bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Evaluation on Detection of Structural Variants by Low-Coverage Long-Read Sequencing

**Li Fang[1], Jiang Hu[1], Depeng Wang[1], Kai Wang[2,3,*]**

1: Grandomics Biosciences, Beijing 102206, China

2: Institute for Genomic Medicine, Columbia University Medical Center, New York, NY 10032, USA

3: Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA

*: Correspondence should be addressed to K.W. (kai@openbioinformatics.org)

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


## Abstract

Structural variants (SVs) in human genome are implicated in a variety of human diseases. Long-read sequencing (such as those from PacBio) delivers much longer read lengths than short-read sequencing (such as those from Illumina) and may greatly improve SV detection. However, due to the relatively high cost of long-read sequencing, users are often faced with issues such as what coverage is needed and how to optimally use the aligners and SV callers. Here, we evaluated SV calling performance of three SV calling algorithms (PBHoney-Tails, PBHoney-Spots and Sniffles) under different PacBio coverages on two personal genomes, NA12878 and HX1. Our results showed that, at 10X coverage, 76% ~ 84% deletions and 80% ~ 92 % insertions in the gold standard set can be detected by PBHoney-Spots. Combining both PBHoney-Spots and Sniffles greatly increased sensitivity, especially under lower coverages such as 6X. We further evaluated the Mendelian errors on an Ashkenazi Jewish trio dataset with low-coverage whole-genome PacBio sequencing. In addition, to automate SV calling, we developed a computational pipeline called NextSV, which integrates PBhoney and Sniffles and generates the union (high sensitivity) or intersection (high specificity) call sets. Our results provide useful guidelines for SV identification from low coverage whole-genome PacBio data and we expect that NextSV will facilitate the analysis of SVs on long-read sequencing data.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Introduction

Structural variants (SVs), including large variations such as deletions, insertions, duplications, inversions, and translocations, make an important contribution to human diversity and disease susceptibility[1,2]. Many inherited diseases and cancers have been associated with a large number of SVs in recent years[3-8]. Recent advances in next-generation sequencing (NGS) technologies have facilitated the analysis of variations such as SNPs and small Indels in unprecedented details, but the discovery of SVs using short reads still remains challenging [9]. Single-molecule, real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio) offers a long read length, making it potentially well-suited for SV detection in personal genomes[9,10]. Most recently, Merker et al. reported the application of low coverage whole genome PacBio sequencing to identify pathogenic structural variants from a patient with autosomal dominant Carney complex, for whom targeted clinical gene testing and whole genome short-read sequencing were negative [11].

Two SV software tools have been developed specifically for long-read sequencing: PBhoney [12] and Sniffles (https://github.com/fritzsedlazeck/Sniffles). PBhoney identifies genomic variants via two algorithms, long-read discordance (PBhoney-Spots) and interrupted mapping (PBhoney-Tails). Sniffles is a SV caller written in C++ and it detects SVs using evidence from split-read alignments, high-mismatch regions, and coverage analysis. Due to the relative high cost of PacBio sequencing, users are often faced with issues such as what coverage is needed and how to get the best use of the available SV callers. In addition, it is unclear which software performs the best in low-coverage settings, and whether the combination of software tools can improve performance of SV calls. Finally, the execution of these software tools is often not straightforward and requires careful re-parameterization given specific coverage of the source data.

Recently, the Genome in a Bottle (GIAB) consortium hosted by National Institute of Standards and Technology (NIST) distributed a set of high-confidence SV calls for the NA12878 genome, an extensively sequenced genome by different platforms, enabling

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

benchmarking of SV callers[13]. They also published sequencing data of seven human genomes, including PacBio data of an Ashkenazi Jewish family trio[14]. Previously, we sequenced a Chinese individual HX1 on the PacBio platform, and generated assembly-based SV call sets[15]. Using data sets of NA12878, HX1 and the AJ trio, we compared the performance of PBhoney-Spots, PBhoney-Tails, Sniffles and their combination under different PacBio coverages. In addition, we provided NextSV, an automated SV calling pipeline using PBHoney-Spots, PBHoney-Tails and Sniffles. NextSV automatically execute these three other software tools with optimized parameters for the specific coverage that user specified, then integrates results of each caller and generates the union (high sensitivity) or intersection (high specificity) call sets. We expect that NextSV will facilitate the detection and analysis of SVs on long-read sequencing data.

