NextSV: a computational pipeline for structural variation analysis from low-coverage long-read sequencing

Li Fang[1,2], 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

*Please address correspondence to:

Kai Wang, Ph.D.

622 W 168th St,

Room PH20-313,

New York, NY 10032

Tel: 212-305-3234

Email: kaichop@gmail.com

Abbreviations: ADI, allele drop-in; AJ, Ashkenazi Jewish; SV, structural variants.

Abstract

Structural variants (SVs) in human genomes are implicated in a variety of human diseases. Long-read sequencing delivers much longer read lengths than short-read sequencing 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 developed NextSV, a meta SV caller and a computational pipeline to perform SV calling from low coverage long-read sequencing data. NextSV integrates three aligners and three SV callers and generates two integrated call sets (sensitive / stringent) for different analysis purpose. We evaluated SV calling performance of NextSV under different PacBio coverages on two personal genomes, NA12878 and HX1. Our results showed that, compared with running any single SV caller, NextSV stringent call set had higher precision and balanced accuracy (F1 value) while NextSV sensitive call set had a higher recall. At 10X coverage, the recall of NextSV sensitive call set was 93.5%~94.1% for deletions and 87.9%~93.2% for insertions, indicating that ~10X coverage might be an optimal coverage to use in practice, considering the balance between the sequencing costs and the recall rates. We further evaluated the Mendelian errors on an Ashkenazi Jewish trio dataset. Our results provide useful guidelines for SV detection from low coverage whole-genome PacBio data and we expect that NextSV will facilitate the analysis of SVs on long-read sequencing data.

Keywords

long-read sequencing, structure variation, structural variants, low coverage, PacBio

1. Introduction

Structural variants (SVs), including large variations such as deletions, insertions, duplications, inversions, and translocations, play important roles in human diversity and disease susceptibility (Feuk et al., 2006; Pang et al., 2010). Many inherited diseases and cancers have been associated with a large number of SVs in recent years (Carvalho and Lupski, 2016; Moncunill et al., 2014; Stankiewicz and Lupski, 2010; Weischenfeldt et al., 2013; Yang et al., 2013; Zhang et al., 2009). 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-read sequencing still remains challenging (English et al., 2015). 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 (Chaisson et al., 2015; English et al., 2015). 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 both negative (Merker et al., 2016). This represents an clear example that long-read sequencing may solve some negative cases in clinical diagnostic settings.

Two popular SV software tools have been developed specifically for long-read sequencing: PBHoney (English et al., 2014) 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. PBHoney uses bam files generated by BLASR (Chaisson and Tesler, 2012) as input while Sniffles requires BAM files from BWA-MEM (Li, 2013) or NGMLR (Rescheneder et al., 2016), a new long-read aligner. 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 aligners and 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.

To address these challenges, we developed NextSV, an automated SV detection pipeline integrating multiple tools. NextSV automatically execute these software tools with optimized parameters for the specific coverage that user specified, then integrates results of each caller and generates a sensitive call set and a stringent call set, for different analysis purpose.

Recently, the Genome in a Bottle (GIAB) consortium and the 1000 Genome Project Consortium released high-confidence SV calls for the NA12878 genome, an extensively sequenced genome by different platforms, enabling benchmarking of SV callers (Parikh et al., 2016; Sudmant et al., 2015). They also published sequencing data of seven human genomes, including PacBio data of an Ashkenazi Jewish (AJ) family trio (Zook et al., 2016). Previously, we sequenced a Chinese individual HX1 on the PacBio platform, and generated assembly-based SV call sets (Shi et al., 2016). Using data sets of NA12878, HX1 and the AJ family trio, we evaluated the performance of four aligner/SV caller combinations (BLASR / PBHoney-Spots, BLASR / PBHoney-Tails, BWA / Sniffles and NGMLR / Sniffles) as well as NextSV under different PacBio coverages. We expect that NextSV will facilitate the detection and analysis of SVs on long-read sequencing data.

2. Results

2.1 NextSV analysis pipeline

As shown in Figure 1, NextSV currently supports four aligner / SV caller combinations: BLASR / PBHoney-Spots, BLASR / PBHoney-Tails, BWA / Sniffles and NGMLR / Sniffles. Some accessory programs (such as SAMtools) are included in NextSV. NextSV extracts FASTQ files from PacBio raw data (.hdf5 or .bam) and performs QC according to users specified settings. Once the aligner / SV caller combination is selected by user, NextSV automatically generates the scripts for alignment, sorting, and SV calling with appropriate parameters. When the analysis is finished, NextSV will format the raw result files (.tails, .spots, or .vcf files) into bed files. If multiple aligner/SV caller combinations are selected, NextSV will integrate the calls to generate a sensitive (by union) and a stringent (by intersection) call set. The output of NextSV is ANNOVAR-compatible, so that users can easily perform downstream annotation using ANNOVAR (Wang et al., 2010). In addition, NextSV also supports job submitting via Sun Grid Engine (SGE), a popular batch-queuing system in cluster environment.

