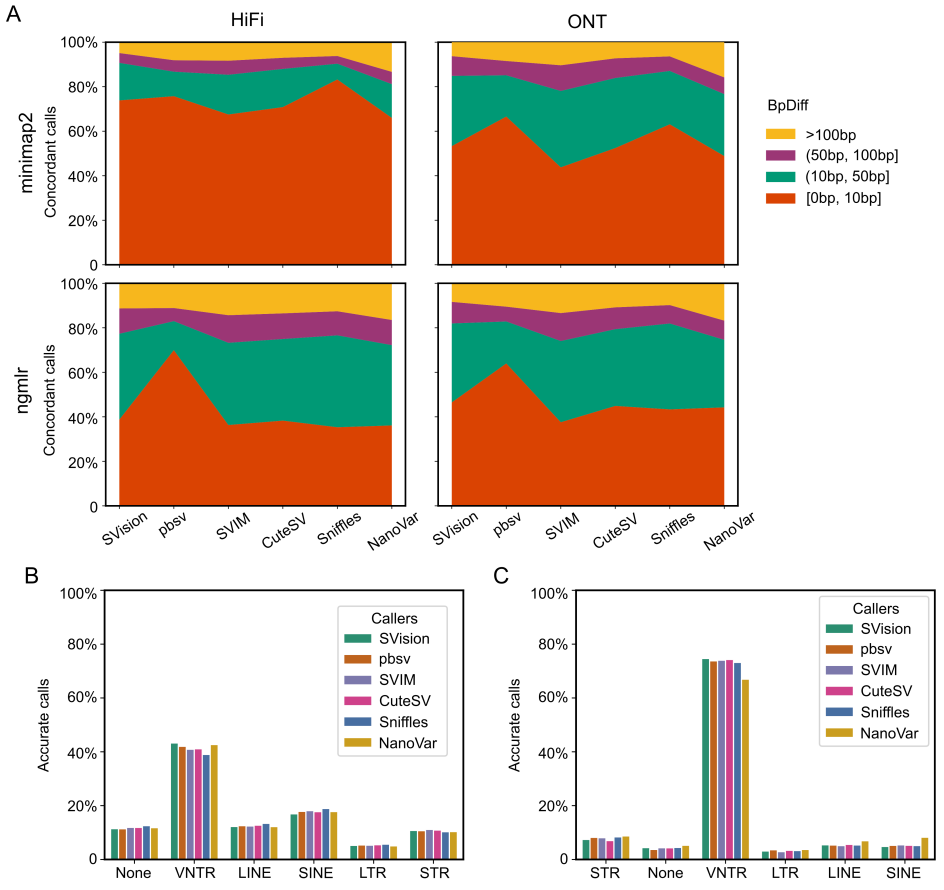

Figure 5.5: Evaluating the breakpoint accuracy of structural variants detected by six callers with Phased Assembly Variant. (A) The breakpoint difference (*BpDiff*) of concordant calls between callers' detections and Phased Assembly Variants (PAVs). (B) The repeat annotation of accurate calls (*BpDiff* $\leq$ 50bp). (C) The repeat annotation of inaccurate calls (*BpDiff* $>$ 50bp).

