frontiers in Pharmacology

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fphar.2021.658072

Comparative Analysis for the Performance of Long-Read-Based Structural Variation Detection Pipelines in Tandem Repeat Regions

*Mingkun Guo[1], Shihai Li[1], Yifan Zhou[1], Menglong Li[1] and Zhining Wen[1,2]**

[1]College of Chemistry, Sichuan University, Chengdu, China, [2]Medical Big Data Center, Sichuan University, Chengdu, China

There has been growing recognition of the vital links between structural variations (SVs) and diverse diseases. Research suggests that, with much longer DNA fragments and abundant contextual information, long-read technologies have advantages in SV detection even in complex repetitive regions. So far, several pipelines for calling SVs from long-read sequencing data have been proposed and used in human genome research. However, the performance of these pipelines is still lack of deep exploration and adequate comparison. In this study, we comprehensively evaluated the performance of three commonly used long-read SV detection pipelines, namely PBSV, Sniffles and PBHoney, especially the performance on detecting the SVs in tandem repeat regions (TRRs). Evaluated by using a robust benchmark for germline SV detection as the gold standard, we thoroughly estimated the precision, recall and F1 score of insertions and deletions detected by the pipelines. Our results revealed that all these pipelines clearly exhibited better performance outside TRRs than that in TRRs. The F1 scores of Sniffles in and outside TRRs were 0.60 and 0.76, respectively. The performance of PBSV was similar to that of Sniffles, and was generally higher than that of PBHoney. In conclusion, our findings can be benefit for choosing the appropriate pipelines in real practice and are good complementary to the application of long-read sequencing technologies in the research of rare diseases.

Keywords: structural variation, tandem repeats, detection pipelines evaluation, rare diseases, long-read sequencing

OPEN ACCESS

Edited by:
Zhichao Liu,
National Center for Toxicological Research (FDA), United States

Reviewed by:
Zhimin Liu,
Janssen Pharmaceuticals, Inc., United States
Bohu Pan,
National Center for Toxicological Research (FDA), United States

****Correspondence:***
Zhining Wen
w_zhining@163.com

Specialty section:
This article was submitted to Pharmacogenetics and Pharmacogenomics, a section of the journal Frontiers in Pharmacology

Received: *25 January 2021*
Accepted: *14 May 2021*
Published: *07 June 2021*

Citation:
Guo M, Li S, Zhou Y, Li M and Wen Z (2021) Comparative Analysis for the Performance of Long-Read-Based Structural Variation Detection Pipelines in Tandem Repeat Regions. Front. Pharmacol. 12:658072. doi: 10.3389/fphar.2021.658072

INTRODUCTION

Previous studies typically defined structural variations as genomic changes at least 50 base pairs (bp) in size. SVs are closely related to diverse human diseases Weischenfeldt et al. (2013); Lupski, (2015), such as autism Pinto et al. (2010); Sanders et al. (2012); Chen et al. (2017) and schizophrenia (Sebat et al., 2007; Stefansson et al., 2008; Walsh et al., 2008; Kirov et al., 2012). Compared with single-nucleotide variations (SNVs), SVs contain more nucleotides and are considered to be higher correlated with evolution, genetic diversity and disease-causing mutations (Stankiewicz and Lupski, 2010; Weischenfeldt et al., 2013; Abel et al., 2020).

Since the size of SV can exceed 1,000 bp, SV detection will be limited by the size of DNA fragments in sequencing. Furthermore, if SVs occur in repetitive regions with high mutation rate, it will be more difficult for detection (Hills et al., 2007; Hastings et al., 2009; Hodgkinson et al., 2012).

In view of the above problems, short-read data may have some difficulties while long-read data can be a good solution (Pollard et al., 2018; Liu et al., 2019). In recent years, long-read technologies

have been developed rapidly Amarasinghe et al. (2020) and used in the discovery of SVs with complex forms (Aneichyk et al., 2018; Song et al., 2018; Ishiura et al., 2019; Zeng et al., 2019; Logsdon et al., 2020). The size of DNA fragment sequenced by long-read technologies is usually larger than 1,000 bp, which can cover the range of large SV and contain much context information (Chaisson et al., 2015). It ensures the advantages of long-read technologies in SV detection, especially in the complex repetitive regions of the genome. Characterized by high incidence rate of SVs and high complexity, repetitive regions are an important and challenging problem in SV detection (Sudmant et al., 2015; Zook et al., 2020). However, the performance of SV detection pipelines based on long-read data applied in repetitive regions still need to be analyzed.

Therefore, in this study, we selected three commonly used long-read-based pipelines Kosugi et al. (2019); Logsdon et al. (2020), namely PBSV Wenger et al. (2019), Sniffles Sedlazeck et al. (2018) and PBHoney English et al. (2014), and comprehensively evaluated their performance on SV detection. Using the benchmark established by the Genome in a Bottle (GIAB) Consortium Zook et al. (2020) as the gold standard, we evaluated the precision, recall and F1 score of these pipelines. The comparison included the comparison between insertions and deletions, the comparison among four size ranges of SVs and the comparison between SVs in TRRs and SVs outside TRRs. The F1 scores of Sniffles were 0.60 in TRRs and 0.76 outside TRRs. Similarly, The F1 scores of PBSV were 0.59 and 0.74 in and outside TRRs, respectively. The performances of the two pipelines were generally higher than that of PBHoney. For the three pipelines, the performances in TRRs were lower than those outside TRRs, which indicated that SV detection in TRRs was more difficult than that outside TRRs. Concerning the type of SVs, it was found that large insertions (> 1,000 bp) were the most difficult to detect while large deletions were easy to precisely detect, especially in TRRs. In addition, we also analyzed the potential performance of three pipelines on detecting *de novo* SVs. The results suggested that long-read technologies and the SV detection pipelines still need further development for the precise detection of *de novo* SVs.