2.2 Performance of SV calling on different coverages of the NA12878 Genome

To determine the optimal coverage for SV detection on PacBio data, we evaluated the performance of NextSV under several different coverages. We downloaded a recently published PacBio data set of NA12878 (Pendleton et al., 2015) and down-sampled the data set to 2X, 4X, 6X, 8X, 10X, 12X, and 15X. SV calling was performed using NextSV under each coverage. All supported aligner/SV caller combinations were run. At least two supporting reads is required for all SV calls. The resulting calls were compared with the gold standard SV set (including 2094 deletion calls and 1114 insertion calls) described in method section.

First, we examined how many calls in the gold set can be discovered. As shown in Figure 1, the recall increased rapidly before 10X coverage but the slope of increase slowed down after 10X. Among the four aligner / SV caller combinations, BLASR / PBHoney-Spots had the highest recall for insertions while NGMLR / Sniffles had the highest recall for deletions. At 10X coverage, BLASR / PBHoney-Spots detected 76.9% of deletions and 81.8% insertions in the gold standard set; NGMLR / Sniffles discovered 90.9% deletions and 75.1% insertions in the gold standard set. BWA / Sniffles had a lower recall for deletions (72.5%) and insertions (51.3%) than NGMLR / Sniffles, indicating NGMLR is a better aligner for Sniffles. PBHoney-Tails only detected 26.6% deletions and 0.09% insertions. NextSV sensitive call set, which was generated by the union call set of BLASR / PBHoney-Spots, BLASR / PBHoney-Tails, and NGMLR / Sniffles, had the highest recall. At 10X coverage, the recall of NextSV sensitive call set is 93.5~94.1% for deletions and 87.9~93.2% for insertions. At 15X coverage, the recall of NextSV sensitive call set increased slightly. Therefore, 10X coverage might be an optimal coverage to use in practice, considering the relatively high sequencing costs and the generally high recall rates.

Second, we examined the precision and balanced accuracy (F1 scores) under different coverages (Figure 2). The precision is calculated as the fraction of detected SVs that matching the gold standard set. For deletions calls, NextSV stringent call set had the second highest precision and highest F1 value. For insertion calls, NextSV stringent call set had the highest precision and F1 value at each coverage. Therefore, NextSV stringent call set performs the best, considering the balance between recall and precision.

2.3 Performance of SV calling on different coverages on the HX1 Genome

To verify the performance of SV detection on different individuals, we also performed evaluation on a Chinese genome HX1, which was sequenced by us recently (Shi et al., 2016) 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 three representative coverages (6X, 10X and 15X). For each coverage, SVs were called using the four pipelines described above and compared to the gold standard set. The results were similar to those of the NA12878 data set (Figure 3). At 10X coverage, NextSV sensitive call set had a recall of 94.1% for deletions and 93.2% for insertions, highest among all the call sets. NextSV stringent call set had the highest precisions and F1 values. Among the four aligner / SV caller combinations, NGMLR / Sniffles discovered the most deletions (91.5%) and BLASR / PBHoney-Spots discovered the most insertions (81.7%) at 10X coverage. BWA / Sniffles had a higher precision but a lower recall and F1 value than NGMLR / Sniffles.

2.4 Evaluation on Mendelian Errors

As the germline mutation rate is very low (Kong et al., 2012; Veltman and Brunner, 2012), Mendelian errors are more likely a result of genotyping errors and can be used as a quality control criteria in genome sequencing (Pilipenko et al., 2014). Due to the lack of gold standard call sets, here, we evaluated the errors of allele drop-in (ADI), which means that the presence of an alleles in offspring that does not appear in either parent. We used a whole genome sequencing data set of an AJ family trio released by NIST (Zook et al., 2016) to do the evaluation. The sequencing data of AJ son, AJ father and AJ mother was down-sampled to 10X coverage. SV detection was performed using NextSV with all supported aligners and SV callers enabled. The calls from AJ son were compared with calls from AJ father and AJ mother. The results showed that, NextSV stringent call set had the lowest ADI rate for both deletions (10.4%) and insertions (23.5%). Among the four aligner/SV caller combinations, NGMLR / Sniffles was the best for both deletions and insertions.

