Comparison and benchmark of structural variants detected from long read and long-read assembly

Lin J., Jia Peng, Wang Songbo, Kosters W.A., Ye K.

OXFORD

Briefings in Bioinformatics, 2023, **24(4)**, 1–12
https://doi.org/10.1093/bib/bbad188
Advance access publication date 17 May 2023
Review

Comparison and benchmark of structural variants detected from long read and long-read assembly

Jiadong Lin, Peng Jia, Songbo Wang, Walter Kosters and Kai Ye
Corresponding author. Kai Ye, School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. Tel.: +86-82664955; Email: kaiye@xjtu.edu.cn

Abstract

Structural variant (SV) detection is essential for genomic studies, and long-read sequencing technologies have advanced our capacity to detect SVs directly from read or de novo assembly, also known as read-based and assembly-based strategy. However, to date, no independent studies have compared and benchmarked the two strategies. Here, on the basis of SVs detected by 20 read-based and eight assembly-based detection pipelines from six datasets of HG002 genome, we investigated the factors that influence the two strategies and assessed their performance with well-curated SVs. We found that up to 80% of the SVs could be detected by both strategies among different long-read datasets, whereas variant type, size, and breakpoint detected by read-based strategy were greatly affected by aligners. For the high-confident insertions and deletions at non-tandem repeat regions, a remarkable subset of them (82% in assembly-based calls and 93% in read-based calls), accounting for around 4000 SVs, could be captured by both reads and assemblies. However, discordance between two strategies was largely caused by complex SVs and inversions, which resulted from inconsistent alignment of reads and assemblies at these loci. Finally, benchmarking with SVs at medically relevant genes, the recall of read-based strategy reached 77% on 5X coverage data, whereas assembly-based strategy required 20X coverage data to achieve similar performance. Therefore, integrating SVs from read and assembly is suggested for general-purpose detection because of inconsistently detected complex SVs and inversions, whereas assembly-based strategy is optional for applications with limited resources.

Keywords: long-read sequencing, structural variant detection, sequence assembly

INTRODUCTION

Structural variants (SVs, ranging from 50 bp to megabases of sequence) comprise different subclasses, such as deletions, insertions, etc., and play important roles in both healthy and disease genomes. Calling SVs between an individual's genome and a reference genome has been the standard practice for genomic studies since the completion of Human Genome Project. Over the past decade, researchers have made great progress in discovering and genotyping SVs in diverse populations with short-read data, but SVs such as insertions or those at repetitive regions remain challenging due to the limited read length [1]. Long-read sequencing technologies, such as Pacific Bioscience (PacBio) and Oxford Nanopore Technology (ONT), have emerged as superior to short-read sequencing for SV detection and thus reveal a number of novel SV functional impacts missed by short-read data [2–5]. Long-read also improved SV detection in genetic diseases [6–8] and cancers [9–15] where SVs are usually undetectable or misinterpreted by short-read [16]. In addition, long-read sequencing have inspired dramatic improvements of assembly methods and promote de novo assembly-based SV detection, such as the study conducted by the Human Genome Structural Variation Consortium, revealing 107 590 SVs with HiFi (High-Fidelity reads generated by PacBio CCS technology) assemblies, of which 68% are not discovered by short-read sequencing [2, 17].

Currently, almost all long-read based studies use either the read-based (i.e. detecting from read alignments) or the assembly-based strategy (i.e. detecting from assembly alignments) for SV detection. The assembly-based strategy is consisted of de novo assembly, alignment and calling, whereas the read-based strategy only contains alignment and calling. The calling step of the two strategies is similar and usually contains two parts. First, the SV 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, whereas inter-read alignments are usually obtained from the supplementary alignments [1]. Second, callers typically cluster and merge similar signatures from multiple aberrant alignments, delineating proximal signatures that support putative SV. Nearly all read-based callers developed in the past 5 years, such as Sniffles [18], pbsv, cuteSV [19], SVIM [20], NanoVar [21], NanoSV [22], and Picky [10], detect SVs through combinations of signatures obtained from inter-read and intra-read alignments but differ in their signature clustering heuristics. Although different from the above methods, SVision [23] applies a deep-learning approach to directly recognize different SV types from the variant signature sequences. Assembly-based callers, such as Phased Assembly Variant (PAV) and SVIM-ASM [20] also collect aberrant inter-contig and intra-contig alignments for SV detection. However, PAV adopts alignment trimming for the detection at complex genomic regions.

Jiadong Lin is an assistant professor of School of Automation Science and Engineering at Xi'an Jiaotong University.
Peng Jia is PhD student of School of Automation Science and Engineering at Xi'an Jiaotong University.
Songbo Wang is a PhD student of School of Automation Science and Engineering at Xi'an Jiaotong University.
Walter Kosters is an associated professor of Leiden Institute of Advanced Computer Science at Leiden University.
Kai Ye is a professor of School of Automation Science and Engineering at Xi'an Jiaotong University.
Received: January 16, 2023. **Revised:** April 25, 2023. **Accepted:** April 26, 2023

The above detection methods have undoubtedly deepened our understanding of SVs and their related biological and pathological process. Although a number of studies have demonstrated the advances of using long-read over short-read data [2, 24], the properties of SVs accessible to long read and long-read assembly are largely unknown. Here, we used HG002 genome and its well-curated benchmarks from Genome in a Bottle Consortium (GIAB) to compare benchmark and establish expectations for these two strategies. We selected four read aligners, three assemblers, one contig aligner, five read-based callers and two assembly-based callers according to methods reviewed by a recent study [25] (Methods). Based on SVs detected from three HiFi and three ONT datasets of different sequencing libraries (Supplementary Table S1), we assessed these two strategies from the following perspectives (Figure 1): (i) the impact of dataset, aligner and assembler on each strategy, (ii) the concordant SVs between read and assembly as well as SVs specifically detected from read and assembly, (iii) the recall and precision of detecting SVs at simple and complex genomic regions. Our findings would help researchers choose appropriate detection strategy for their genomic studies, and they also help developers to know how dataset-specific attributes, various methods and variant types influence the performance of these detection strategies.

RESULTS

Impact of dataset, aligners and assemblers on each strategy

Overall, assembly-based strategy detects more SVs, especially insertions, than read-based strategy and the difference was even remarkable on ONT datasets (Figure 2A, Figure S1). For both strategies, around 78 and 82% of the WGS-SVs (SVs at whole genome scale) and ExTD-SVs (SVs outside of tandem repeat regions) were concordant among three HiFi datasets, respectively (Figure 2B, Figure S2A). On the contrary, the percentage of concordant SVs detected from ONT datasets was much lower and divergent for assembly-based strategy. For example, the percentage of concordant WGS-SVs was around 30 and 65% when detected from assemblies created by shasta and flye, respectively (Figure 2B, Figure S2B). HiFi datasets also enabled accurate breakpoint identification, where both strategies could detect 15% more BSD-0 (BSD equals 0 bp, BSD indicates breakpoint standard deviation) WGS-SVs and 27% more BSD-0 ExTD-SVs than that detected on ONT datasets (Figure 2C). Remarkably, for assembly-based strategy, the percentage of BSD-0 WGS-SVs was similar to that of BSD-0 ExTD-SVs detected on HiFi datasets, suggesting TD-SVs (i.e. SVs at tandem repeat regions) were also consistently detected from assemblies generated on different datasets (Figure 2C). For read-based strategy, only pbsv and Sniffles detected the most accurate breakpoint for datasets concordant SVs, and it was comparable to assembly-based strategy (Figure 2C).

In addition, when detecting SVs with the same caller but different aligners or assemblers on a dataset, the read-based strategy was greatly affected by the choice of read aligners comparing to the influence of assemblers on assembly-based strategy. For example, calling from HiFi datasets, around 76 and 48% of the WGS-SVs were concordant among two assemblers and four read aligners, respectively (Figure 2D). We further compared two aligners and two assemblers to avoid bias caused by an unequal number of methods, where the SV concordant rate of two assemblers was still higher than the highest one achieved by any aligners pair (Figure S3A). We then assessed the breakpoint identity of the assembler and read aligner concordant SVs. Specifically, most of the concordant SVs detected by assembly-based strategy shared identical breakpoints, whereas similar performance could only be achieved by read-based strategy when pbsv was used for calling (Figure 2E). On ONT dataset, the percentage of BSD-0 SVs among concordant SVs detected by pbsv, PAV and SVIM-asm increased as the read length increases. Interestingly, given that incremented versions of Guppy were used as base caller for three ONT datasets, the percentage of breakpoint identical concordant SVs was also positively correlated with the read length (Figure 2E). This is because breakpoint identity largely relies on read mapping accuracy, which could be improved by the higher base-calling accuracy.