MATERIALS AND METHODS

Datasets

The long-read sequencing data of an Ashkenazim Jewish trio Zook et al. (2016) were used in our study. Subreads datasets of the son (HG002), the father (HG003) and the mother (HG004) were downloaded from GIAB (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/). The average coverages of the trio are approximately 69X, 32X and 30X, and their N50 subread lengths are 11,087, 10,728, and 10,629 bp.

Benchmark

The benchmark is established by GIAB for HG002 on GRCh37, which was downloaded from GIAB FTP site (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/). The benchmark dataset contains close to 100% of true insertions and deletions in the specific regions. According to the guidance of the benchmark, we used the SVs with the FILTER field "PASS" in the Tier 1 vcf, including 12,745 isolated, sequence-resolved insertion (7,281) and deletion (5,464) calls. The benchmark regions include 34,830 large regions, of which 15% are within 1,000–10,000 bp and 82% are over 10,000 bp. Through the manual inspection in the benchmark work, it was found that approximately 5% of true insertions in the benchmark regions might be missing. Therefore, when comparing callsets (especially from long-read data) with the benchmark, it is possible to misjudge some true insertions. When making the comparison, we first selected the SVs in the benchmark regions, and then compared these SVs with the benchmark SVs.

Structural Variation Detection Pipelines

In this study, we used three long-read-based pipelines named PBSV (version 2.2.2; https://github.com/PacificBiosciences/pbsv), Sniffles (version 1.0.11; https://github.com/fritzsedlazeck/Sniffles) and PBHoney (in PBSuite-15.8.24; http://sourceforge.net/projects/pb-jelly/). For PBSV and Sniffles, subreads were aligned to reference genomes GRCh37 by PBMM2 and NGMLR, respectively. After the help of SAMtools, SVs were called by PBSV and Sniffles. PBHoney includes two parts of results, namely Tails (based on interrupted mapping) and Spots (based on intra-read discordance). There were too few results in the Tails part to compare with other pipelines, thus the result of the Tails part was separately shown in **Supplementary Figure S1**. Because of the complexity of parameter optimization in Spots and the time-consume of recommended aligner BLASR, following a previous work Kosugi et al. (2019), we used NGMLR to align the subreads and detected SVs with custom-made parameters for insertions and deletions. SVs with < 0.2 of the value, which was calculated by dividing the szCount tag with the coverage tag, were filtered out.

In these callsets, we only summarized the variations ≥ 50 bp. SVs with the type "BND" (breakpoint end) were excluded. In this study, we only focused on the SVs on the autosomes and sex chromosomes.

The Metrics for Comparison

During comparison, we mainly considered the type consistency, the distance between breakpoints and the proportion of the reciprocal overlap. For compared insertions, if the distance of breakpoints was within 200 bp, they were considered the same. For compared deletions, the called SV needed to exhibit ≥ 50% reciprocal overlap with the reference SV. When comparing the callset with regions (i.e. the benchmark regions and tandem repeat regions), it was only required that breakpoints overlapped with these regions. When comparing the overlap among pipelines, the callset with more SVs was chosen as the comparison benchmark. The code used for comparison are available at GitHub (https://github.com/cic-gmk/DNSV).

When comparing the callsets with the benchmark, the precision, recall and F1 score were calculated *via* the following equations:

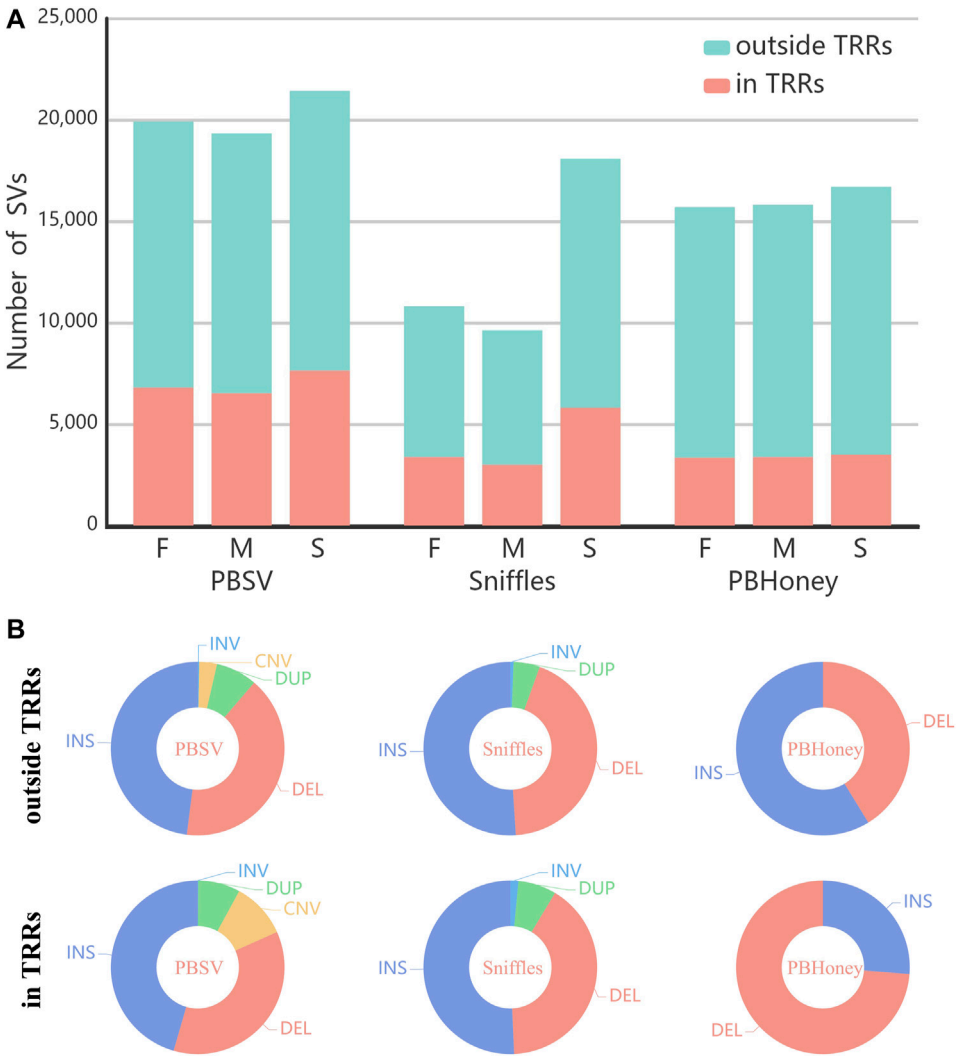

FIGURE 1 | (A) The number of SVs in the callsets of the parents and the son (F, Father; M, Mother; S, Son) detected by PBSV, Sniffles and PBHoney. **(B)** The type distribution of the whole callsets of the trio detected by three pipelines in/outside TRRs.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, FP and FN are the numbers of true positives, false positives and false negatives. TP + FP is equal to the number of the called SVs. TP + FN is equal to the number of the benchmark SVs.

Tandem Repeat Regions

The repeats used in our study were annotated in the annotation file of hg19, which can be obtained at the download site of UCSC Genome Browser (Fernandes et al., 2020; Navarro Gonzalez et al., 2020) (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz). "Simple repeats" and "Satellites" were selected as the TRRs from the file. "Simple repeats" are short pattern tandem repeats and "Satellites" are medium to long pattern tandem repeats. SVs were divided into two parts according to whether they were in TRRs or not.

RESULTS

The Landscape of Structural Variation Callsets

The numbers of SVs detected by the three pipelines are shown in **Figure 1A**. Among the three pipelines, PBSV detected the largest number of SVs and Sniffles detected the least number of SVs. For all the pipelines, the numbers of detected SVs of the son were more than those of the parents mainly due to the higher coverage of the sequencing data of the son. For Sniffles, largest difference in