2.5 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 NGMLR needs 11.2 hours, which is the fastest among the three aligners. The SV calling step is much faster. PBHoney-Spots and Sniffles take about 1 hour, while PBHoney-Tails needs 0.27 hour. In total, the BLASR / PBHoney combination takes 80.8 hours while the NGMLR / Sniffles combination takes 12.5 hours, 84.5% less than the former one. Since BLASR/PBHoney-Spots and NGMLR / Sniffles have good performance on SV calling and running PBHoney-Tails is very fast given the BLASR output, the NextSV pipeline will execute the three methods by default for generating the final results.

3. Discussion

Long-read sequencing such as PacBio sequencing has clear advantages over short-read sequencing on SV discovery (English et al., 2015). However, its application in real-world setting is often limited due to the relatively high sequencing cost and hence the relatively low sequencing coverage. In this study, we developed NextSV, a computational pipeline integrating multiple aligners and SV callers to improve SV discovery on low-coverage PacBio data sets. Our results showed that, NextSV stringent call set had the highest precisions and F1 values while NextSV sensitive call set had the highest recall. At 10X coverage, the recall of NextSV sensitive call set was 93.5%~94.1% for deletions and 87.9%~93.2% for insertions. At 15X coverage, there is only a slight increase in recall. Therefore, ~10X coverage can be an optimal coverage to use in practice, considering the balance between the sequencing costs and the recall rates.

There is often a trade-off between recall and precision. NextSV generates a sensitive call set and a stringent call set, for different purposes. NextSV sensitive call set is suitable for users who consider recall more important than precision and who can afford extensive downstream analysis (such as Sanger sequencing) to validate the candidate variants. This is often the case when doing disease-casual variant discovery on personal genomes. NextSV stringent call set has the highest precision, F1 value and Mendelian error. It is suitable for users who aim to perform genome-wide analysis of SVs on a collection of samples, with limited downstream validation.

The performance of SV callers are affected by the parameter settings. By default, PBHoney requires a minimal read support of 3 for an SV event and Sniffles requires a minimal read support of 10 for an SV event. However, this may be too high for low coverage data set. In our evaluation of recall and precision, we changed this setting to require a minimal read support of 2. This allows detection SVs from very low coverage regions, with an acceptable precision. This result in substantially higher number of true positives and less variants of interest would be missed. The increased false positive calls can be removed by downstream validation or using call sets of two SV callers (e.g. using the NextSV stringent call set).

In addition to test recalls and precisions, we examined the allele drop-in errors, which represent the SV calls that in the offspring but not appear in either parent. The allele drop-in errors can come from two sources: false positive calls of the offspring or false negatives in the parents, though in very rare cases it could be due to de novo mutations. So this measure is related to both recall and precision. Since we consider 10X coverage as a good choice, we did the evaluation on a family trio data set with ~10X coverage. In our results, NextSV stringent call set has the lowest allele drop-in error, which is consistent with the results that it has the highest F1 value.

NextSV currently supports four aligner / SV caller combinations: BLASR / PBHoney-Spots, BLASR / PBHoney-Tails, BWA / Sniffles, NGMLR / Sniffles, but we expect to continuously expand the support for other aligner / caller combinations. Users can choose to run any of them. By default, NextSV will enable BLASR / PBHoney-Spots, BLASR / PBHoney-Tails and NGMLR / Sniffles and integrate the results to generate the sensitive calls and stringent calls. We do not enable BWA / Sniffles by default because Sniffles works better with NGMLR in our evaluation and alignment is a time consuming step. SVs that are shorter than reads may result in intra-read discordances while larger SVs may result in soft-clipped tails of long reads. We suggest running both PBHoney-Spots and PBHoney-Tails because they are two complementary algorithms designed to detect intra-read discordances and soft-clipped tails, respectively. Sniffles uses multiple evidences to detect SV so it should be suitable for both small and large SVs.

In this study, we only evaluated the performance for insertions and deletions because we only have the gold standard calls of insertions and deletions. This is another limitation of the study. We will evaluate the performance on other types of SVs in the future when more gold standard SV calls are available. Nonetheless, NextSV generates SV calls of all types. The output of NextSV is in ANNOVAR-compatible bed format. Users can easily perform downstream annotation using ANNOVAR and disease gene discovery using Phenolyzer (Yang et al., 2015). NextSV is available at http://github.com/Nextomics/NextSV and can be installed by one simple command. We believe that NextSV will facilitate the detection of structural variants from low coverage long-read sequencing data.