## Materials and Methods

### *PacBio data sets used for this study*

Five whole-genome PacBio sequencing data sets were used to test the performance of SV calling pipelines (Table 1). Data sets of NA12878 and HX1 genome were obtained from NCBI SRA database. Data sets of the Ashkenazi Jewish (AJ) family trio were downloaded from ftp site of NIST (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/). After we obtained raw data, we extracted subreads using the SMRT Portal software (Pacific Biosciences, Menlo Park, CA) with default settings. The subreads were mapped to the reference genome using BLASR [16] or BWA-MEM [17]. The bam files were down-sampled to different coverages using SAMtools (samtools view -s). The down-sampled coverages and mean read lengths of the data sets are shown in Table 1.

### *SV detection using PBHoney*

PacBio subreads were iteratively aligned with the human reference genome (GRCh38 for HX1, GRCh37 for NA12878 and AJ trio genomes, depending on the reference of gold standard set) using the BLASR aligner (parameter: -bestn 1). Each read's single best alignment was stored in the SAM output. Unmapped portions of each read were extracted from the alignments and remapped to the reference genome. The alignments

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in SAM format were converted to BAM format and sorted by SAMtools. PBHoney-Tails and PBHoney-Spots were run with slightly modified parameters (minimal read support 2, instead of 3 and consensus polishing disabled) to increase sensitivity and discover SVs under low coverages (2~15X).

***SV detection using Sniffles***

PacBio subreads were aligned to the reference genome, using BWA-MEM with parameters modified for PacBio reads (bwa mem -M -x pacbio), to generate the BAM file. The BAM file was used as input of Sniffles. Sniffles was run with slightly modified parameters (minimal read support 2, instead of 10) to increase sensitivity and discover SVs under low fold of coverages (2~15X).

***Comparing two SV call sets***

Calls which reciprocally overlapped by more than 50% (bedtools intersect -f 0.5 -r) were considered to be the same SV and merged into a single call. For insertion calls, a padding of 500 bp was added before intersection. When merging two SVs, the average start and end positions were used.

***Gold standard SV call set***

The gold standard SV call set for NA12878 was retrieved from the GIAB consortium [13], in which most of the calls were refined by experimental validation or other independent technologies. For the HX1 genome, we used the SV calls from a previously validated local assembly approach[10], as the initial high-quality calls. We also detected SVs on 100X coverage PacBio data set of the HX1 genome using PBHoney-Tails, PBHoney-Spots and Sniffles. The initial high-quality calls that overlapped with one of the three 100X call sets (PBHoney-Tails, PBHoney-Spots or Sniffles) were retained as final gold standard calls. SVs with length less than 200 bp were not considered. Number of SVs in the gold standard sets is shown in Table 2.

***Performance Evaluation of SV callers***

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The SV calls of each caller were compared with the gold standard SV set. Precision, recall, and F1 score were used to evaluate the performance of the callers. Precision, recall, and F1 were calculated as

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

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

$$\text{F1} = 2 \cdot \frac{precision \cdot recall}{precision+recall},$$

where TP is the number of true positives (variants called by a variant caller and matching the gold standard set), FP is the number of false positives (variants called by a variant caller but not in the gold standard set), and FN is the number of false negatives (variants in the gold standard set but not called by a variant caller).

# Results

## *Performance of SV calling under different PacBio coverage*

To determine what sequencing coverage is needed for SV detection using PacBio data, we evaluated the performance of SV callers under several different coverages. We downloaded a recently published 22X PacBio data set of NA12878 [18] and down-sampled the data set to 2X, 4X, 6X, 8X, 10X, 12X, and 15X. SV calling was performed using PBHoney and Sniffles under each coverage. The resulting calls were compared with the gold standard SV set (including 2094 deletion calls and 68 insertion calls) from the Genome In A Bottle (GIAB) consortium[18].