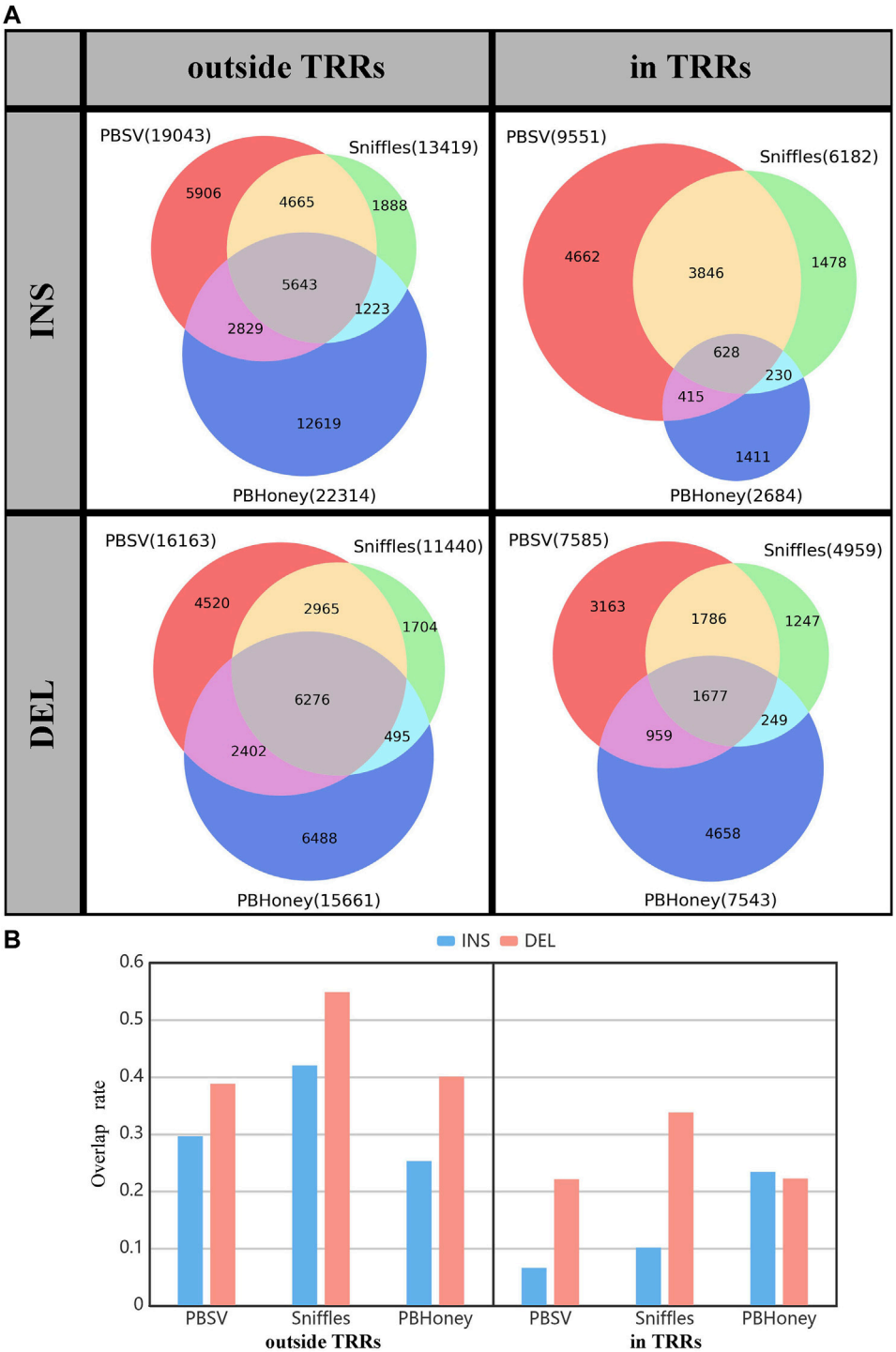

FIGURE 2 | (A) The overlap among callsets detected by three pipelines divided by the type and the location relation with TRRs of SVs. **(B)** The overlap rate of three pipelines. The gray part in **Figure 2A** shows the overlap SVs.

the numbers of detected SVs existed between the son and the parents. Because of the high mutation rate of TRRs Hills et al. (2007); Hastings et al. (2009), although the abundance of TRRs accounts for only about 10% of the human genome Benson, (1999), a number of SVs were still detected in TRRs in the callsets of the trio by three pipelines (PBSV 35%, Sniffles 32%, PBHoney 21%). **Figure 1B** shows the type distribution of all SVs detected by each of the three pipelines. Although insertions are more difficult

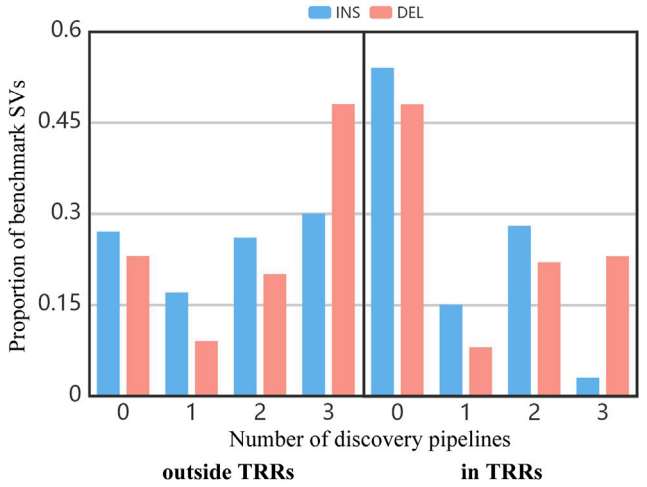

FIGURE 3 | The proportion of benchmark SVs concurrently detected by different number of pipelines.

to detect Zook et al. (2020), the proportions of insertions detected by three pipelines were higher than those of deletions, except for PBHoney in TRRs. The size distribution of SVs detected by each of the three pipelines is provided in **Supplementary Figure S2**. It was found that the number of detected SVs decreased fast as the size of SVs increased. Insertions were generally in the majority when the size < 1,000 bp, but the proportion of deletions increased with the increase of size. In addition, we also investigated the distribution of the percentage of SVs across chromosomes for pipelines (**Supplementary Figure S3**).

We summed the overlap among the callsets of each person detected by three pipelines for comparison (**Figure 2**). For the SVs outside TRRs, the overlap proportion of SVs detected by Sniffles was the highest, and close to 42% (5,643/13,419) of insertions and 55% (6,267/11,440) of deletions can be detected by the other two pipelines. For SVs in TRRs, when comparing Sniffles with PBSV, about 72% ((3,846 + 628)/6,182) of insertions and 70% ((1,786 + 1,677)/4,959) of deletions identified by Sniffles can be detected by PBSV. However, only 628 insertions detected by PBHoney in TRRs were involved in the callsets of PBSV and Sniffles due to the insufficient ability of PBHoney for detecting insertion in TRRs. It can be seen from **Figure 2B** that, except for PBHoney in TRRs, the overlap rates of insertions were lower than those of deletions. For the three pipelines, the overlap rates in TRRs were lower than those outside TRRs, suggesting that the difference among the callsets from different pipelines in TRRs was large.

Evaluation on the Performance of Pipelines

The benchmark used in our study defines the comparing regions, in which the benchmark contains close to 100% of true insertions and deletions. Therefore, we compared the callsets detected by three pipelines with the benchmark callset in the comparing regions.

Figure 3 shows the proportion of benchmark SVs concurrently detected by different number of pipelines. In the whole benchmark callset, close to 25% of benchmark SVs outside TRRs and 51% of benchmark SVs in TRRs cannot be discovered by any pipeline (the bar of "0"). It suggested that SVs in TRRs were more difficult to detect. Similarly, the proportion of benchmark SVs concurrently detected by three pipelines in TRRs was obviously lower than that outside TRRs (the bar of "3"), which agreed with the overlap results among three pipelines (**Figure 2**).