We next investigated the impact of aligners and assemblers on SV types. For read-based strategy, each caller was applied to the alignments generated by the four aligners, from which read-aligner-specific SVs were obtained for further analysis (Figure S3B). Overall, the size distribution of aligner-specific SVs differed between read aligners. For example, an SV size peak at 300 bp was observed for SVs detected from ngmlr-aligned HiFi and ONT reads, while more SVs smaller than 100 bp were detected from lra-aligned ONT reads (Figure 2F). In addition, around 17, 39, 38 and 33% of the aligner-specific ExTD-SVs were deletions when detected from ngmlr, minimap2, lra and winnowmap alignments, respectively (Figure 2G). Notably, 37% of the ngmlr-specific ExTD-SV were duplications, but they were seldom observed in other aligners' specific ExTD-SVs. Such bias was also observed in pairwise aligner comparisons (Figure S3C). This aligner-induced SV type and size bias was largely due to the mapping strategy adopted by ngmlr, which splits reads into non-overlapping 256 bp sub-reads and maps them independently of each other [18]. As for the impact of different aligners on each caller, the SV types predicted by pbsv were less consistent than other callers among different alignments (Figure S4A). Moreover, we examined whether different callers would predict the SV of the same type from one alignment. As a result, the percentage of discordant SV type was 11.83, 12.08, 12.29 and 18.22% when SVs were detected from lra, minimap2, winnowmap and ngmlr alignments, where insertions and duplications contributed to the majority of the discordant SV types (Figure S4B). For assembly-based strategy, the impact of contig aligner and assembler on SV types was examined separately. In general, we found that the SV type bias was more likely to be affected by sequencing platforms rather than contig aligners or assemblers. For instance, among the assembler-specific ExTD-SVs (Figure S5A), we only noticed variability caused by assemblers on the ONT dataset. Specifically, 88% of the specific ExTD-SVs detected from shasta assemblies were insertions, which was 13% higher than the average percentage of assembler-specific insertions detected from HiFi assemblies (Figure S5B), whereas an opposite trend for the distribution of insertions and deletions was observed on ONT dataset among contig aligner-specific ExTD-SVs (Figure S5C).

Accordingly, we concluded that both strategies were able to detect SVs consistently among HiFi datasets, but assembly-based strategies identified more accurate breakpoints without being affected by the dataset or assembler. Though read-based strategy was versatile to different sequencing technologies, the reported variant size, type and breakpoints were greatly affected by aligners, even for SVs outside of tandem repeat regions.

Analysis of SVs captured by both assembly and read

On each dataset, 20 read-based callsets and 4 assembly-based callsets were obtained from different callers, and subsequent comparisons were made between read-based callsets and assembly-based callsets (Methods). On average, a remarkable subset of ExTD-SVs (81% for assembly-based and 82% for read-based)

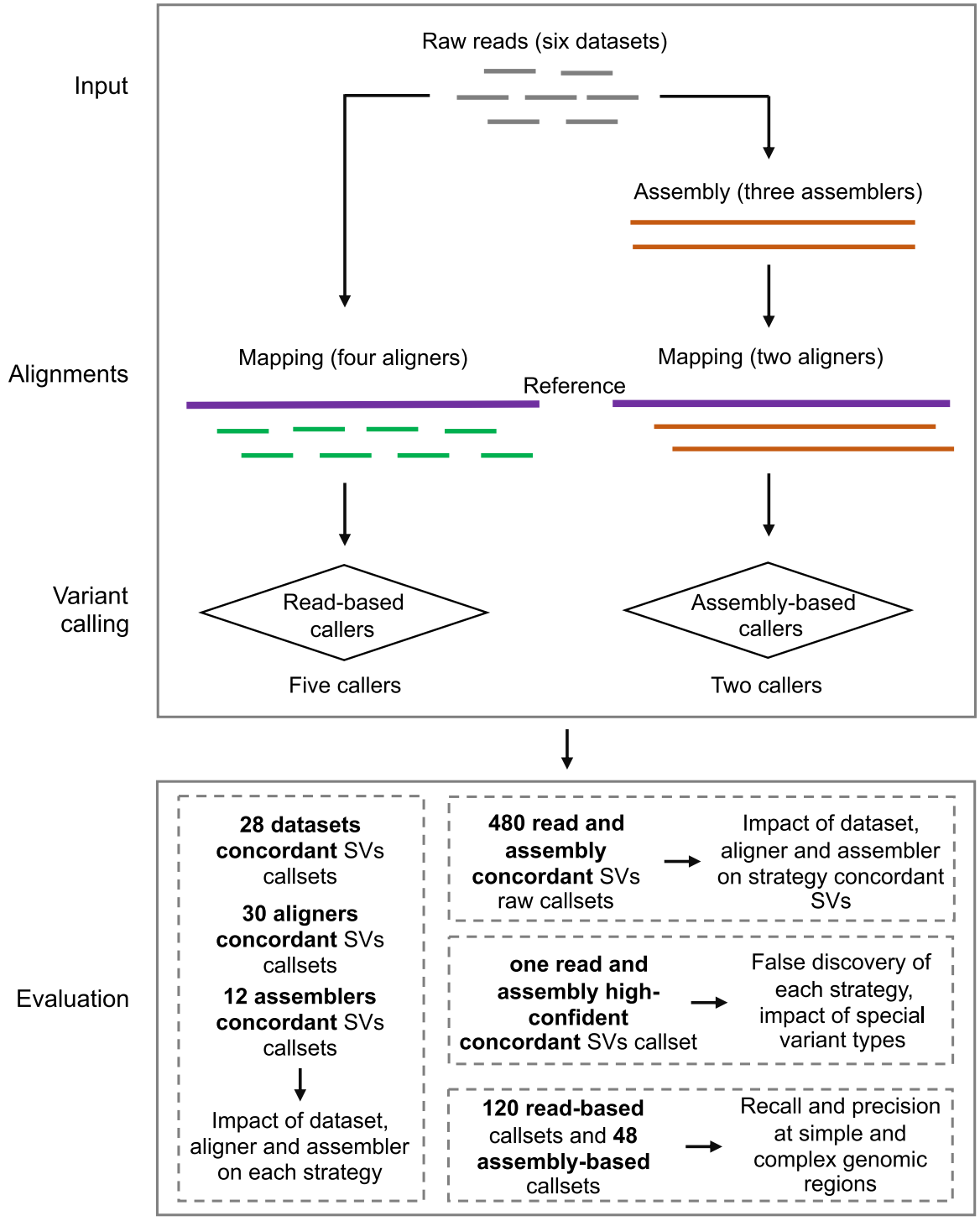

Figure 1. Schematic summaries of assessing two strategies. Based on three HiFi and three ONT datasets, we used four read aligners (i.e. mininimap2, ngmlr, winnowmap and lra) and five read-based callers (i.e. SVIM, pbsv, SVision, cuteSV and Sniffles) to generate the 120 read-based callsets. Moreover, we used two assembly aligners (i.e. minimap2 and lra), three assemblers (i.e. hifiasm and flye for HiFi data, flye and shasta for ONT data) and two assembly-based callers (i.e. SVIM-asm and PAV) to generate 48 assembly-based callsets. On the basis of the 120 read-based and 48 assembly-based callsets, we (1) examined the impact of dataset, aligner and assembler on each strategy, (2) examined the impact of dataset, aligner and assembler on concordant SVs captured by read and assembly, (3) examined the false discoveries and the impact of variant types on high-confident insertions and deletions captured by both read and assembly and (4) measured the recall and precision of detecting SVs at simple and complex genomic regions.