First, we examined how many calls in the gold set can be discovered. As shown in Figure 1A and 1B, the recall increased rapidly before 6X coverage but the slope of increase slowed down after 10X. Among the three callers, PBHoney-Spots discovered more SV calls than Sniffles and PBHoney-Tails. At 10X coverage, PBHoney-Spots detected 76% of deletions and 80% insertions in the gold standard set; Sniffles discovered 63% deletions and 25% insertions in the gold standard set; PBHoney-Tails recalled 26% deletions and 3% insertions. At 15X coverage, the recall of PBHoney-Spots was 80% for deletion calls and 87% for insertion calls, which is only 6% ~ 9% higher than the recall at 10X.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Second, we examined the precision and F1 scores of callers under different coverage. We calculated precision as the fraction of detected SVs that matching the gold standard set. As shown in Figure 1C, Sniffles has higher precision than PBHoney-Spots and PBHoney-Tails. The precision of Sniffles for deletion calls was 70% at 6X coverage, and decreased slightly as the coverage increased. F1 score, the harmonic mean of precision and recall, increased before 10X and then kept stable at higher coverage (Figure 1D). Precision for insertion calls was not assessed because there were only 86 insertion calls in the GIAB gold standard set, which was one order of magnitude smaller than the number of deletion calls, with potentially high false negative rates.

To verify the performance of SV detection on different individuals, we also did evaluation on a Chinese genome HX1, which was sequenced by us recently[15] at 103X PacBio coverage. The genome was sequenced using a newer version of chemical reagents and thus the mean read length of HX1 was 40% longer than that of NA12878 (Table 1). The total data set was down-sampled to 6X, 10X and 15X coverage. For each coverage data set, SVs were called and compared to the gold standard set. The results (Figure 3) were similar to those of the NA12878 data set. At 10X coverage, 84% deletions and 92% insertions in the gold standard set can be detected by PBHoney-Spots. The precisions at 10X coverage range from 54% ~ 60% for deletion calls and 31% ~ 43% for insertion calls. At 15X coverage, the recall increased slightly but precision decreased. Thus, 10X may be an optimal coverage to use in practice, considering the sequencing costs and the balance of recall and precision.

***Performance of SV calling using a combination of PBHoney and Sniffles***

Although PBHoney-Spots detected most of the variants, we examined whether we can improve the recall rates by running both PBHoney-Spots and Sniffles, especially under low fold coverages. As shown in Figure 2, at 6X coverage, the union set of both callers discovered 77% deletions in the NA12878 gold standard set, which was 23% more than running PBHoney-Spots alone at 6X coverage and comparable to running PBHoney-

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spots alone at 10X. At 15X coverage, the union set recalled 93% deletions and 88% insertions.

In addition, we tested whether we can get high confidence calls by running both callers. We evaluated precision of the intersection call sets of both callers on 6X, 10X and 15X data sets of the HX1 genome (Figure 3 B, D). The precision of the intersection sets was 87% ~ 90% for deletion calls and 64% ~ 73% for insertion calls, which was half to one-fold higher than that of PBHoney-Spots only.

***Evaluation on Mendelian Errors***

As the germline mutation rate is very low [19,20], Mendelian errors are more likely a result of genotyping errors and can be used as a quality control criteria in genome sequencing [21]. Here, we evaluated the errors of allele drop-in (ADI), which means that an offspring presents an allele that does not appear in either parent, using a whole genome sequencing data set of an Ashkenazi Jewish (AJ) family trio released by NIST [14]. The sequencing data of AJ son, AJ father and AJ mother was down-sampled to 10X coverage. SVs were called using PBHoney-Tails, PBHoney-Spots and Sniffles. The calls from AJ son were compared with calls from AJ father and AJ mother. ADI rate was calculated as the proportion of calls in offspring not matching any call from either parent. The result shows that PBHoney-Spots returns the most calls. For deletion calls, PBHoney-Spots gives us a lowest ADI rate (14.1%), while the ADI rates for insertion calls are considerable higher (31.8% ~ 41.8). Therefore, further validation or manual inspection of the calls is needed when analyzing SVs that may be associated with diseases with low coverage sequencing.