Using the benchmark as the golden standard, the precision, recall and F1 score of three pipelines are shown in **Figure 4**. The precisions achieved by PBSV and Sniffles were higher than 80% both in and outside TRRs, indicating that the SVs detected by these two pipelines were relatively precise. The precision of PBHoney was the lowest, suggesting that more false positives existed in the callset of PBHoney. For all the pipelines, the recalls were under 80 and 50% outside and in TRRs, respectively. It suggested that there were still a number of SVs omitted by the three pipelines. The recall of insertions detected by PBHoney in TRRs was especially low (8%), suggesting that its detection ability of insertions in TRRs was suboptimal. For all the three pipelines, the F1 scores in TRRs were obviously lower than those outside TRRs, indicating the detection of SVs in TRRs was more challenging. In addition, because the son's SVs are inherited from the parents, we also made comparison between the callsets of the parents and the benchmark (**Supplementary Figure S4**). The nominal precisions of the callsets of the parents were clearly lower than those of the son mainly because the benchmark was constructed only based on the sequencing data of the son. It suggested that the benchmark construction in future need to consider the diversity of the population.

Figure 5 shows the impact of the size of SVs on the detection ability of three pipeline. The F1 scores of PBSV and Sniffles were relatively stable with the increase of the size of SVs. However, the size of SVs induced a clear impact on PBHoney. Especially when the size of SVs was more than 1,000 bp, PBHoney can hardly detect true insertions. For all the three pipelines, the F1 scores of large insertion detection (>1,000 bp) were obviously lower than those of large deletion detection, suggesting that the detection for large insertions were more challenging.

Potential Performance of Three Pipelines on Detecting *de novo* Structural Variations

The mutations that only occurred in the child rather than the parents are generally called *de novo* mutations (Conrad et al., 2011; Veltman and Brunner, 2012). We calculated the rate of "*de novo*" SVs by dividing the number of the SVs only detected from the son by the number of all SVs detected from the son. As shown in **Figure 6**, the "*de novo*" rate of Sniffles was 32%, which was higher than that of PBSV (18%) and PBHoney (14%). These rates were much higher than the actual *de novo* rate (Veltman and Brunner, 2012). It indicated that a large number of false positive *de novo* SVs existed in the callsets, which may also be attributed to Mendelian inheritance errors Pilipenko et al. (2014); Kothiyal et al. (2019), the false positive SVs of the son and the false negative SVs of the parents. Our results suggested that long-read

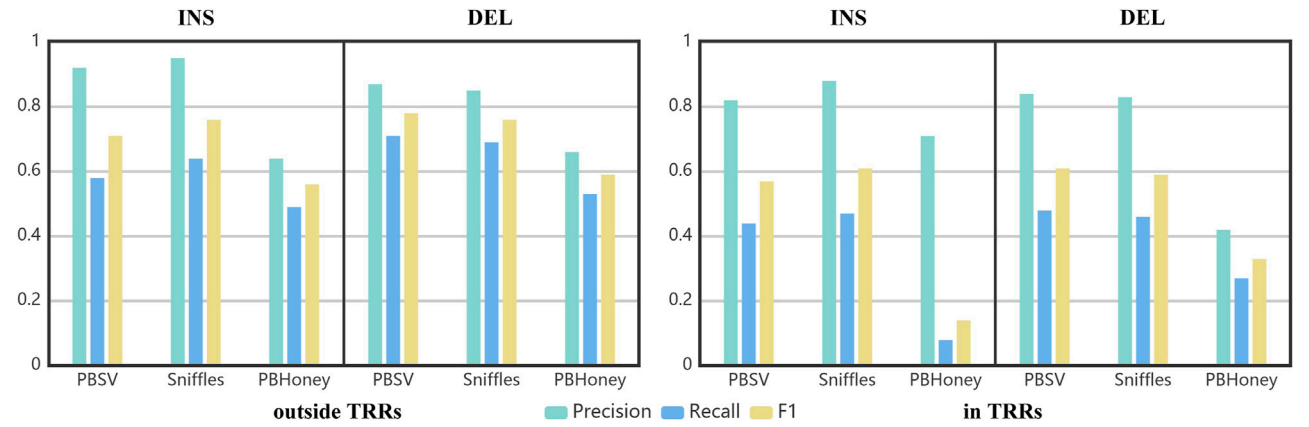

FIGURE 4 | The precision, recall and F1 of three pipelines in detecting insertions and deletions in/outside TRRs.

