Genome Biology

METHOD Open Access

A benchmark of structural variation detection by long reads through a realistic simulated model

Nicolas Dierckxsens[1,2*], Tong Li[2], Joris R. Vermeesch[1] and Zhi Xie[2*]

*Correspondence:
nicolasdierckxsens@hotmail.com;
xiezhi@gmail.com
[1] Center for Human Genetics, University Hospital Leuven and KU Leuven, Leuven, Belgium
[2] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

Abstract

Accurate simulations of structural variation distributions and sequencing data are crucial for the development and benchmarking of new tools. We develop Sim-it, a straightforward tool for the simulation of both structural variation and long-read data. These simulations from Sim-it reveal the strengths and weaknesses for current available structural variation callers and long-read sequencing platforms. With these findings, we develop a new method (combiSV) that can combine the results from structural variation callers into a superior call set with increased recall and precision, which is also observed for the latest structural variation benchmark set developed by the GIAB Consortium.

Keywords: Structural variation, Long-read sequencing, Benchmark, Simulated model

Background

In order to decipher the genetic basis of human disease, a comprehensive knowledge of all genetic variation between human genomes is needed. Until recently, the emphasis has been on single-nucleotide polymorphisms, as these variants are easier to trace with current sequencing technologies and algorithms [1, 2]. Over the past 20 years, we gained a better view on the prevalence of structural variation (SV), which changed our perspective on the impact it has on genomic disorders. We now know that structural variation contributes more to inter-individual genetic variation at the nucleotide level than single nucleotide polymorphisms (SNPs) and short indels together [3, 4]. Structural variation covers insertions, deletions, inversions, duplications and translocations that are at least 50 bp in size. The limited length of Next-Generation Sequencing (NGS) reads ($\leq$ 300 bp) hampers the detection of SVs, especially for insertions [3, 5]. These technical limitations can be partially overcome by the third-generation sequencing, which is capable of producing far longer read lengths [6, 7]. The race for dominance on the third-generation sequencing market has significantly reduced the costs per Mb and increased the throughput and accuracy, which makes these technologies (Pacific Biosciences (PacBio) [8] and

Oxford Nanopore Technologies (ONT) [9, 10]) currently the best options for structural variance detection [11].

The introduction of long sequencing reads required a revolution in new computational tools for sequencing analysis. Even though several algorithms for SV detection were developed over the past decade, there is a large discrepancy between their outputs [3]. Assessing the performance of SV detection tools is not straightforward, as there is no gold standard method to accurately identify structural variation in the human genome. To overcome this shortcoming, the Genome in a Bottle (GIAB) Consortium recently published a sequence-resolved benchmark set for identification of SVs, though it only includes deletions and insertions not located in segmental duplications [12]. For as long as there is no completely resolved benchmark available, it is crucial to simulate a human genome with a set of structural variations that resembles reality as close as possible. There are a wide range of structural variation and long sequencing reads simulators available, yet without a thorough benchmark, it is impossible to know which tools are best suited to design the model you want to simulate. Therefore we compared several structural variance and long-read simulators for their system requirements and available features. Furthermore, we introduce Sim-it, a new SV and long-read simulator that we designed for the assessment of SV detection with long-read technologies.

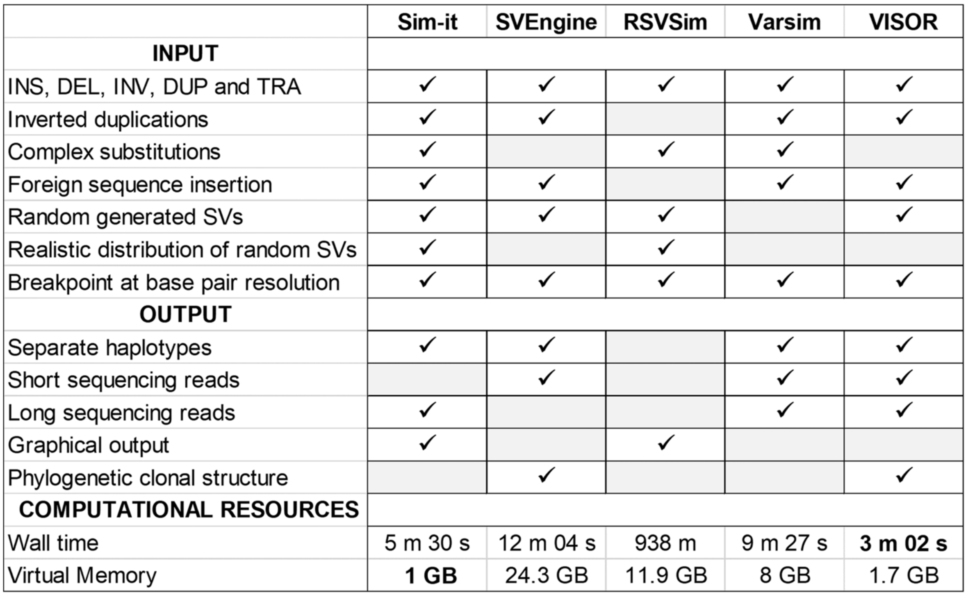
The most extensive structural variance detection study to date identified around 25,000 SVs for each individual by combining a wide range of sequencing platforms [3]. The large amount of sequencing data used for this study makes it too costly to reproduce it on a larger scale, but it can be used to estimate the number of SVs in a human genome. We used the results of this study to produce a realistic model for the evaluation of the available SV detection algorithms and to develop a new script that can improve SV detection by combining the results of existing tools.