***Automated pipeline for SV calling using PBhoney and Sniffles***

Although we can get highly confident calls at low PacBio coverage using PBhoney and Sniffles, there are still challenges for installation, execution and integration of the aligners and SV callers for average users. Therefore, we developed NextSV, an automated computational pipeline that allows SV calling from PacBio sequencing data using PBhoney and Sniffles. The workflow of NextSV is shown in Figure 4. Two

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mapping tools (BWA-MEM, BLASR), three SV callers (PBHoney-Tails, PBHoney-Spots and Sniffles) and some accessory programs (such as SAMtools, BEDtools) were included in NextSV. NextSV takes FASTA or FASTQ files as input. Once the SV caller is selected, NextSV automatically chooses the compatible aligner and performs mapping. The alignments will be automatically sorted and then presented to the SV caller with appropriate parameters. When the analysis is finished, NextSV will examine the FASTA/FASTQ, BAM, and result files and generate a report showing various statistics. If more than one caller is selected, NextSV will format the raw result files (.tails, .spots, or .vcf files) into bed files and generate the intersection or union call set for the purpose of higher accuracy or sensitivity. In addition, NextSV also supports analyzing high coverage samples via Sun Grid Engine (SGE), a popular batch-queuing system in cluster environment. NextSV splits the input FASTA/FASTQ file into several files of equal sizes and generates mapping task for each file. The mapping tasks are then submitted to the queue. After mapping is done, the alignments are automatically merged and subjected to the caller.

***Computational Performance of NextSV***

To evaluate the computational resources consumed by NextSV, we used the whole genome sequencing data set of HX1 (10X coverage) for benchmarking. All aligners and SV callers in NextSV were tested using a machine equipped with 12-core Intel Xeon 2.66 GHz CPU and 48 Gigabytes of memory. As shown in Table 5, mapping is the most time-consuming step. BLASR takes about 80 hours to map the reads, whereas BWA-MEM needs 27 hours. The SV calling step is much faster. PBHoney-Spots and Sniffles take about 1 hour, while PBHoney-Tails needs 0.27 hours. In total, the BLASR / PBHoney-Spots pipeline takes 80.8 hours while the BWA-MEM / Sniffles pipeline takes 28.1 hours, two thirds less than the former one. Since the BLASR/PBHoney-Spots pipeline has improved performance on SV calling and the BWA-MEM/Sniffles pipeline is faster and complementary of PBHoney, we suggest running both to get the best results in practice.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Discussion

Depth of coverage is often a key consideration in genomic analyses [22]. In this study, we evaluated SV calling performance of three SV calling algorithms, PBHoney-Tails, PBHoney-Spots and Sniffles, at various PacBio coverages of 2 ~ 15X. Our results showed that, at 10X coverage, 76% ~ 84% deletions and 80% ~ 92 % insertions were detected by running PBHoney-Spots. By running both PBHoney-Spots and Sniffles, comparable recall can be achieved at coverage as low as 6X. At more than 10X coverage, the recall slightly increased. Thus, 10X can be an optimal PacBio coverage for efficient SV detection, yet 6X may also be an economic choice under limited budget.

Given the long read length, structural variants can be spanned by reads. In our results, the “Spots” algorithm of PBHoney, which was specifically designed for detection of intra-read SV events, uncovered the most calls among the three algorithms. Sniffles was a newly designed SV caller, and its pre-publication release version was tested in our study. There are several advantages of running both PBHoney and Sniffles. First, the overlapping calls are more accurate. In our results, the precisions of the intersection sets were half to one-fold higher than those of PBHoney-Spots only. The recall of the intersection set was 45% at 10X coverage, meaning that 45% calls can be detected at a very high accuracy. Second, more calls can be discovered by running both, especially for deletion calls. In our results, under 6X coverage, the union call set of two callers covered 77% deletions in the NA12878 gold standard set, which was 23% more than the call set of PBHoney-Spots alone. In addition, by running both BLASR/PBHoney and BWA-MEM/Sniffles, we can have two BAM files for necessary manual inspection, potentially eliminating the mapping artifacts that are specific to one aligner.