4. Materials and Methods

4.1 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 downloaded from NCBI SRA database. Data sets of the 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 (reads that can be used for analysis) using the SMRT Portal software (Pacific Biosciences, Menlo Park, CA) with filtering parameters (minReadScore=0.75, minLength=500). The subreads were mapped to the reference genome using BLASR (Chaisson and Tesler, 2012), BWA-MEM (Li, 2013) or NGMLR (Rescheneder et al., 2016). 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.

4.2 SV detection using BLASR / PBHoney-Spots and BLASR / PBHoney-Tails

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 in SAM format were converted to BAM format and sorted by SAMtools. PBHoney-Tails and PBHoney-Spots (from PBSuite-15.8.24) were run

with slightly modified parameters (minimal read support 2, instead of 3 and consensus polishing disabled) to increase sensitivity and to discover SVs under low coverages (2~15X).

4.3 SV detection using BWA / Sniffles and NGMLR / Sniffles

PacBio subreads were aligned to the reference genome, using BWA-MEM (bwa mem -M -x pacbio) or NGMLR (default parameters) to generate the BAM file. The BAM file was sorted by SAMtools, then used as input of Sniffles (version 1.0.5). 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).

4.4 NextSV sensitive call set and NextSV stringent call set

NextSV sensitive call set is generated as

$$\text{SNIF} \cup (\text{SPOT} \cup \text{TAIL}),$$

and NextSV stringent call set is generated as

$$\text{SNIF} \cap (\text{SPOT} \cup \text{TAIL}),$$

where SNI denotes the call set of NGMLR / Sniffles, SPOT denotes the call set of BLASR / PBHoney-Spots and TAIL denotes the call set of BLASR / PBHoney-Tails.

4.5 Comparing two SV call sets

Calls which reciprocally overlapped by more than 50% (bedtools intersect -f 0.5 -F 0.5) were considered to be the concordant SV calls and were 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 taken.

4.6 Gold standard SV call set

The gold standard deletion call set of the NA12878 genome was release by the Genome In A Bottle (GIAB) consortium (Parikh et al., 2016), in which most of the calls were refined by experimental validation or other independent technologies. The gold standard insertion call set of the NA12878 genome was obtained by merging the high-confidence insertion calls of 1000 Genome phase 3 (Sudmant et al., 2015) and high-confidence insertion calls from GIAB. For the HX1 genome, due to the availability of high-coverage (>100X) data, we used the SV calls from a previously validated

local assembly-based approach (Chaisson et al., 2015) as the initial high-quality calls. We also detected SVs on 100X coverage PacBio data set of the HX1 genome using BLASR / PBHoney-Spots, BLASR / PBHoney-Tails, BWA / Sniffles and NGMLR / Sniffles (minimal read support=20 for each SV caller). The initial high-quality calls that overlapped with one of the four 103X call sets 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.

4.7 Performance Evaluation of SV callers

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).

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 software. J.H. and D.W. tested the software and advised on the study. K.W. conceived and supervised the study, and revised the manuscript.

Competing Interests

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

References

Carvalho, C.M., Lupski, J.R., 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17, 224-238.

Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J.M., Stamatoyannopoulos, J.A., Hunkapiller, M.W., Korlach, J., Eichler, E.E., 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608-611.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238.

English, A.C., Salerno, W.J., Hampton, O.A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D.I., Beck, C.R., Davis, C.F., Dahdouli, M., Ma, S., Carroll, A., Veeraraghavan, N., Bruestle, J., Drees, B., Hastie, A., Lam, E.T., White, S., Mishra, P., Wang, M., Han, Y., Zhang, F., Stankiewicz, P., Wheeler, D.A., Reid, J.G., Muzny, D.M., Rogers, J., Sabo, A., Worley, K.C., Lupski, J.R., Boerwinkle, E., Gibbs, R.A., 2015. Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Genomics 16, 286.

English, A.C., Salerno, W.J., Reid, J.G., 2014. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics 15, 180.

Feuk, L., Carson, A.R., Scherer, S.W., 2006. Structural variation in the human genome. Nat Rev Genet 7, 85-97.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W.S., Sigurdsson, G., Walters, G.B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D.F., Helgason, A., Magnusson, O.T., Thorsteinsdottir, U., Stefansson, K., 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471-475.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997v2 [q-bio.GN].

Merker, J., Wenger, A.M., Sneddon, T., Grove, M., Waggott, D., Utiramerur, S., Hou, Y., Lambert, C.C., Eng, K.S., Hickey, L., Korlach, J., Ford, J., Ashley, E.A., 2016. Long-read whole genome sequencing identifies causal structural variation in a Mendelian disease. bioRxiv doi:10.1101/090985.