were captured by both read-based and assembly-based callers when detected on HiFi datasets (Figure 3A). The percentage of concordant WGS-SVs and ExTD-SVs was also positively correlated with the read length because assemblers essentially created longer DNA sequences for SV detection (Figure 3A). We also noticed that minimap2 paired with hifiasm led to the highest percentage of concordant WGS-SVs and ExTD-SVs on HiFi datasets, whereas the highest percentage was achieved by flye paired with minimap2 on ONT datasets (Figure 3B). Besides, aligner was also critical for detecting consistent breakpoints

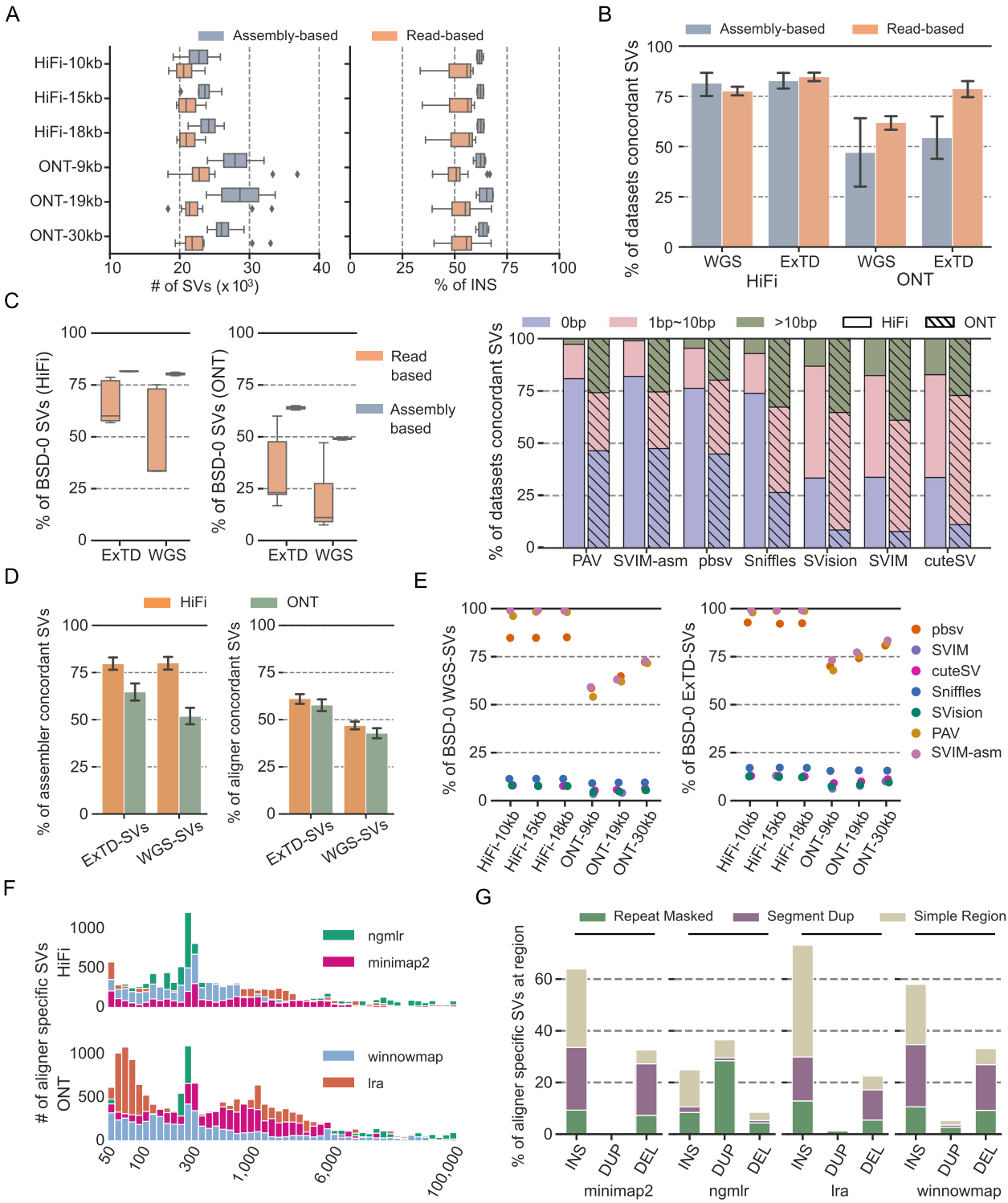

Figure 2. Overview of detected SVs and the detection variability of each strategy. (**A**) The number of SVs and the percentage of insertions detected by two strategies on each dataset. (**B**) The percentage of dataset concordant WGS-SVs and ExTD-SVs detected by two strategies on HiFi and ONT datasets. (**C**) The percentage of breakpoint identically (i.e. breakpoint standard deviation equals 0 bp, BSD-0 SVs) detected concordant WGS-SVs and ExTD-SVs among ONT and HiFi datasets as well as the distribution of concordant SVs' BSD of each caller. (**D**) The percentage of aligner and assembler concordant WGS-SVs and ExTD-SVs. (**E**) For each caller, the percentage of BSD-0 aligner concordant SVs and assembler concordant SVs detected on different HiFi and ONT datasets. (**F**) The size distribution of aligner-specific ExTD-SVs. (**G**) The SV types of aligner-specific ExTD-SVs at different genomic regions. WGS-SVs: SVs at whole genome scale, ExTD-SVs: SVs outside of tandem repeat regions.