Besides installation of the aligners and callers, several steps are required to perform SV detection using the combination of PBHoney and Sniffles, including quality check, mapping, sorting, SV calling, generating union/intersection call set, and generating summary statistics. In addition, several issues need to be considered during analysis. PBHoney typically takes alignments from BLASR as input but Sniffles requires output from BWA-MEM. The output files of PBHoney in tails or spots format should be

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

converted to standard format (such as bed or vcf) for the convenience of further analysis. When two calls are merged, original information from each caller should be retained. Therefore, we developed NextSV, a comprehensive solution to address this. NextSV is available at http://github.com/Nextomics/NextSV. We believe that NextSV will facilitate the detection of structural variants from low fold of PacBio sequencing data.

## Acknowledgments

The authors wish to thank the National Institute of Standards and Technology and Genome in a Bottle Consortium for making the reference data on PacBio sequencing available to benchmark bioinformatics software tools. We also thank members of Grandomics to test the software tools and offering valuable feedback.

## Author Contributions

L.F. performed the evaluation and wrote the NextSV software. J.H. and D.W. tested the software and advised on the study. K.W. conceived of and supervised the study.

## Competing Interests

L.F., J.H. and D.W. are employees and K.W. is a consultant for Grandomics Biosciences.

## References

1 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).
2 Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**, R52 (2010).
3 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**, 437-455 (2010).
4 Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125-138 (2013).
5 Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929 (2013).
6 Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotechnol* **32**, 1106-1112 (2014).
7 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481 (2009).
8 Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**, 224-238 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

9 English, A. C. *et al.* Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* **16**, 286 (2015).
10 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611 (2015).
11 Merker, J. *et al.* Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease. *bioRxiv* doi:10.1101/090985 (2016).
12 English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 180 (2014).
13 Parikh, H. *et al.* svclassify: a method to establish benchmark structural variant calls. *BMC Genomics* **17**, 64 (2016).
14 Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**, 160025 (2016).
15 Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**, 12065 (2016).
16 Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
17 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v2 [q-bio.GN] (2013).
18 Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**, 780-786 (2015).
19 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).
20 Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat Rev Genet* **13**, 565-575 (2012).
21 Pilipenko, V. V. *et al.* Using Mendelian inheritance errors as quality control criteria in whole genome sequencing data set. *BMC Proc* **8**, S21 (2014).
22 Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121-132 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Figure and Tables