Moncunill, V., Gonzalez, S., Bea, S., Andrieux, L.O., Salaverria, I., Royo, C., Martinez, L., Puiggros, M., Segura-Wang, M., Stutz, A.M., Navarro, A., Royo, R., Gelpi, J.L., Gut, I.G., Lopez-Otin, C., Orozco, M., Korbel, J.O., Campo, E., Puente, X.S., Torrents, D., 2014. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nat Biotechnol 32, 1106-1112.

Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., Kirkness, E.F., Levy, S., Feuk, L., Scherer, S.W., 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11, R52.

Parikh, H., Mohiyuddin, M., Lam, H.Y., Iyer, H., Chen, D., Pratt, M., Bartha, G., Spies, N., Losert, W., Zook, J.M., Salit, M., 2016. svclassify: a method to establish benchmark structural variant calls. BMC Genomics 17, 64.

Pendleton, M., Sebra, R., Pang, A.W., Ummat, A., Franzen, O., Rausch, T., Stutz, A.M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M.H., Cao, H., Cohain, A., Deikus, G., Durrett, R.E., Blanchard, S.C., Altman, R., Chin, C.S., Guo, Y., Paxinos, E.E., Korbel, J.O., Darnell, R.B., McCombie, W.R., Kwok, P.Y., Mason, C.E., Schadt, E.E., Bashir, A., 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12, 780-786.

Pilipenko, V.V., He, H., Kurowski, B.G., Alexander, E.S., Zhang, X., Ding, L., Mersha, T.B., Kottyan, L., Fardo, D.W., Martin, L.J., 2014. Using Mendelian inheritance errors as quality control criteria in whole genome sequencing data set. BMC Proc 8, S21.

Rescheneder, P., Sedlazeck, F.J., Haeseler, A.V., Schatz, M.C., 2016. NGMLR: Highly accurate read mapping of third generation sequencing reads for improved structural variation analysis, Genome Informatics 2016, Wellcome Genome Campus Conference Centre, Hinxton, Cambridge, UK.

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner, K.E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G.J., Guan, Y., Shen, Y., Evgrafov, O.V., Knowles, J.A., Thibaud-Nissen, F., Schneider, V., Yu, C.Y., Zhou, L., Eichler, E.E., So, K.F., Wang, K., 2016. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7, 12065.

Stankiewicz, P., Lupski, J.R., 2010. Structural variation in the human genome and its role in disease. Annu Rev Med 61, 437-455.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stutz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J.M., Kong, Y., Lameijer, E.W., McCarthy, S., Flicek, P., Gibbs, R.A., Marth, G., Mason, C.E., Menelaou, A., Muzny, D.M., Nelson, B.J., Noor, A., Parrish, N.F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E.E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A.A., Untergasser, A., Walker, J.A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., Genomes Project, C., Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E., Korbel, J.O., 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81.

Veltman, J.A., Brunner, H.G., 2012. De novo mutations in human genetic disease. Nat Rev Genet 13, 565-575.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.

Weischenfeldt, J., Symmons, O., Spitz, F., Korbel, J.O., 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14, 125-138.

Yang, H., Robinson, P.N., Wang, K., 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods 12, 841-843.

Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R., Lee, C., Park, P.J., 2013. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153, 919-929.

Zhang, F., Gu, W., Hurles, M.E., Lupski, J.R., 2009. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10, 451-481.

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., Henaff, E., McIntyre, A.B., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R.M., Chang, C.C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S.T., Zaranek, A.W., Ball, M., Bobe, J., Estep, P., Church, G.M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G.X., Schnall-Levin, M., Ordonez, H.S., Mudivarti, P.A., Giorda, K., Sheng, Y., Rypdal, K.B., Salit, M., 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3, 160025.

Tables

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	(Pendleton et al., 2015)
SRX1424851	HX1	6~15X	7.0 kb	(Shi et al., 2016)
NIST	AJ son	10X	8.0 kb	(Zook et al., 2016)
NIST	AJ father	10X	7.3 kb	(Zook et al., 2016)
NIST	AJ mother	10X	7.8 kb	(Zook et al., 2016)

Table 2. Number of calls in gold standard SV set

Genome	Platform	Number of Deletions (≥ 200bp)	Number of Insertions (≥ 200bp)	Reference
NA12878	Illumina	2094	1114	(Parikh et al., 2016; Sudmant et al., 2015)
HX1	PacBio	2387	2937	(Shi et al., 2016)