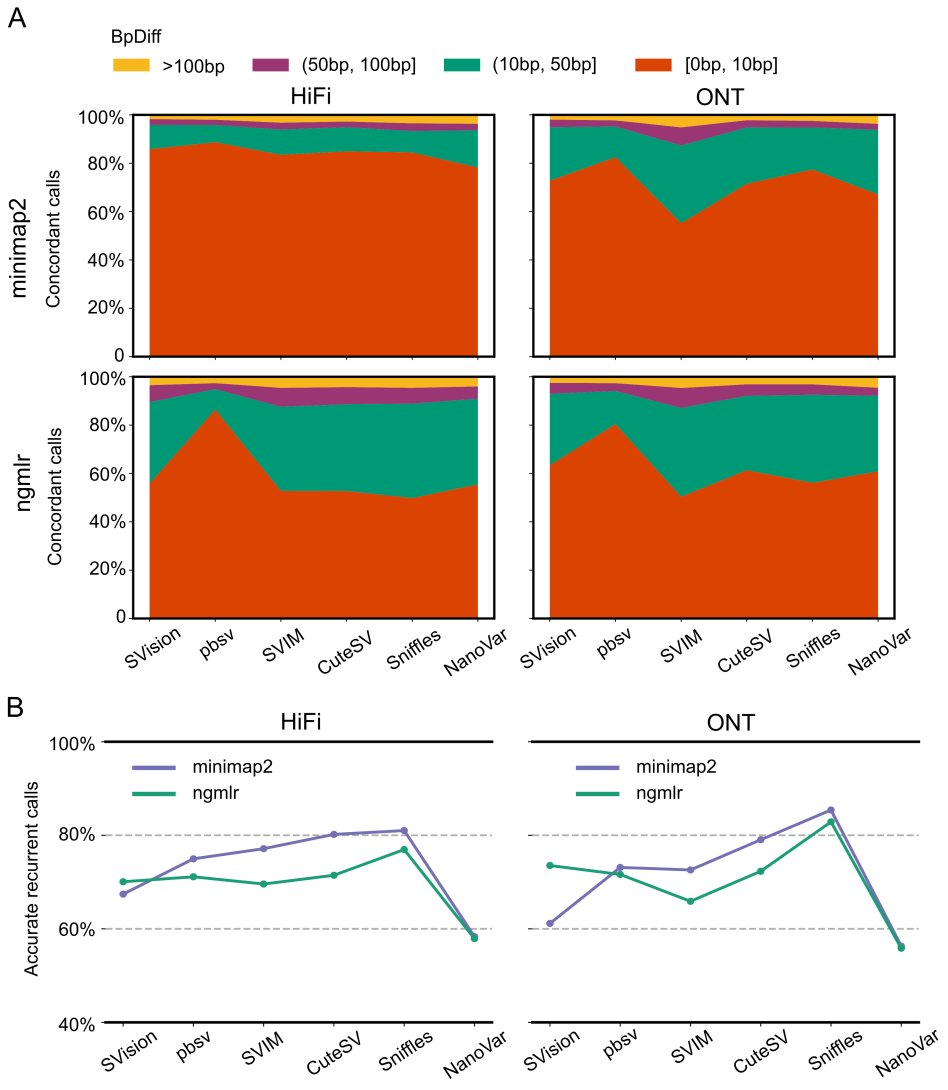

Figure 5.6: Evaluating the breakpoint accuracy with short-read data and assessing breakpoints of recurrent structural variants. (A) The breakpoint difference (BpDiff) of structural variants (SVs) detected by six callers and those detected by short-read data. (B) The breakpoint accuracy of recurrent SVs among three samples (i.e., NA19240, HG00733 and HG00514).

and it was independent of aligners (Figure 5.7C). Furthermore, 1,500 putatively somatic translocations were identified from pbsv calls using ngmlr alignments, which was 15 times more than translocations identified from minimap2 alignments. In addition, we assessed the putative somatic SVs with the COLO829 somatic benchmark, containing 78 (i.e., 38 deletions, 13 translocations, 7 duplications, 7 inversions and 3 insertions) high-quality SVs released by a multi-platform study. As a result, though 57 ground truth somatic SVs were missed by one of the five callers, all somatic insertions were correctly detected. In addition, 35 out of 57 ground truth SVs, consisting of six translocations, 21 deletions, five inversions and three duplications, could not be detected by any combination of callers and aligners. We thus reasoned that integration of discoveries from different callers might substantially increase the detection sensitivity.

5.4 Conclusion

SVs are important types of genomic alterations to form population diversity [5] and to drive disease progression, such as tumorigenesis [6], but are more difficult to detect than small variants from short-read data due to the limited read length. In the past five years, the long-read sequencing technologies and the newly developed algorithms greatly facilitate the detection of SVs from both healthy [113] and tumor genomes [114], improving our understanding of the functional impact of SVs. Remarkably, the SV detection based on haplotype-resolved assembly enables the haplotype-aware germline SV detection, and significantly improves the detection at complex genomic regions, such as segmental duplication and variable number tandem repeat (VNTR) [9, 10]. Though studies have attempted to evaluate the performance of routine SV detection algorithms, we explored the major factors affecting the ability of different algorithms in detecting SVs in complex genomic regions and somatic SVs. Overall, using public HiFi and ONT data from four healthy genomes and ONT data from a normal-tumor paired sample, we evaluated multiple aligners and SV callers to assess the routine SV detection algorithms by comparing with PAV calls and high-quality somatic truth set.

In this chapter, we examined the performance of each SV caller with two aligners (i.e., minimap2 and ngmlr). The alignment time and memory usage had been systematically evaluated in other studies [115], which was out of the scope of this study. For both HiFi and ONT platforms, all callers tend to detect more SVs on minimap2 aligned data than that of ngmlr, while SVIM produced more ONT unique calls on both aligners. Since the same parameters