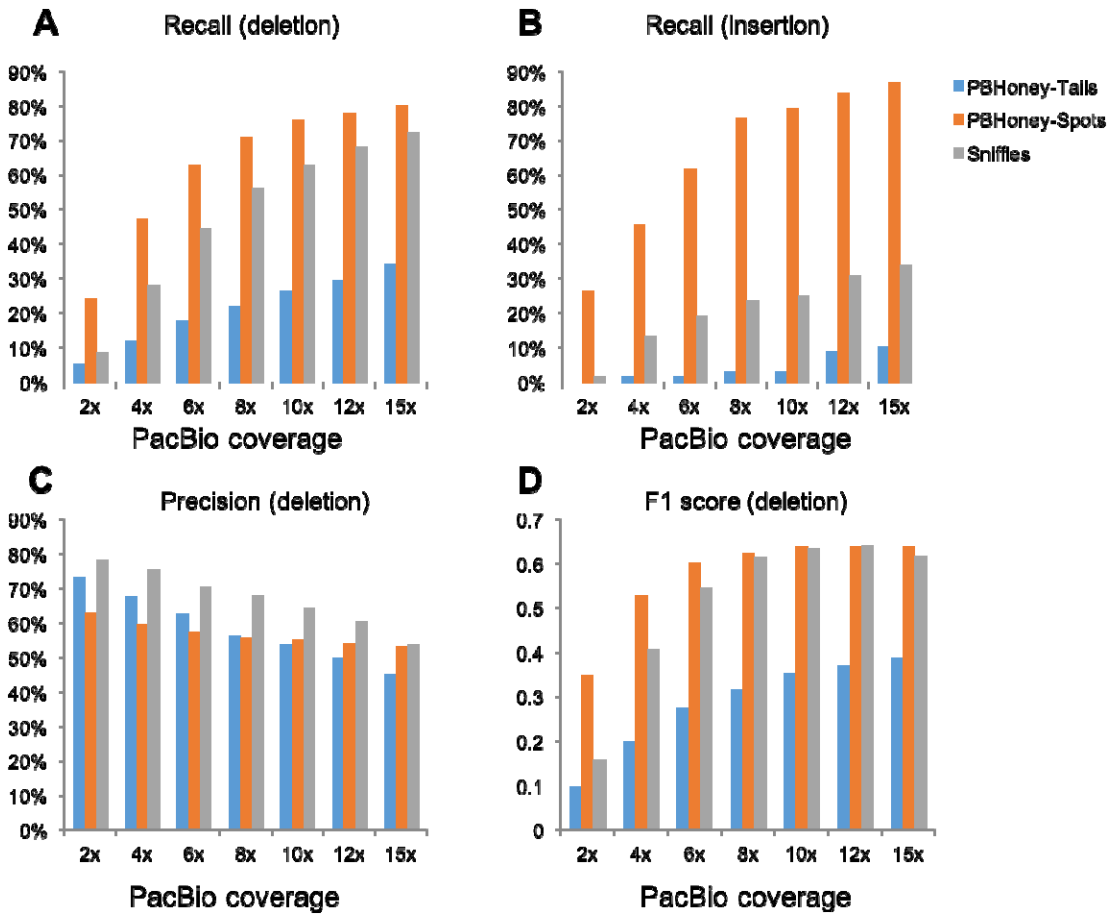


Figure 1. SV calling performance for each SV caller under different coverage on the NA12878 genome.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

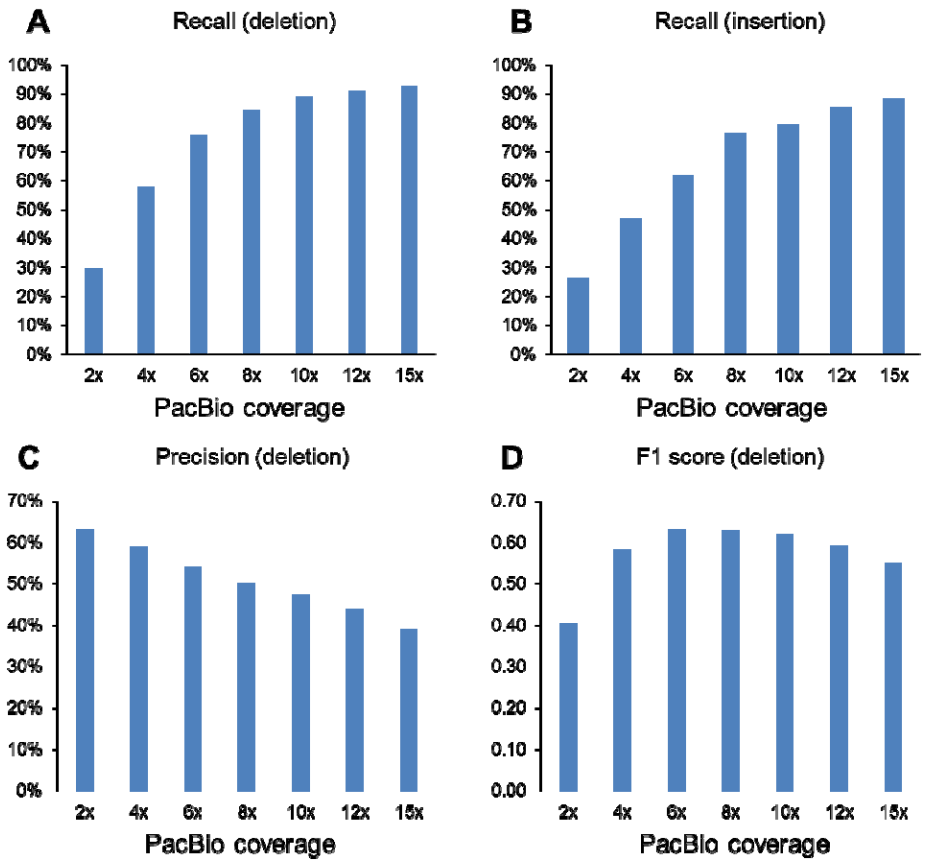


Figure 2. SV calling performance for the union call set of PBHoney-Spots and Sniffles under different coverage on the NA12878 genome.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