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

	BLASR / PBHoney-Spots	BLASR / PBHoney-Tails	BWA / Sniffles	NGMLR / Sniffles	NextSV Sensitive Calls	NextSV Stringent Calls
No. of calls (AJ father)	2943	775	2173	3109	4172	2342
No. of calls (AJ mother)	3090	789	2008	3169	4299	2399
No. of calls (AJ son)	3120	727	1976	3182	4246	2444
No. of calls inherited from father	2047	306	1238	2151	2812	1684
No. of calls inherited from mother	2166	295	1235	2232	2929	1747
No. of ADI	447	296	335	376	600	253
ADI rate	14.3%	40.7%	17.0%	11.8%	14.1%	10.4%

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

	BLASR / PBHoney-Spots	BLASR / PBHoney-Tails	BWA / Sniffles	NGMLR / Sniffles	NextSV Sensitive Calls	NextSV Stringent Calls
No. of calls (AJ father)	4817	18	764	2601	5326	2103
No. of calls (AJ mother)	5151	20	855	2781	5708	2237
No. of calls (AJ son)	5341	17	903	2708	5815	2243
No. of calls inherited from father	2837	5	255	1393	3142	1182
No. of calls inherited from mother	2907	6	247	1458	3228	1238
No. of ADI	1625	10	520	711	1756	528
ADI rate	30.4%	58.8%	57.6%	26.3%	30.2%	23.5%

435

436

437 **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) 0.96 (Spots)	80.8
Sniffles	BWA-MEM	12	27.0	1.1	28.1
Sniffles	NGMLR	12	11.2	1.3	12.5

438

439

Figures

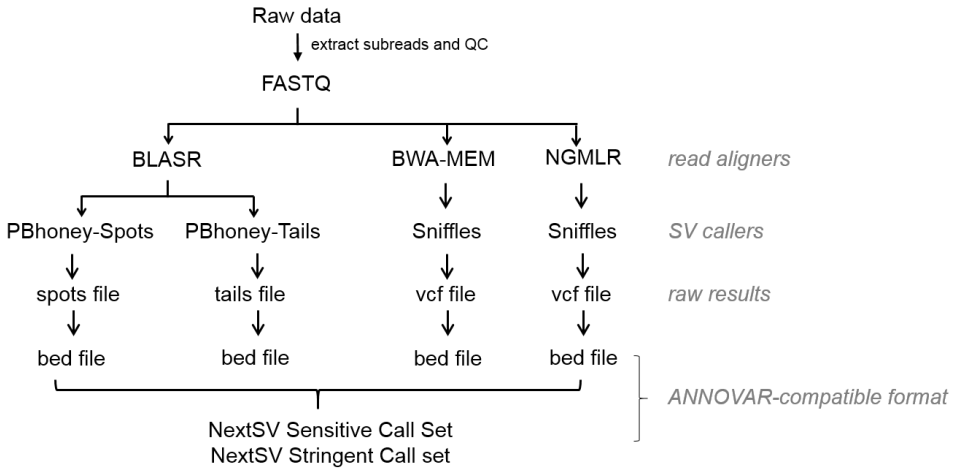

Figure 1. Scheme of NextSV workflow.