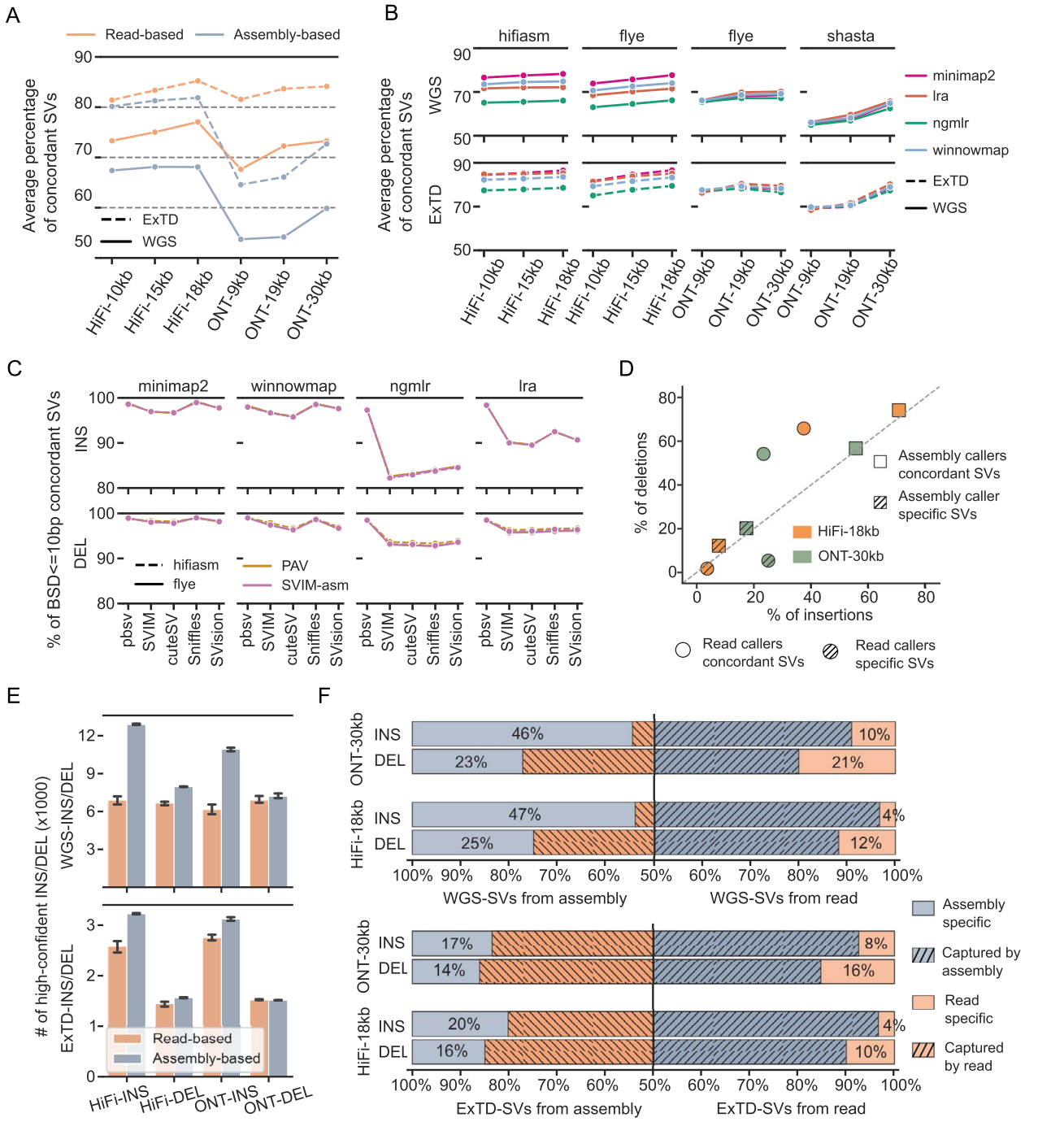

Figure 3. The analysis of SVs detected by both read and assembly. (**A**) The average percentage of concordant SVs among total number of SVs detected by read and assembly. (**B**) The impact of aligner and assembler on the percentage of concordant SVs between read and assembly. (**C**) The impact of aligner on the breakpoint of SVs captured by both read and assembly. (**D**) Using minimap2 for alignment and assemblies created by hifiasm (HiFi) and flye (ONT), the percentage of insertions and deletions detected by one of the callers or all callers. (**E**) The number of high-confident WGS-INS/DEL and ExTD-INS/DEL detected from HiFi and ONT datasets. (**F**) The percentage of concordant SVs and specific SVs detected on ONT-30 kb and HiFi-18 kb datasets. WGS-INS/DEL: insertions and deletions at whole genome scale, ExTD-INS/DEL: insertions and deletions outside of tandem repeat regions.

of SVs captured by both read and assembly. For instance, on HiFi-18 kb dataset, the highest percentage of BSD-10 (i.e. breakpoint standard deviation smaller than 10 bp) ExTD-INS/DEL (i.e. insertions and deletions outside of tandem repeats) was achieved when detected from minimap2-aligned reads and assemblies (Figure 3C). Therefore, minimap2 combined with hifiasm or flye would minimize the impact of assemblers and aligners on the number of concordant SVs and their breakpoints' identity.

However, even using minimap2 for alignment and hifiasm and flye for sequence assembly, around 60 and 65% of the deletions were detected by five-read based callers and two assembly-based callers, respectively, and it was even lower for insertions (Figure 3D). Thus, to avoid caller bias, we compared high-confident callset of insertions and deletions that were detected by all read-based callers and assembly-based callers (Figure 3E, Figure S6A). Further comparison revealed that the majority of the

insertions (90% on ONT-30 kb and 96% on HiFi-18 kb) detected by reads were also captured by assemblies, whereas only half of the insertions (46% on ONT-30 kb and 47% for HiFi-18 kb) detected by assembly were captured by read (Figure 3F, Figure S6B). For the ExTD-INS/DEL, a large subset of insertions and deletions could be captured by both assembly and read and we only observed slightly difference between sequencing technologies (Figure 3F). Of note, at non-tandem repeat regions, the percentage of assembly insertions captured by read increased to 80 and 83% on HiFi-18 kb and ONT-30 kb datasets, respectively. This result was significantly higher than previous reported insertions captured by both short-read and assembly at non-tandem repeat regions [24]. Therefore, we reasoned that most of the ExTD-INS/DEL were able to be detected by read and assembly without technology bias, whereas assembly-based strategy could resolve more SVs at tandem repeat regions.

Assessing SVs only accessible to assembly or read

We further investigated the factors that have the greatest influence on false discoveries and discordance between assembly and read calls. To accomplish this, we proposed an *in silico* trio-based SV assessment workflow, where paternal and maternal data and two evaluation tools (VaPoR [26] and TT-mars [27]) were used to examine the correctness of assembly and read specific ExTD-INS/DEL. Considering that some SVs were not able to be evaluated by VaPoR or TT-mars, our procedure classified these specifically detected SVs into: (i) Valid SVs were SVs validated in both parents or in one of the parents; (ii) Invalid SVs were SVs not validated in parents; (iii) Inconclusive SVs were SVs only evaluated in one of the parents or none of the parents (Methods). Excluding inconclusive SVs, the average validation rate for read-specific ExTD-INS/DEL was higher than assembly-specific ExTD-INS/DEL, especially on ONT-30 kb dataset. For read-specific ExTD-INS/DEL, 89.2 (207 out of 232 evaluated SVs) and 88.3% (379 out of 429 evaluated SVs) of them were validated on HiFi-18 kb and ONT-30 kb, respectively (Figure 4A, Figure S6C). In contrast, the validation rates for assembly-specific ExTD-INS/DEL were 85.3 (635 out of 745 evaluated SVs on HiFi-18 kb) and 75% (533 out of 644 evaluated SVs on ONT-30 kb) (Figure 4B). We also observed a big difference between the percentage of validated deletions on ONT-30 kb dataset, where 59.5 and 86.8% of the assembly-specific and read-specific deletions were correct discoveries, respectively (Figure 4A and B). Moreover, the genomic hotspots for those invalid assembly ExTD-INS/DEL also varied between sequencing technologies, such as more than half of the invalid assembly-specific ExTD-SVs detected from HiFi-18 kb were enriched at RepeatMasker (RM) annotated regions, but 90% of the ONT-30 kb invalid SVs were located in segmental duplication (SD) regions (Figure 4C).

Furthermore, we analyzed the HG002's read alignments at 110 (43 deletions and 67 insertions detected on HiFi-18 kb) invalid assembly-specific SV loci and revealed that 60 out of 110 loci were not supported by any SV signature reads, indicating they were substantial false discoveries (Figure 4D). For the rest of the 50 SV loci containing signature reads, our analysis revealed that seven SVs were supported by less than five read signatures, eight SVs were misclassified as assembly-specifics because these SVs were not detected by all five read-based callers, and there were only two assembly novel SVs (Figure 4D). Notably, 33 of the 50 SV loci contained inversion and complex SV (CSV) signatures derived from read alignments, but more than one SV signatures were identified from assembly alignments (Figure 4D, Supplementary File 1). For instance, a ~1.3 kb inversion was detected as 683 bp insertion and 684 bp deletion by assembly-based callers (Figure 4E). Moreover, a CSV locus of ~3kbp, consisting of a deletion and two segment switch, was detected as three separate SVs (a 320 bp insertion, a 305 bp deletion and a 1867 bp deletion) from assembly (Figure 4F). This phenomenon was mainly caused by ambiguous alignments at inversion and CSV loci, such that the 33 assembly-specific SVs were considered as different calls from those detected from read alignments. Taken together, the above results suggested that around 90% of the read and assembly-specific ExTD-SVs were true positive discoveries when detected from HiFi dataset, thereby indicating the integration of read and assembly calls was necessary for comprehensive SV profile. In addition, SV loci containing inversion or CSV were not only difficult to assess by existing evaluation tools but also challenging to detect consistently from read and assembly.