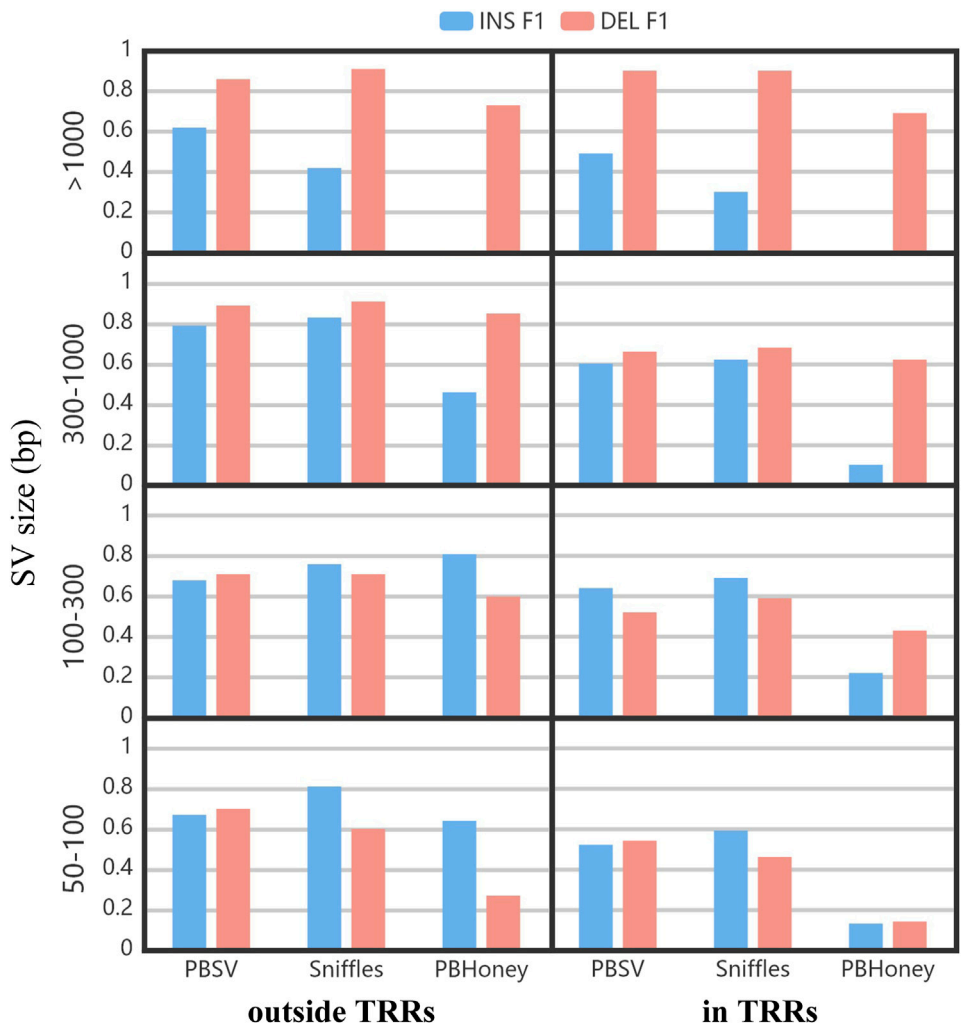

FIGURE 5 | The F1 of insertions and deletions divided by the size and the location relation with TRRs of SVs. The size ranges included 50–100 bp, 100–300 bp, 300–1,000 bp and > 1,000 bp.

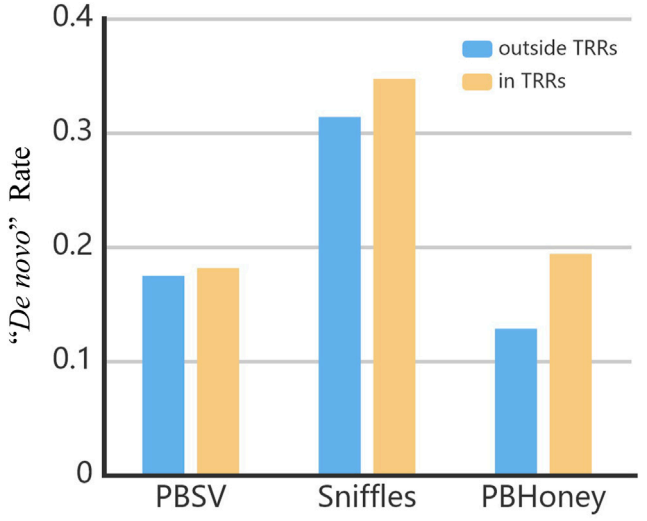

FIGURE 6 | The "*de novo*" rate in the callsets of three pipelines in and outside TRRs.

technologies and the detection pipelines still need further improvement for detecting *de novo* SVs when applying to the exploration of mechanisms of rare diseases. It is a considerable challenge to reduce the false positives and false negatives in SV detection in future study.

DISCUSSION

The relatively large size of SVs and the complex repetitive context make SV detection challenging. Because long-read sequencing data contain abundant context information, it can perform well in SV detection. Therefore, we comprehensively analyzed the performance of three commonly used long-read SV detection pipelines. Our results showed that the overlap proportion among the callsets of the pipelines in TRRs was generally lower than that outside TRRs. Comparing callsets with the benchmark, the precisions, recalls and F1 scores of these pipelines in TRRs were obviously lower than those outside TRRs. These results suggested that the detection of SVs in TRRs was more difficult than that outside TRRs.

As shown in **Figure 4**, the F1 scores of PBSV and Sniffles were similar, and higher than that of PBHoney. With the default recommended parameter, preferable results can be obtained by PBSV and Sniffles. As shown in **Figure 5**, the F1 scores of PBSV and Sniffles did not change a lot with the increase of the size of SVs except for the detection of insertions larger than 1,000 bp. But the detection of both insertions and deletions with PBHoney was clearly influenced by the size of SVs. In fact, as shown in **Supplementary Figure S2**, it was difficult for PBHoney to detect SVs above 4,000 bp. For PBHoney, it was necessary to make proper settings and filter process for SVs with different types. In addition, there were 3% of the son's callset of PBSV and 45% of the son's callset of Sniffles marked with the label "IMPRECISE", which indicated the probably insufficient precision of SVs. These SVs were mainly composed of insertions (95% for PBSV and 71% for Sniffles). Interestingly, the precision of SVs tagged "IMPRECISE" in PBSV was really low (15%), but for Sniffles, the precision was still high (81%), which meant there was no need to filter these SVs in Sniffles.

In our study, we selected SVs (≥ 50 bp) from the benchmark for comparison. If more variations was selected, such as using the cutoff of variations ≥ 30 bp, it would lead to higher precision and lower recall (Kosugi et al., 2019). Our results showed that, even with high precision, no pipeline can achieve very high recall in SV detection. Therefore, it is necessary to integrate different pipelines for generating a comprehensive callset. Integrating Sniffles and PBHoney, NextSV Fang et al. (2018) had been developed to detect SVs from low-coverage long-read sequencing data and achieved better performance than a single pipeline. In addition, a pipeline with multiple algorithms can be developed to optimize and simplify the procession of SV detection.