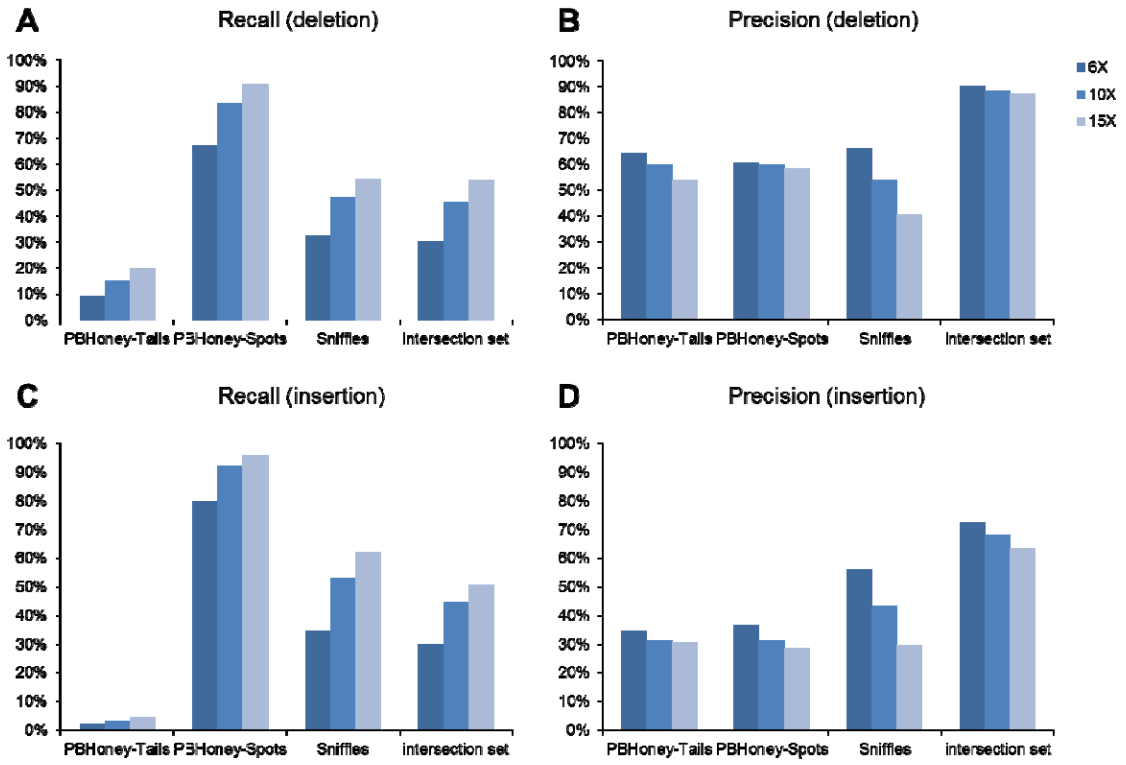


Figure 3. SV calling performance on the HX1 genome. Recalls and precisions of PBHoney-Tails, PBHoney-Spots , Sniffles and the intersection set of the latter two are shown.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