Performance at genomic regions of different complexity

So far, GIAB had released one truthset containing SVs at simple regions, and another one at medically relevant genes (CMRGs) surrounded by complex genomic context. To evaluate, SVs were called on reads and assemblies aligned with minimap2 to avoid the performance bias caused by aligners (Methods). For the SVs at simple regions, the highest recall was achieved by assembly-based strategy but read-based strategy resulted in the highest precision (Figure 5A). Moreover, the recall of both strategies was positively correlated with read length on both HiFi and ONT datasets, but large precision variance was observed for assembly-based calls on ONT datasets (Figure 5A, Supplementary Table S2). As for SVs at CMRGs, the highest recall of the assembly-based strategy was 96% and it was 7% higher than the highest one achieved by read-based strategy (Figure 5B, Supplementary Table S3). We further compared the false negative (i.e. missed benchmark SVs) and false positive (i.e. novel SVs outside of benchmark) discoveries assessed by SVs at CMRGs. For example, 71 (54/76, read-based) and 53% (26/49, assembly-based) of the false negative discoveries were concordant among HiFi datasets, whereas only 30% of the false negatives detected from assemblies were concordant among ONT datasets (Figure 5C, Figure S7A). In addition, the percentage of the false positive SVs only accessible to one dataset was even higher for assembly-based strategy, i.e. 73% for HiFi and 80% for ONT (Figure 5C, Figure S7B), suggesting that it was difficult to create consistent assemblies at CMRGs consisted of repetitive elements. We also noticed that the two strategies detected more concordant false negatives, whereas false positives were found to be strategy specific (Figure 5D).

In addition, CMRGs were well-documented genes across multiple diseases but often excluded from standard targeted or whole-genome sequencing analysis [28], enabling the evaluation for potential clinical application. The above analysis used the 35X coverage datasets, which was not applicable to clinical settings due to the high sequencing cost and computational cost. Therefore, we subsampled the 35X coverage datasets to 5X, 10X and 20X coverage and examined the performance of each strategy. Overall, the read-based strategy outperformed the assembly-based strategy on both HiFi and ONT datasets when the coverage was below 20X (Figure 5E, Figure S7C, Supplementary Table S4). Especially for 5X ultra-low coverage data, the average recall of the read-based strategy was 78% for HiFi-18 kb and ONT-30 kb, where SVision and cuteSV outperformed others (Figure 5F). Accordingly, we reasoned that the assembly-based strategy required at least 20X coverage data to perform similarly to the read-based strategy on 5X coverage data, thereby suggesting a great potential of the read-based strategy for clinical applications.

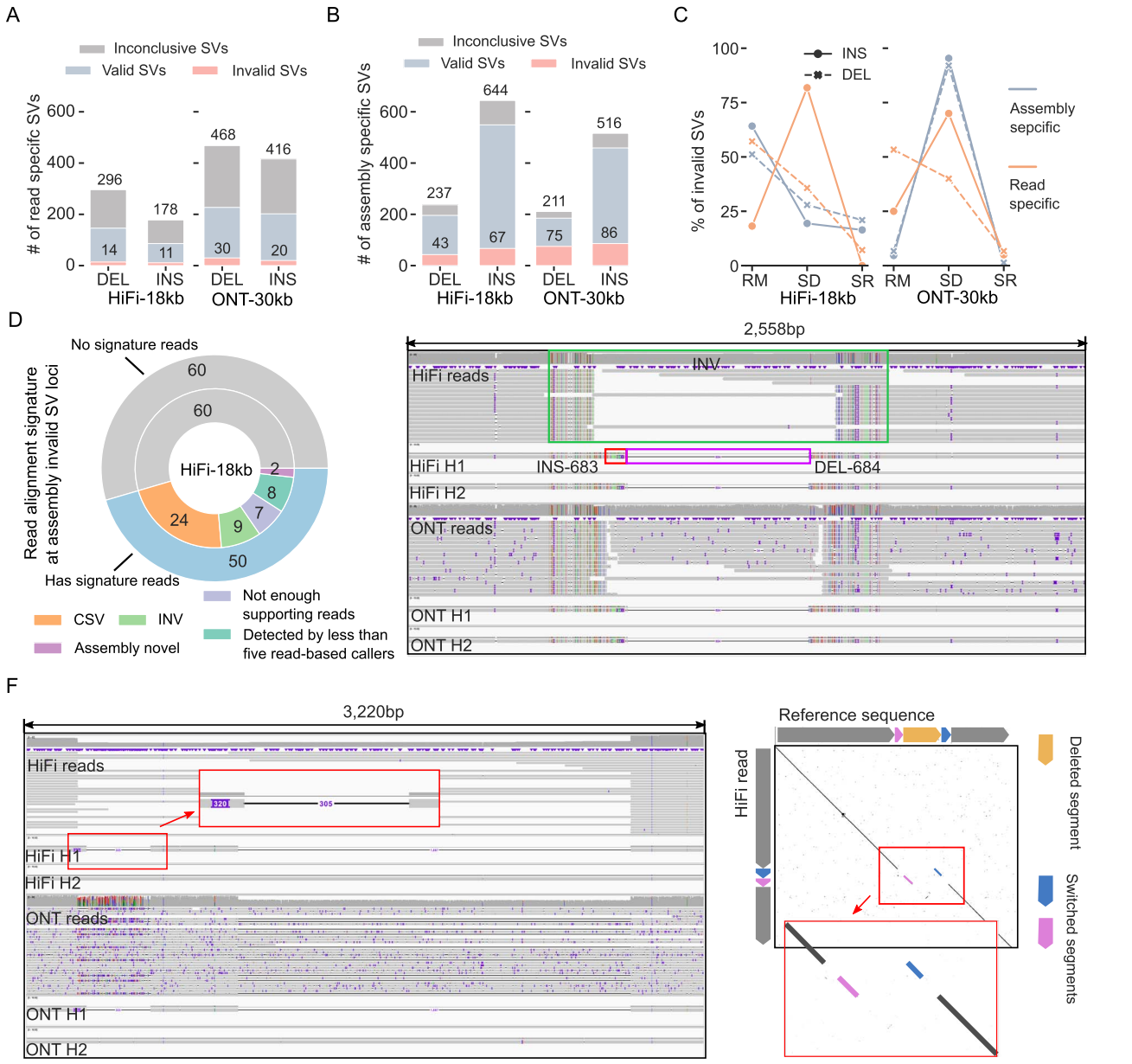

Figure 4. The validation and analysis of strategy-specific SVs. (**A**) The results summary of validating read-specific SVs with VaPoR, including inconclusive SVs, valid SVs and invalid SVs. (**B**) The results summary of validating assembly-specific SVs with TT-mars, including inconclusive SVs, valid SVs and invalid SVs. (**C**) The genomic region distribution of invalid SVs (RM: RepeatMakser annotated regions, SD: segmental duplication, SR: simple regions). (**D**) The analysis of invalid assembly-specific SVs. (**E**) The IGV alignment view of an inversion, where the read alignment signature is inconsistent with the SV size detected by assembly-based callers (i.e. a 683 bp insertion and a 684 bp deletion). (**F**) The IGV alignment view (left) and the diagram (right) of a CSV, spanning 3 kb of the genome, which is detected as three single SVs (a 320 bp insertion, a 305 bp deletion and a 1867 bp deletion) by assembly-based strategy.

CONCLUSION

In this study, we compared and investigated the factors that affect the most widely used long-read based SV detection strategies. We also analyzed and emphasized the discordance caused by CSVs and inversions, which were excluded in the previous comparison study of short-read and assembly-based detection. This is an important step toward the in-depth understanding of the usability and stability of detecting SVs from read or assembly.