Previous studies have found that long-read sequencing can identify pathogenic SVs of rare genetic diseases which cannot be identified by short-read sequencing, such as the pathogenic SVs of Carney complex Merker et al. (2018) and progressive myoclonic epilepsy (Mizuguchi et al., 2019). In the study of Carney complex, the pathogenic SV was identified by pipeline detection followed by manual screening and analysis. Since the SV was not detected from the parents, the pathogenic SV was also proved to be a *de novo* SV. However, it was difficult to identify pathogenic SVs from "*de novo*" SVs using long-read sequencing. It is known that the number of *de novo* mutations in heredity is very small (Conrad et al., 2011; Veltman and Brunner, 2012). But compared with the number of *de novo* single-nucleotide variations detected from short-read data Liang et al. (2019), the number of "*de novo*" SVs detected from long-read data was too large. And the precision was too high when comparing the "*de novo*" SVs with the benchmark (**Supplementary Figure S5**). Therefore, these SVs cannot be simply regarded as true *de novo* SVs. In order to analyze true *de novo* SVs, more true SVs need to be detected from parents. As shown in **Figure 1A**, in each pipeline, the numbers of SVs of the parents were obviously smaller than that of the son due to the lower coverages of the parents. It suggested that the sequencing coverages of the trio need to be ensured. Also, for high precision and recall of the detection of *de novo* SVs, the precision and recall of SV detection pipelines still need to be improved.

CONCLUSION

In this study, we thoroughly compared three commonly used SV detection pipelines and found that the precisions of PBSV and Sniffles were generally similar, and higher than PBHoney. The recalls of the three pipelines were still suboptimal. The performances of PBSV and Sniffles were relatively stable with the increase of the size of SVs, while the performance of PBHoney varied largely. The performances of the three pipelines in TRRs were obviously lower than those outside TRRs, indicating that SV detection in TRRs was more difficult. Comparing insertions with deletions, the detection of large insertions was obviously more

difficult than that of large deletions. Our findings can be helpful for conducting the SV detection in the mechanism exploration of rare diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZW designed the experiments. MG did the entire experiment. MG, SL, and YZ analyzed the data. ZW and MG wrote the main manuscript text and prepared all the figures. All authors discussed the results and revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This project was supported by the grant from the National Natural Science Foundation of China (No. 21575094).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2021.658072/full#supplementary-material

REFERENCES

Abel, H. J., Larson, D. E., Larson, D. E., Regier, A. A., Chiang, C., Das, I., et al. (2020). Mapping and Characterization of Structural Variation in 17,795 Human Genomes. *Nature* 583, 83–89. doi:10.1038/s41586-020-2371-0

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biol.* 21 (1), 30. doi:10.1186/s13059-020-1935-5

Aneichyk, T., Hendriks, W. T., Yadav, R., Shin, D., Gao, D., Vaine, C. A., et al. (2018). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* 172, 897–909. doi:10.1016/j.cell.2018.02.011

Benson, G. (1999). Tandem Repeats Finder: a Program to Analyze DNA Sequences. *Nucleic Acids Res.* 27, 573–580. doi:10.1093/nar/27.2.573

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing. *Nature* 517, 608–611. doi:10.1038/nature13907

Chen, C. H., Chen, H. I., Chien, W. H., Li, L. H., Wu, Y. Y., Chiu, Y. N., et al. (2017). High Resolution Analysis of Rare Copy Number Variants in Patients with Autism Spectrum Disorder from Taiwan. *Scientific Rep.* 7 (1), 11919. doi:10.1038/s41598-017-12081-4

Conrad, D. F., Keebler, J. E., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., et al. (2011). Variation in Genome-wide Mutation Rates within and between Human Families. *Nat. Genet.* 43, 712–714. doi:10.1038/ng.862

English, A. C., Salerno, W. J., and Reid, J. G. (2014). PBHoney: Identifying Genomic Variants via Long-Read Discordance and Interrupted Mapping. *Bmc Bioinf.* 15, 180. doi:10.1186/1471-2105-15-180

Fang, L., Hu, J., Wang, D., and Wang, K. (2018). NextSV: a Meta-Caller for Structural Variants from Low-Coverage Long-Read Sequencing Data. *Bmc Bioinf.* 19 (1), 180. doi:10.1186/s12859-018-2207-1

Fernandes, J. D., Zamudio-Hurtado, A., Clawson, H., Kent, W. J., Haussler, D., Salama, S. R., et al. (2020). The UCSC Repeat Browser Allows Discovery and Visualization of Evolutionary Conflict across Repeat Families. *Mobile DNA* 11, 13. doi:10.1186/s13100-020-00208-w

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of Change in Gene Copy Number. *Nat. Rev. Genet.* 10, 551–564. doi:10.1038/nrg2593

Hills, M., Jeyapalan, J. N., Foxon, J. L., and Royle, N. J. (2007). Mutation Mechanisms that Underlie Turnover of a Human Telomere-Adjacent Segmental Duplication Containing an Unstable Minisatellite. *Genomics* 89, 480–489. doi:10.1016/j.ygeno.2006.12.011

Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The Large-Scale Distribution of Somatic Mutations in Cancer Genomes. *Hum. Mutat.* 33, 136–143. doi:10.1002/humu.21616