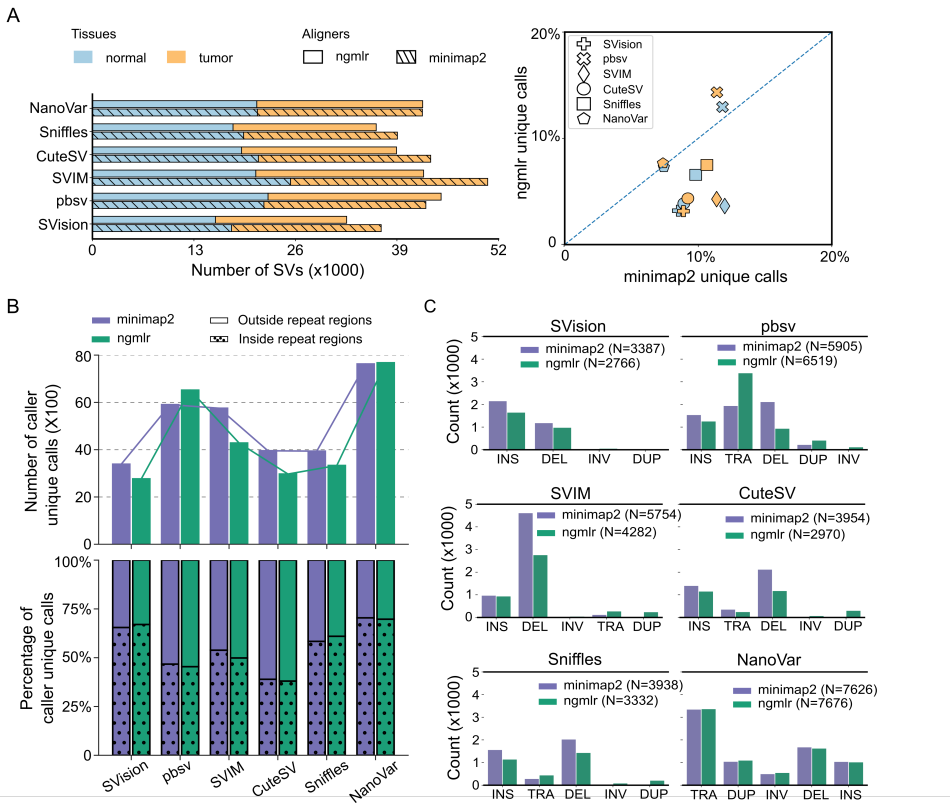

Figure 5.7: Evaluating the filtering-based somatic structural variants detection of the six callers. (A) Comparison of structural variants (SVs) detected from both tumor and normal tissues (left panel) and percentage of aligner unique calls detected by each caller (right panel). (B) Percentage of tissue unique calls detected by each caller (top panel) and repeat annotation of caller unique calls (bottom panel). (C) SV types of tumor unique calls identified from each caller.

were used for each caller on different platforms and aligners, SVIM might need specific parameter tuning for ONT data. Moreover, we found that aligner was the major factor affecting the number of detected SVs and their breakpoint accuracy, whereas the breakpoint of pbsv were less affected by aligners and platforms. Therefore, we recommend using pbsv with either minimap2 or ngmlr for the initial SV for a new sample. In terms of the recall and precision of callers, both the GIAB and PAV benchmarking suggested the bias of minimap2 paired with HiFi data. Though these two benchmarks showed limitations for evaluation, they suggested that SVision, Sniffles, pbsv, CuteSV and SVIM showed similar performance and outperformed NanoVar. In addition, Sniffles and CuteSV showed the highest precision for all of the HiFi and ONT data tested, while SVision call sets generally had a higher recall rate. Therefore, Sniffles and CuteSV should be used when high precision was the priority, pbsv was recommended when accurate breakpoint were required, and SVision should be considered if high sensitivity was desired.

Additionally, and uniquely to this study, we investigated the features of PAV calls missed by read-based detection to assess whether read-based calling was capable of generating comprehensive call set. It was expected that most of the missed PAV calls were found at VNTR regions and insertion was the major SV type missed by read-based detection. While our results suggested that the majority of the missed PAV loci contained SV signature reads, and most importantly, this was not depending on aligners, indicating the read-based detection would recover most of the PAV calls.

Moreover, since we also observed high SV breakpoint concordance on different platforms using identical aligner, the selection of sequencing platform would have less impact on SV detection for a new sample. However, it should be noted that the majority of the inaccurate and inconsistent calls were found at tandem repeat regions, so that disease associated with repeat expansion requires extra downstream analysis or specific algorithms, such as Straglr [116] and NanoSatellite [101]. Another critical step in studying tumor genomes was to characterize the somatic SVs, which were considered closely related to the tumorigenesis. Due to lack of long-read based somatic SV detection algorithms, we only evaluated the recall of detecting somatic SVs in tumor unique calls. However, this approach identified ground truth somatic SVs in low precision, suggesting an urgent demand of standalone somatic SV detection algorithms in the community.

Altogether, our analysis suggested that alignment-based callers would uncover a near comprehensive and high-quality call set of a genome, while the filtering-based approach for somatic SV discovery was suboptimal, leading to high false positive rate. Thus, as the detection of SVs from long-reads becomes

routine and gradually applied to investigate tumor genomes, it is imperative to start to consider and work towards developing robust pipelines or algorithms for SV detection in tumors. Moreover, we expect resources from ONT and PacBio to accumulate as the technology improves and the sequencing price decreases, which leaves great opportunities for better somatic benchmark generation and future algorithm development for clinical applications.