In general, our analysis first showed that the assembly-based strategy was able to detect SVs consistently from HiFi assemblies without assembler bias, especially for SVs at complex genomic regions. In contrast, the read-based strategy was versatile to different sequencing technologies, but it was greatly affected by aligners, such as an unexpected large number of duplications detected from ngmlr-aligned reads. Second, by comparing SVs detected from read and assemblies, our analysis revealed a positive correlation between the percentage of concordant SVs and read length. Incorporating recent achievements in generating reads of 4Mbp and longer [29], the percentage of concordant SV is expected to be even higher. In addition, at non-tandem repeat regions, encompassing most the protein-coding sequences, a remarkable subset of high-confident ExTD-INS/DEL were captured by both read and assembly, and 90% of the read and assembly-specific SVs detected from HiFi dataset were validated

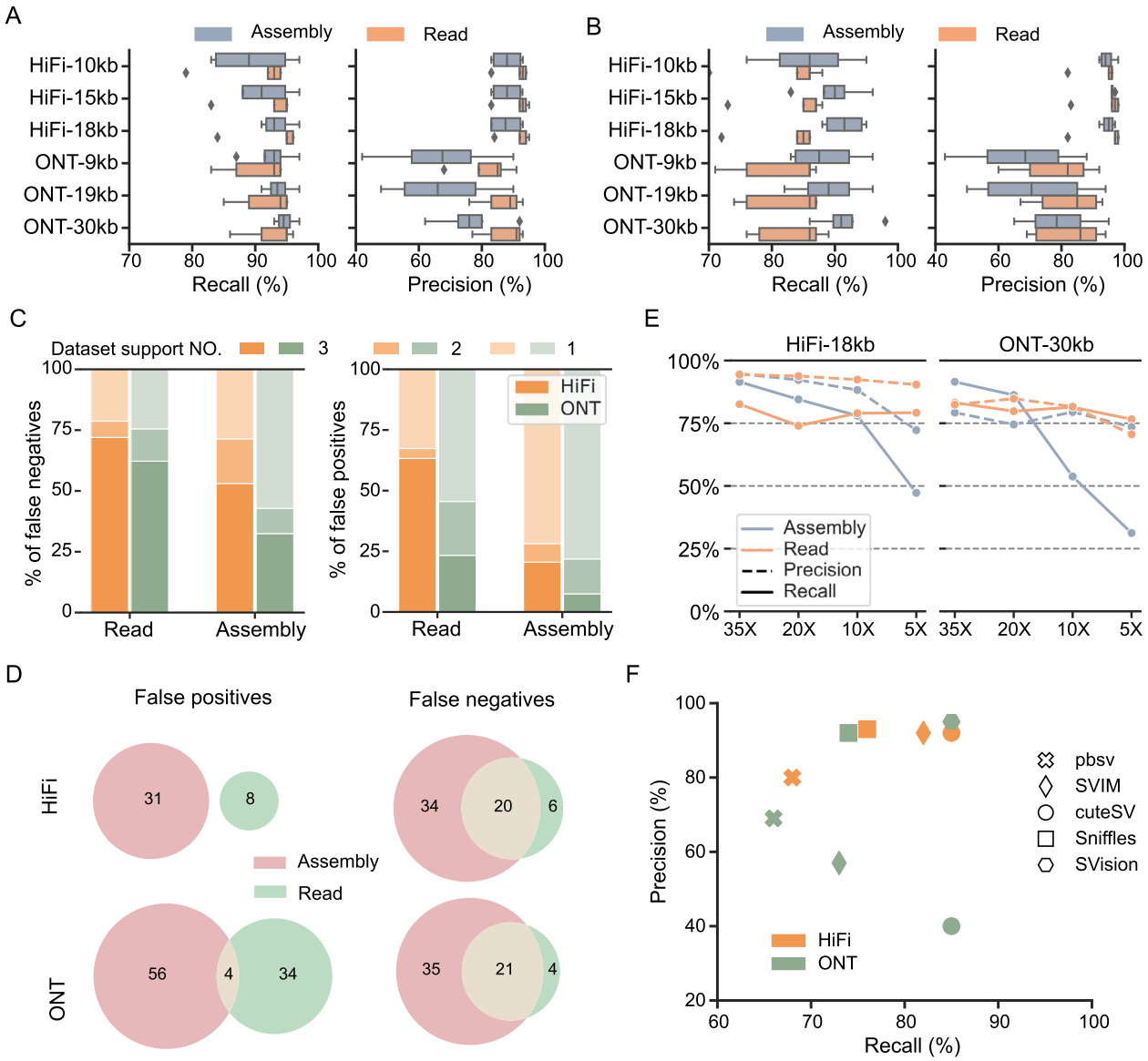

Figure 5. Summaries of benchmarking callers with well-curated SVs. (**A**) The recall and precision of detecting SVs at simple regions. (**B**) The recall and precision of detecting SVs at challenging medically relevant autosomal genes (CMRGs) surrounded by repetitive elements. (**C**) Evaluated with SVs at CMRGs, percentage of false positive and false negative SVs among HiFi and ONT datasets i.e. SVs in three, two and one dataset. (**D**) The Venn-diagram of false positive and false negatives detected by both strategies on HiFi and ONT datasets. (**E**) The impact of sequencing coverage on the recall and precision of detecting SVs at CMRGs. (**F**) At 5X coverage, the recall and precision of each read-based callers benchmarked with SVs at CMRGs.

by our *in silico* assessment workflow. Most importantly, for those invalid assembly-specific SVs supported by SV read signatures, multiple alignment signatures induced by CSV and inversion were identified in 66% of these loci and thus made them difficult to detect consistently. Finally, considering the sequencing and computational cost for future clinical application, the read-based strategy was able to effectively detect SVs at CMRGs with 5X coverage data, and the sensitivity was 25% higher than assembly-based strategy. As for the potential limitations, a collection of high-confident simple and complex regions is anticipated to further benchmark the two strategies, because the current GIAB high-confident simple genomic regions are unable to assess SV detection at complex regions. The evaluation is also limited to diploid genomes, whereas the performance of two strategies on cancers, affecting by purity, heterogeneity and aneuploidy, requires further investigation. Moreover, with the recent advances in graph-based pangenome study, it is also critical to assess several new strategies in the future, such as aligning reads or assemblies to a graph-based pangenome.

Taken together, the assembly-based strategy will access complex genomic regions that are intractable to read-based strategy, which is expected to provide exciting new insights into the SVs at such regions. We confirmed that reads and assemblies are usually inconsistently aligned at genomic loci containing inversions and CSVs, thereby leading to discordance or even false discoveries. Therefore, integration of read-based and assembly-based calls is necessary for comprehensive SV detection, even for SVs at non-tandem repeat regions. Moreover, the performance of assembly-based strategy is especially pronounced at complex genomic regions, such as SVs at CMRGs could be more accurately detected from assembly than read. As the long-read sequencing price drops to $1000 for a human genome, we expect this work

will help users to select proper SV detection strategy for different applications and foster future development of SV detection algorithms at complex genomic regions.

METHODS

Read mapping and sequence assembly

The three HiFi datasets (i.e. HiFi-10 kb, HiFi-15 kb and HiFi-18 kb) and the three ONT datasets (i.e. ONT-9 kb, ONT-19 kb, ONT-30 kb) are all publicly available. Based on a recent review [25], aligners containing minimap2, lra, winnowmap and ngmlr were included in our study, and assemblers including hifiasm, flye and shasta were used.

First of all, HiFi and ONT reads were mapped to human reference genome hg19 with minimap2 (v2.20), lra (v1.3.2), winnowmap (v2.03) and ngmlr (v0.2.7). Parameters used for each mapper were as follows:

- 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, and '*-a -z 600,200 -x map-ont*' were used for ONT reads.
- ngmlr: parameters '*-x pacbio*' and '*-x ont*' were used to align HiFi and ONT reads, respectively.
- winnowmap: parameters '*-ax map-ont*' and '*-ax map-pb*' of winnowmap were used to map ONT and HiFi reads, respectively.
- lra: '*-CCS*' and '*-ONT*" were set to map HiFi and ONT reads, respectively. We then applied each read-based caller with default parameters except for the minimum number of SV supporting reads. Since the sequencing coverage was around 35X for all datasets, the minimum SV supporting read for each read-based caller was set to five for the detection of both homozygous and heterozygous SVs. For 5X coverage, the minimum SV supporting read for each read-based caller was set to one.

For sequence assembly, we use minimap2-aligned reads and phased SNPs released by GIAB to obtain phased reads via the whatshap [30] 'haplotag' option. Those unphased reads are randomly assigned as either haplotype 1 and haplotype 2, which are also used in further sequence assembly. Given the phased reads, we then apply assemblers with default parameters to create the haplotype-aware assemblies.

SVs detection and post-processing

To detect SVs, methods were further excluded from the recent review [25] based on several criteria: (a) lack of detailed user manual; (b) no programming interface; (c) reported bias on aligners; (d) unresolved errors during wrapping. In the end, cuteSV (v1.0.10), pbsv (v2.2.2), SVIM (v1.4.0), Sniffles (v1.0.12) and SVision (v1.3.6) were selected as callers for read-based strategy, whereas PAV and SVIM-asm were selected for assembly-based strategy. Note that SVIM and SVIM-asm are two independent methods, where SVIM-asm requires genome assemblies as input for SV detection.