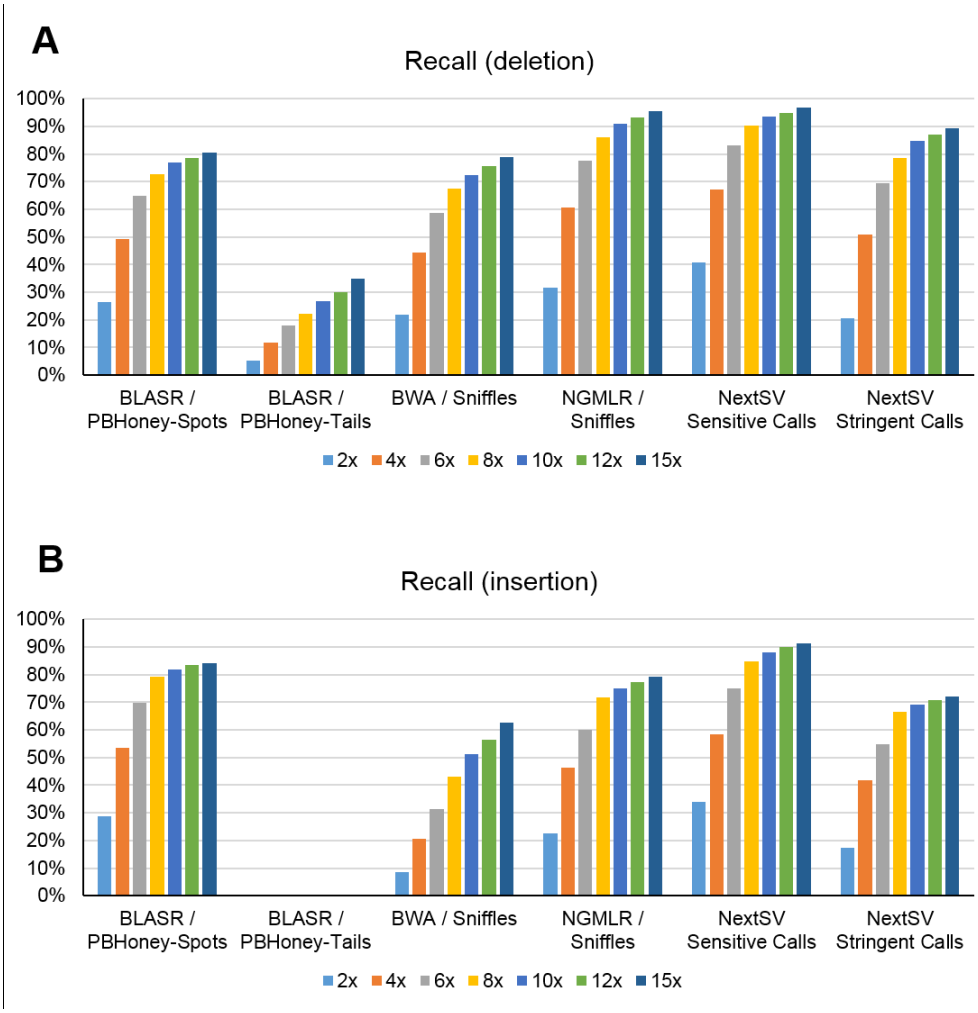

Figure 2. Evaluation of recall rates under different coverages on the NA12878 genome.