Ishiura, H., Shibata, S., Yoshimura, J., Suzuki, Y., Qu, W., Doi, K., et al. (2019). Noncoding CGG Repeat Expansions in Neuronal Intranuclear Inclusion Disease, Oculopharyngodistal Myopathy and an Overlapping Disease. *Nat. Genet.* 51, 1222, 1232. doi:10.1038/s41588-019-0458-z

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., et al. (2012). De Novo CNV Analysis Implicates Specific Abnormalities of Postsynaptic Signalling Complexes in the Pathogenesis of Schizophrenia. *Mol. Psychiatry* 17, 142–153. doi:10.1038/mp.2011.154

Kosugi, S., Momozawa, Y., Liu, X. X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing. *Genome Biol.* 20 (1), 117. doi:10.1186/s13059-019-1720-5

Kothiyal, P., Wong, W. S. W., Bodian, D. L., and Niederhuber, J. E. (2019). Mendelian Inconsistent Signatures from 1314 Ancestrally Diverse Family Trios Distinguish Biological Variation from Sequencing Error. *J. Comput. Biol.* 26, 405–419. doi:10.1089/cmb.2018.0253

Liang, Y., He, L., Zhao, Y. R., Hao, Y. Y., Zhou, Y. F., Li, M. L., et al. (2019). Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the De Novo Mutations in Humans. *Front. Pharmacol.* 10, 358. doi:10.3389/fphar.2019.00358

Liu, Z., Zhu, L., Roberts, R., and Tong, W. (2019). Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? *Trends Genet.* 35, 852–867. doi:10.1016/j.tig.2019.08.006

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read Human Genome Sequencing and its Applications. *Nat. Rev. Genet.* 21, 597–614. doi:10.1038/s41576-020-0236-x

Lupski, J. R. (2015). Structural Variation Mutagenesis of the Human Genome: Impact on Disease and Evolution. *Environ. Mol. Mutagen.* 56, 419–436. doi:10.1002/em.21943

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., et al. (2018). Long-read Genome Sequencing Identifies Causal Structural Variation in a Mendelian Disease. *Genet. Med.* 20, 159–163. doi:10.1038/gim.2017.86

Mizuguchi, T., Suzuki, T., Abe, C., Umemura, A., Tokunaga, K., Kawai, Y., et al. (2019). A 12-kb Structural Variation in Progressive Myoclonic Epilepsy Was Newly Identified by Long-Read Whole-Genome Sequencing. *J. Hum. Genet.* 64, 359–368. doi:10.1038/s10038-019-0569-5

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., and Raney, B. J. (2020). The UCSC Genome Browser Database: 2021 Update. *Nucleic Acids Res.* 49 (D1), D1046–D1057. doi:10.1093/nar/gkaa1070

Pilipenko, V. V., He, H., Kurowski, B. G., Alexander, E. S., Zhang, X., Ding, L., et al. (2014). Using Mendelian Inheritance Errors as Quality Control Criteria in Whole Genome Sequencing Data Set. *BMC Proc.* 8, S21. doi:10.1186/1753-6561-8-s1-s21

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., et al. (2010). Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorders. *Nature* 466, 368–372. doi:10.1038/nature09146

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. (2018). Long Reads: Their Purpose and Place. *Hum. Mol. Genet.* 27, R234–R241. doi:10.1093/hmg/ddy177

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De Novo mutations Revealed by Whole-Exome Sequencing Are Strongly Associated with Autism. *Nature* 485, 237–241. doi:10.1038/nature10945

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong Association of De Novo Copy Number Mutations with Autism. *Science* 316, 445–449. doi:10.1126/science.1138659

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., et al. (2018). Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing. *Nat. Methods* 15, 461–468. doi:10.1038/s41592-018-0001-7

Song, J. H. T., Lowe, C. B., and Kingsley, D. M. (2018). Characterization of a Human-specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am. J. Hum. Genet.* 103, 421–430. doi:10.1016/j.ajhg.2018.07.011

Stankiewicz, P., and Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annu. Rev. Med.* 61, 437–455. doi:10.1146/annurev-med-100708-204735

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., et al. (2008). Large Recurrent Microdeletions Associated with Schizophrenia. *Nature* 455, 232–236. doi:10.1038/nature07229

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* 526**,** 75, 81. doi:10.1038/nature15394

Veltman, J. A., and Brunner, H. G. (2012). De Novo mutations in Human Genetic Disease. *Nat. Rev. Genet.* 13, 565–575. doi:10.1038/nrg3241

Walsh, T., Mcclellan, J. M., Mccarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., et al. (2008). Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* 320, 539–543. doi:10.1126/science.1155174

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease. *Nat. Rev. Genet.* 14, 125–138. doi:10.1038/nrg3373

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome. *Nat. Biotechnol.* 37, 1155–1162. doi:10.1038/s41587-019-0217-9

Zeng, S., Zhang, M.-Y., Wang, X.-J., Hu, Z.-M., Li, J.-C., Li, N., et al. (2019). Long-read Sequencing Identified Intronic Repeat Expansions inSAMD12from Chinese Pedigrees Affected with Familial Cortical Myoclonic Tremor with Epilepsy. *J. Med. Genet.* 56, 265–270. doi:10.1136/jmedgenet-2018-105484

Zook, J. M., Catoe, D., Mcdaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive Sequencing of Seven Human Genomes to Characterize Benchmark Reference Materials. *Scientific Data* 3, 160025. doi:10.1038/sdata.2016.25

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., et al. (2020). A Robust Benchmark for Detection of Germline Large Deletions and Insertions. *Nat. Biotechnol.* 38 (11), 1347–1355. doi:10.1038/s41587-020-0538-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.