Read-based callers were directly applied to reads aligned by minimap2, ngmlr, lra and winnowmap with default parameters. Note that the minimum SV supporting read is set to five so that both homozygous and heterozygous germline SVs can be effectively detected from the 35X coverage datasets. For the assembly-based strategy, five most-widely used assembly aligners (i.e. LAST, MUMmer, minimap2, Cactus and SibeliaZ) were reviewed by Wouter De Coster et al, whereas PAV and SVIM-asm were only compatible with the alignments generated by minimap2 or lra. For PAV, the phased assemblies were directly used as input for the detection, and we run PAV with default parameters. For SVIM-asm, assemblies were first mapped to reference hg19 with minimap2 parameters '*-x asm20 -m 10000 -z 10000,50 -r 50000 –end-bonus=100 –secondary=no -O 5,56 -E 4,1 -B 5 -a*', these parameters were used in minimap2 embedded in PAV. Then, we run SVIM-asm with parameters '*svim-asm diploid –tandem_duplications_as_insertions –interspersed_duplications_as_insertions*' for SV detection.

For each callsets, a BED file obtained from a publication [31] was used to exclude SVs located at centromere and other low mapping quality regions. SVs overlapped with regions in the BED file were ignored in the downstream analysis. For the rest of the autosome SVs, we then annotated their associated repetitive elements using Tandem Repeat Finder, RM and SD results provided by UCSC Genome Browser. The original files downloaded from the genome browser were first processed based on scripts introduced by CAMPHOR [32] (https://github.com/afujimoto/CAMPHOR). Repeat element associated with each SV is assigned based on a recent publication [33]. In particular, Variable Number Tandem Repeat (VNTR) was assigned if the length of the repeat unit was longer than 7 bp; otherwise, we considered it as Short Tandem Repeat (STR). It should be noted that simple repeat annotated by RM was also classified into VNTR and STR. For SVs overlapping repetitive element, we prioritized the highest percentage of overlaps on the entire length of SV when multiple repeat types are annotated. For example, if 70% of an SV was composed of STR and 50% of the SV overlapped by ALU, then STR was assigned correspondingly. Moreover, according to the repetitive elements, we divided the genome into four different regions i.e. Tandem Repeat, Repeat Masked, Segment Dup and Simple Region. Tandem Repeat represented regions containing either VNTR or STR. Repeat Masked were those annotated as SINE, LINE, etc., by RM. Segment Dup represented regions overlapping with SDs. The rest of the genomic regions outside of Tandem Repeats, Repeat Masked and Segment Dup were termed as Simple Regions.

Analysis of concordant and specific SVs

The SV comparison tools are selected based on a recent review [25], which introduces Jasmine, SVanalyer, Truvari [34] and SURVIVOR. Jasmine is selected because of the following two reasons. First, we need to compare more than two callsets at once, whereas SVanalyzer and Truvari is only applicable to two callsets. Second, the Mendelian discordance of Jasmine is lower than that of SURVIVOR [35], despite their similar appearance.

According to different comparison purpose, we first obtained the nonredundant SVs of different callsets by the running command '*Jasmine file_list=vcf_list.txt out_file=nonredundant_SVs.vcf max_dist=1000 spec_len=50 spec_reads=1*'. Then, using the VCF file generated by Jasmine, we were able to identify concordant and specific calls as well as the breakpoint standard deviation of concordant SVs. The breakpoint standard deviation was indicated in 'STARTVARIANCE' and 'ENDVARIANCE' in the Jasmine merged VCF file, which were directly used to analyze the breakpoint consistency of concordant SVs. The major steps for analyzing SV reproducibility among datasets and strategies were listed as below:

- Dataset concordant and specific: Each caller was applied to six datasets for SV detection, and a nonredundant SV set was generated via Jasmine accordingly. SVs reproduced in three HiFi or ONT datasets were indicated by 'SUPP = 3', whereas dataset-specific calls were indicated by 'SUPP = 1'. Then, the percentage of dataset concordant SVs among

total number of SVs in the nonredundant SV callset was calculated.

- Aligner concordant and specific: On each dataset, the reads were aligned with four aligners and SVs were detected subsequently with each caller. For a caller, we merged its four callsets originated from four aligners and obtained the nonredundant SV callset, from which aligner concordant SVs were obtained with 'SUPP = 4' and aligner-specific SVs were labeled by 'SUPP = 1'. Then, the percentage of aligner concordant SVs among total number of SVs in the nonredundant SV callset was calculated.
- Assembler concordant and specific: On HiFi dataset, the reads were assembled by two assemblers (i.e. hifiasm, flye), and the assemblies were mapped with minimap2. For a caller, we merged its two callsets originated from two assemblers and obtained the nonredundant SV callset, from which assembler concordant SVs were obtained with 'SUPP = 2' and assembler-specific SVs were labeled by 'SUPP = 1'. Similar process was applied to ONT dataset, but the assemblies were created by flye and shasta. Then, the percentage of assembler concordant SVs among total number of SVs in the nonredundant SV callset was calculated.
- Callers concordant and specific: On each dataset, we obtained a nonredundant SV set between a read-based caller and an assembly-based caller. From each nonredundant SV set, callers concordant and specific SVs were marked as 'SUPP = 2' and 'SUPP = 1', respectively. Then, the percentage of callers concordant SVs among total number of SVs detected by a caller was calculated.

Obtain and compare high-confident callsets

This study aims to evaluate SV detection at whole genome scale, including complex genomic regions such as tandem repeats and SDs. Thus, the high-confident callset at whole genome scale was created by integrating outputs from different callers, whereas the GIAB high-confident regions were not used because they were limited to simple genomic regions.

To obtain the high-confident callset, insertions and duplications were first written to the insertion VCF file, and deletions were written to a separate VCF file. Note that (1) read-based insertions and deletions were detected from minimap2-aligned reads and (2) assembly-based insertions and deletions were detected from minimap2-aligned contigs created by hifiasm and flye on HiFi and ONT dataset, respectively. Afterwards, we followed the following process to obtain the high-confident insertions and deletions:

- High-confident insertion: Since the insertion VCF file contains two types i.e. insertion and duplication, we first run *'Jasmine file_list=vcf_list.txt out_file=nonredundant_SVs.vcf max_dist=1000 –dup_to_ins genome_file=hs37d5.fa spec_len=50 spec_reads=1'* and obtained the integrated callset of callers. Then, from the integrated callset of five read-based callers, high-confident insertions were those marked by 'SUPP = 5'. The assembly-based insertions were marked by 'SUPP = 2' from the integrated callset of two read-based callers.
- High-confident deletion: We first run *'Jasmine file_list=vcf_list.txt out_file=nonredundant_SVs.vcf max_dist=1000 spec_len=50 spec_reads=1'* and obtained an integrated callset of callers. Then, we followed the same procedure as above to obtain high-confident deletions from the integrated callsets.

Given the read and assembly high-confident callsets, we first obtained the concordant high-confident insertions by running *'Jasmine file_list=vcf_list.txt out_file=nonredundant_SVs.vcf max_dist=1000 –dup_to_ins genome_file=hs37d5.fa spec_len=50 spec_reads=1'*, concordant deletions were obtained via command *'Jasmine file_list=vcf_list.txt out_file=nonredundant_SVs.vcf max_dist=1000 spec_len=50 spec_reads=1'*. Second, in the integrated VCF file, the read and assembly concordant insertions and deletions were those marked as 'SUPP = 2', whereas 'SUPP = 1' indicated the read and assembly-specific calls.

Trio-based validation of read and assembly-specific SVs

The HiFi sequencing data of HG003 and HG004 were obtained from the GIAB. The HG003 and HG004 assemblies were created by hifiasm via default parameters for further usage. The details of the validation steps for specifically detected SVs were listed as below:

- Read specific: SVs were validated via VaPoR with HG003 and HG004 HiFi reads separately. Those SVs not able to be processed by VaPoR were considered as 'Inconclusive SVs'. For SVs that could be evaluated by VaPoR on both maternal and paternal data, if a SV was scored zero (i.e.'VaPoR' = 0) in both parents, it was classified as 'Invalid SVs', whereas others were considered as 'Valid SVs'.
- Assembly specific: SVs were validated via TT-mars with HG003 and HG004 HiFi assemblies separately. Those SVs not able to be processed by TT-mars were considered as 'Inconclusive SVs'. For SVs that could be evaluated by TT-mars on both maternal and paternal data, if a SV was labeled as 'False' in both parents, it was classified as 'Invalid SVs', whereas others were considered as 'Valid SVs'.