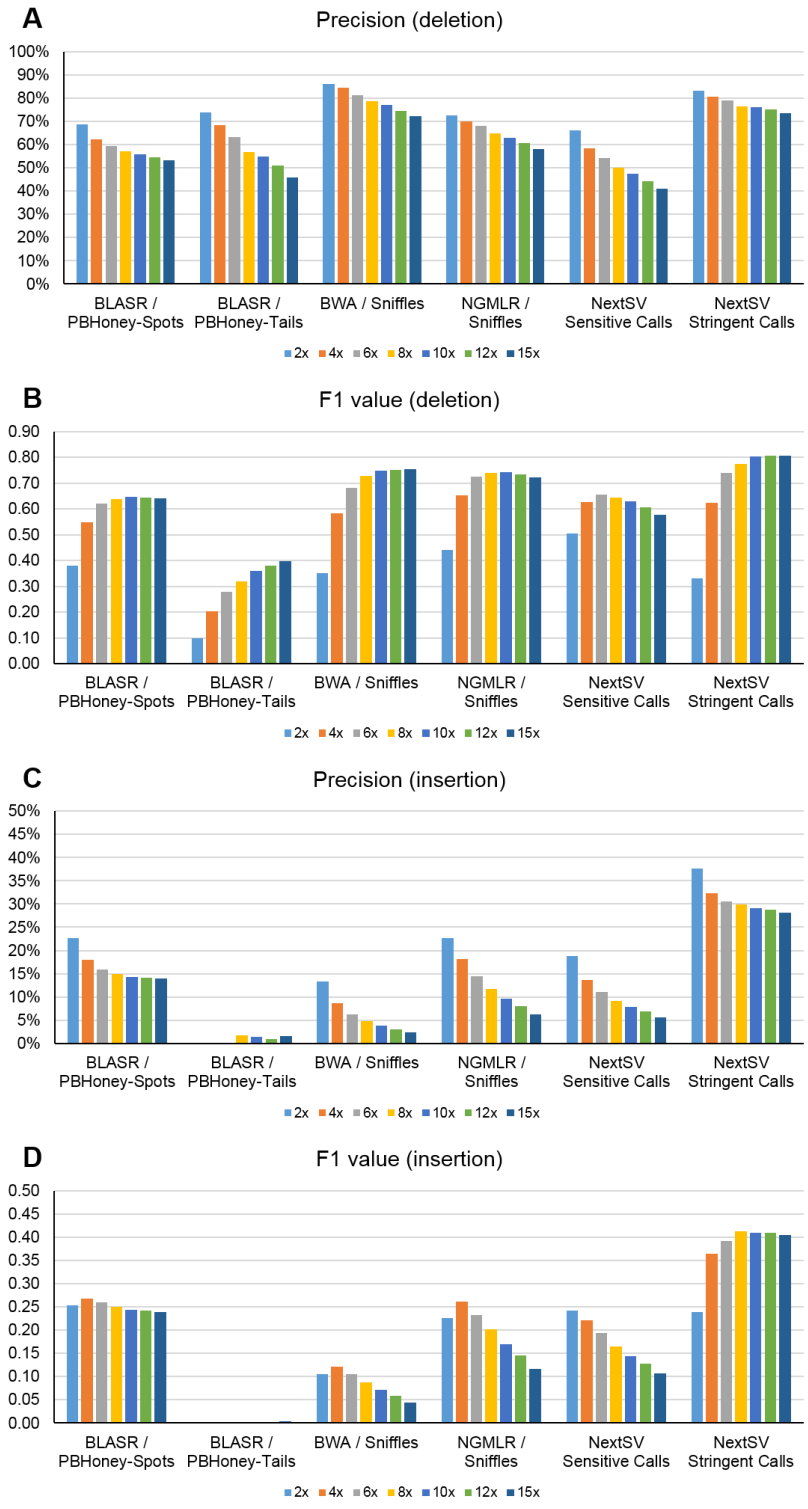

Figure 3. Evaluation of precisions and F1 values under different coverages on the NA12878 genome.

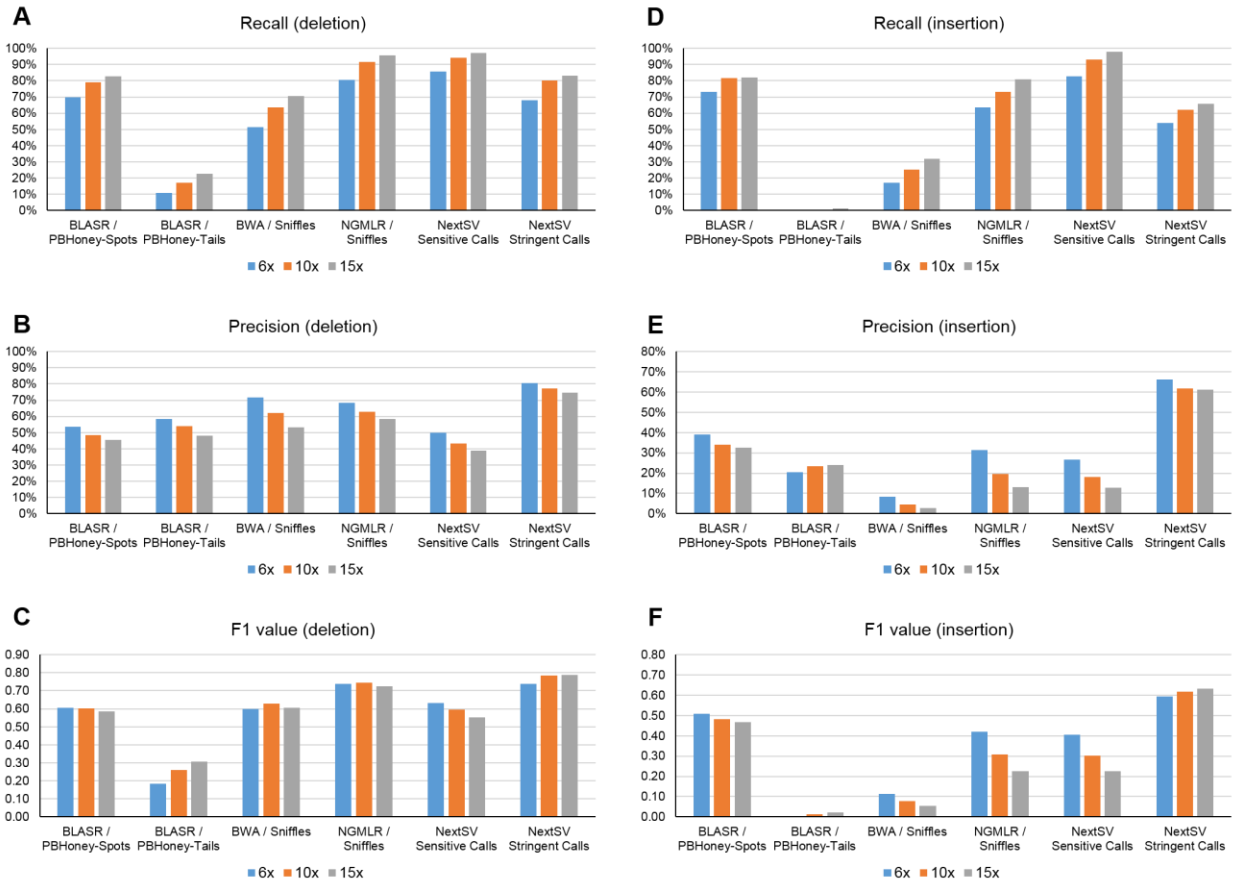

Figure 4. SV calling performance on the HX1 genome.