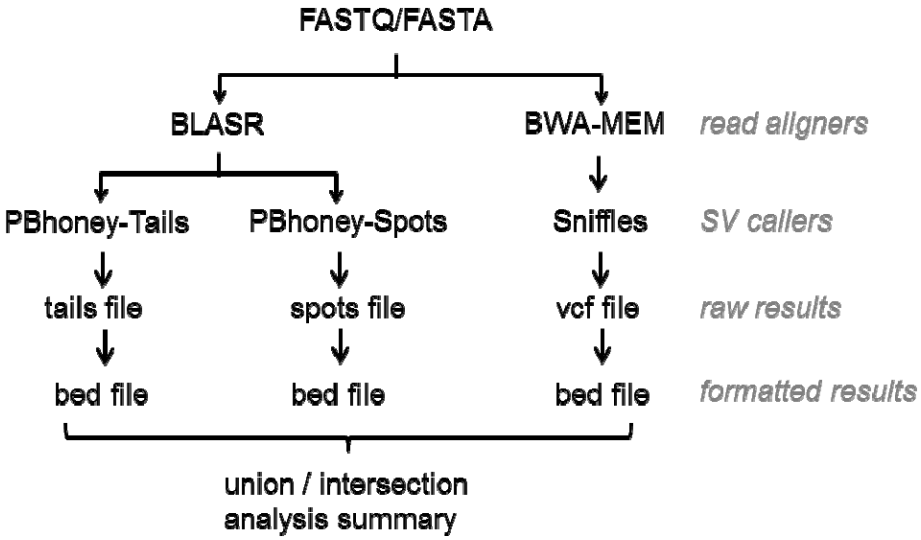


Figure 4. Scheme of NextSV workflow.

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Table 1. Description of PacBio data sets used for this study.**

| Data Source / Accession | Genome | Down-sampled Coverage | Mean Read Length | Reference |
|---|---|---|---|---|
| SRX627421 | NA12878 | 2~15X | 4.9 kb | [18] |
| SRX1424851 | HX1 | 6~15X | 7.0 kb | [15] |
| NIST | AJ son | 10X | 8.0 kb | [14] |
| NIST | AJ father | 10X | 7.3 kb | [14] |
| NIST | AJ mother | 10X | 7.8 kb | [14] |

**Table 2. Number of calls in gold standard SV set**

| Genome | Platform | Number of Deletions (≥ 200bp) | Number of Insertions (≥ 200bp) | Reference |
|---|---|---|---|---|
| NA12878 | Illumina | 2094 | 68 | [13] |
| HX1 | PacBio | 2976 | 2944 | [15] |

**Table 3. Mendelian error of deletion calls under 10X coverage**

| | PBhoney-Tails | PBhoney-Spots | Sniffles | Union set |
|---|---|---|---|---|
| No. of calls (AJ father) | 775 | 2944 | 2206 | 4020 |
| No. of calls (AJ mother) | 789 | 3091 | 2178 | 4165 |
| No. of calls (AJ son) | 728 | 3121 | 2198 | 4090 |
| No. of calls inherited from father | 370 | 1867 | 1006 | 2356 |
| No. of calls inherited from mother | 375 | 2095 | 987 | 2539 |
| No. of ADI | 282 | 441 | 814 | 937 |
| ADI rate | 38.6% | 14.1% | 37.0% | 22.9% |

**Table 4. Mendelian error of insertion calls under 10X coverage**

| | PBhoney-Tails | PBhoney-Spots | Sniffles | Union set |
|---|---|---|---|---|
| No. of calls (AJ father) | 168 | 6691 | 1096 | 6952 |
| No. of calls (AJ mother) | 148 | 7183 | 1181 | 7476 |
| No. of calls (AJ son) | 151 | 7522 | 1148 | 7778 |
| No. of calls inherited from father | 104 | 2952 | 452 | 3897 |
| No. of calls inherited from mother | 87 | 3541 | 476 | 3986 |
| No. of ADI | 49 | 2721 | 479 | 2911 |
| ADI rate | 31.8% | 36.2% | 41.8% | 37.4 |

bioRxiv preprint doi: https://doi.org/10.1101/092544; this version posted December 17, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

390 **Table 5. Time consumption for each steps in the NextSV pipeline for 10X PacBio data set**

| SV caller | Aligner | CPU (number of threads) | Alignment time (hour) | SV calling time (hour) | Total Time (hour) |
|---|---|---|---|---|---|
| PBhoney | BLASR | 12 | 79.6 | 0.27 (Tails)<br>0.96 (Spots) | 80.8 |
| Sniffles | BWA-MEM | 12 | 27.0 | 1.08 | 28.1 |

391

392