Read alignment analysis for assembly-specific SVs

We applied the following steps to examine whether assembly-specific SV loci contain aberrant read alignment i.e. the abnormal inter-read and intra-read alignments used to detect SVs by read-based callers.

- Step1. The assembly-based strategy specifically detected SVs were classified into three types of regions according to the average read mapping quality (**avg_mapq**) obtained from minimap2-aligned reads:
  1) No read mapping region (No_reads)
  2) Low-mapping quality regions (Low_mapq, **avg_mapq** < 20)
  3) High-confident mapping regions (High_mapq, **avg_mapq** ≥ 20).
- The average mapping quality threshold 20 was set according to the default minimum read quality used for SV detection.
- Step2. The potential SV signature reads of those assembly-specific SVs at high-confident mapping quality regions were identified. In general, the 'I' and 'D' tags in the CIGAR string, and the primary reads and their supplementary were collected and used to identify deletion (DEL), insertion (INS), inversion (INV) and duplication (DUP) signatures. The total number of reads containing SV signature was referred to signature count. Moreover, we calculated the start position standard deviation and size standard deviation of all signature reads.

Accordingly, if assembly-specific loci (i) did not contain signature read or (ii) did not contain mapped reads or (iii) located at low mapping quality region, these loci were classified as 'No signature reads' and the others were labeled as 'Has signature reads'. For the 'Has signature reads' loci, we manually inspected their local alignment via IGV and the sequence Dotplot created by Gepard [36] to determine their characteristics.

Evaluating each strategy with well curated SVs

For 35X coverage datasets HiFi-18 kb and ONT-30 kb, we downsample them to 5X, 10X and 20X with SAMtools [37]. Afterwards, each caller is applied to the 5X, 10X and 20X datasets with default parameters except for the number of minimum SV supporting reads, which is set to 1, 2 and 5 for 5X, 10X and 20X datasets, respectively. These values are set to enable effective detection of both homozygous and heterozygous germline SVs. The final VCF files are sorted, compressed and indexed for further evaluation. Furthermore, two benchmarks released by GIAB were used to assess both strategies of detecting SVs at true INS/DEL regions and CMRGs. The recall and precision were measured by Truvari (https://github.com/ACEnglish/truvari) with parameters '*-p 0.00 -r 1000 –passonly –giabreport*', but the genotype accuracy was not considered in our evaluation and all commands were available at https://github.com/jiadong324/ComparStra-Parser.

Key Points

- A comparison and evaluation of SVs detected from read and assembly was performed on six long-read datasets of HG002 genome.
- Up to 80% of the SVs were concordant among different long-read datasets, whereas the breakpoints and type of SVs detected from read were greatly affected by aligners when detecting on a dataset.
- A remarkable subset of SVs at non-tandem repeat regions could be captured by both read and assembly, whereas the discordance was largely caused by complex SVs and inversions due to inconsistent alignment of read and assembly at these loci.
- Benchmarking with SVs at medically relevant genes, detecting from assembly required 20X coverage data to achieve similar performance as read-based detection at 5X coverage data.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

ACKNOWLEDGEMENTS

The authors thank the comments from Bo Wang and other colleagues in the lab as well as those from X.Z., C.R.B. and P.A.A.

FUNDING

This work is supported by the National Science Foundation of China (32125009 and 32070663).

AUTHORS' CONTRIBUTIONS

K.Y. conceptualized and supervised the study. J.L. designed the framework and performed data analysis with help from S.W. and P.J.. J.L., W.K. and K.Y. wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

AVAILABILITY OF CODE AND DATA

All related commands, analysis scripts are available at https://github.com/jiadong324/ComparStra-Parser. All information and links for HiFi and ONT data are listed in Supplementary Table S1. The mutations used for phasing reads are available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/. The structural variant calls used in this study are available at zenodo.7856049.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

REFERENCES

1. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet* 2020;**21**:171–89.
2. Chaisson MJP, Sanders AD, Zhao X, *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019;**10**:1784.
3. Kosugi S, Momozawa Y, Liu X, *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 2019;**20**:117.
4. Wu Z, Jiang Z, Li T, *et al.* Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* 2021;**12**:6501.
5. Beyter D, Ingimundardottir H, Oddsson A, *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* 2021;**53**:779–86.
6. Sone J, Mitsuhashi S, Fujita A, *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* 2019;**51**:1215–21.
7. Hiatt SM, Lawlor JMJ, Handley LH, *et al.* Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGG Adv* 2021;**2**:100023.
8. Pauper M, Kucuk E, Wenger AM, *et al.* Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet* 2021;**29**:637–48.
9. Aganezov S, Goodwin S, Sherman RM, *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res* 2020;**30**:1258–73.
10. Gong L, Wong CH, Cheng WC, *et al.* Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat Methods* 2018;**15**:455–60.
11. Nattestad M, Goodwin S, Ng K, *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* 2018;**28**:1126–35.
12. Zhou B, Ho SS, Greer SU, *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res* 2019;**29**:472–84.
13. Sakamoto Y, Xu L, Seki M, *et al.* Long-read sequencing for non-small-cell lung cancer genomes. *Genome Res* 2020;**30**:1243–57.

14. Zhou B, Ho SS, Greer SU, *et al.* Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res* 2019;**47**:3846–61.
15. Peneau C, Imbeaud S, La Bella T, *et al.* Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. *Gut* 2021;616–26.
16. De Roeck A, De Coster W, Bossaerts L, *et al.* NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol* 2019;**20**:239.
17. Ebert P, Audano PA, Zhu Q, *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021;**372**:170–75.
18. Sedlazeck FJ, Rescheneder P, Smolka M, *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;**15**:461–8.
19. Jiang T, Liu Y, Jiang Y, *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020;**21**:189.
20. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* 2019;**35**:2907–15.
21. Tham CY, Tirado-Magallanes R, Goh Y, *et al.* NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol* 2020;**21**:56.
22. Cretu Stancu M, van Roosmalen MJ, Renkens I, *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun* 2017;**8**:1326.
23. Lin J, Wang S, Audano PA, *et al.* SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* 2022;**19**: 1230–3.
24. Zhao X, Collins RL, Lee WP, *et al.* Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am J Hum Genet* 2021;**108**:919–28.
25. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. *Nat Rev Genet* 2021;**22**: 572–87.
26. Zhao X, Weber AM, Mills RE. A recurrence-based approach for validating structural variation using long-read sequencing technology. *Gigascience* 2017;**6**:1–9.
27. Yang J, Chaisson M. TT-Mars: structural variants assessment based on haplotype-resolved assemblies. *Genome Biology* 2022;**23**:110.
28. Wagner J, Olson ND, Harris L, *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* 2022;**40**:672–80.
29. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 2019;**35**:2193–8.
30. Patterson M, Marschall T, Pisanti N, *et al.* WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* 2015;**22**:498–509.
31. Zhao X, Emery SB, Myers B, *et al.* Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol* 2016;**17**:126.
32. Fujimoto A, Wong JH, Yoshii Y, *et al.* Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. *Genome Med* 2021;**13**:65.
33. Audano PA, Sulovari A, Graves-Lindsay TA, *et al.* Characterizing the major structural variant alleles of the human genome. *Cell* 2019;**176**(663–675):e619.
34. English AC, Menon VK, Gibbs RA, *et al.* Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol* 2022;**23**:271.
35. Kirsche M, Prabhu G, Sherman R, *et al.* Jasmine and iris: population-scale structural variant comparison and analysis. *Nat Methods* 2023;408–17.
36. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 2007;**23**: 1026–8.
37. Li H, Handsaker B, Wysoker A, *et al.* Genome project data processing S: the sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.