Result

Structural variation simulation benchmark

We compared the features and computational resources of five structural variation simulators, as shown in Table 1 and described in the "Methods" section "Benchmark of structural variation simulators". Although all simulators can simulate the most common types of structural variation (insertions, deletions, duplications, inversions, and translocations), more complex SV events need to be included in order to reproduce a realistic SV detection model. For Sim-it, we also included complex substitutions and inverted duplications, both common types of variation in germline and somatic genomes [5, 13–15]. A complex substitution is defined as a region which been deleted and replaced with another region of the genome, while an inverted duplication is a tandem duplication of an inverted segment. Additionally, it is possible to combine random generated SV events with a defined list of SVs at base pair resolution. Random generated SVs will be distributed realistically across the genome with higher prevalence around the telomeres. As output, Sim-it produces a sequence file in FASTA format and optionally long sequencing reads (PacBio or ONT). Additional files to draw gnuplot [16] figures of the length distributions from the simulated SVs are provided (Figs. S4-6). Although none of the other tools has a proprietary method to simulate long reads, Varsim can generate long reads through PBSIM or LongISLND. Currently, Sim-it does not support short read or phylogenetic clonal structure simulation. As for computational resources, Sim-it performed

Table 1 Available features and system requirements of structural variation simulators



	Sim-it	SVEngine	RSVSim	Varsim	VISOR
INPUT					
INS, DEL, INV, DUP and TRA	✓	✓	✓	✓	✓
Inverted duplications	✓	✓		✓	✓
Complex substitutions	✓		✓	✓	
Foreign sequence insertion	✓	✓		✓	✓
Random generated SVs	✓	✓	✓		✓
Realistic distribution of random SVs	✓		✓		
Breakpoint at base pair resolution	✓	✓	✓	✓	✓
OUTPUT					
Separate haplotypes	✓	✓		✓	✓
Short sequencing reads		✓		✓	✓
Long sequencing reads	✓			✓	✓
Graphical output	✓		✓		
Phylogenetic clonal structure		✓			✓
COMPUTATIONAL RESOURCES					
Wall time	5 m 30 s	12 m 04 s	938 m	9 m 27 s	**3 m 02 s**
Virtual Memory	**1 GB**	24.3 GB	11.9 GB	8 GB	1.7 GB

*SCNVsim and SURVIVOR was excluded from the benchmark

best on peak memory consumption and runtime. With 1 GB as peak memory consumption and 5 min 30 s as runtime (single core) to simulate 24,600 SV events, Sim-it can be implemented for any set of SVs on a small desktop or laptop. SVEngine and Varsim also have relatively low runtimes, though a peak memory consumption of respectively 24.3 GB and 8 GB limits it's use on machines with limited computational resources. SCNVsim and SURVIVOR were excluded as they do not accept a set list of SVs as input and have an upper limit of SVs (600 for SCNVsim, less than 24,000 for SURVIVOR) for random simulation.

Long-read simulation benchmark

We assessed the quality of the simulated long reads by comparing their error profiles to those of real PacBio and ONT sequencing reads. Additionally, we compared the features and system requirements for each tool, as described in the "Methods" section "Benchmark of the long-read simulators".

Several systems of ONT and PacBio technologies have been released in the last decade, each with different specifications for the sequencing reads. This complicates an accurate simulation as a specific error profile is needed for each released system. From the 9 tested simulators, only Sim-it, Badread, SURVIVOR and LongISLND support simulations for both ONT and PacBio. Sim-it provides error profiles for ONT, PacBio RS II, PacBio Sequel II, and Pacbio Sequel HiFi systems, while other simulators are limited to one or two error profiles. This shortcoming can be overcome by training a new model for a system, a feature supported by all simulators apart from PBSIM and SimLoRD. This is more laborious and a real dataset along with an accurate reference sequence is required to train a new model. Not all updates require a completely new error profile, therefore we provide the option to adjust the overall accuracy and read length independently from the error profile. Sequencing depths can fluctuate strongly in real datasets, Sim-it can imitate this with a sequencing depth profile file. Such a file can be created with Samtools [17] from an alignment file. As for computational resources, PBSIM performed the best with just 5 min and 0.25 GB of RAM to simulate 15x coverage for chromosome 1 of GRCh38. Besides for DeepSimulator, Badread and NanoSim, computational resources stayed within a reasonable range. Sim-it needed 35 min for chromosome 1, yet this does not represent the real speed of Sim-it. From version 1.3 on, Sim-it uses multiple threads to simulate each chromosome and haplotype in parallel. A complete overview of the features for each long read simulator can be found in Additional file 2: Table S1.

Available features and computational resources determine the suitability and user-friendliness of the simulators, but not the accuracy of the simulation. Therefore, we compared the context-specific error patterns of the simulated reads to real long sequencing datasets. Figure 1A shows the context-specific errors derived from real data from Nanopore PromethION and PacBio Sequel II sequencing reads, as well from their respective simulations by Sim-it. These context-specific error heatmaps were generated for each of the 9 simulators and can be found in Additional file 1: Figs. S1-3. NanoSim generated random errors in stead of a context-specific error pattern, while SURVIVOR, PBSIM and SimLoRD have simplified patterns. For Sim-it, the length of deletions and insertions closely match the real data (Fig. 1C, D). LongISLND has proportionally too many single nucleotide deletions, while the asymmetry for DeepSimulator is caused by a low absolute number of deletions, which is not adjustable. Besides the heatmaps, three more tables can be found in Additional file 1. Two tables (Additional file 1: Tables S2 and S3) with general statistics of the simulated reads and a table with the Euclidean distances for the context-specific errors and for the length distributions of deletion and insertion errors (Additional file 1: Table S3). The Euclidean values confirm the heatmaps and Fig. 1C, D, with LongISLND and Sim-it as the most accurate simulations. LongISLND has the most accurate context-specific errors, while Sim-it has the most accurate length distributions of deletion and insertion errors.

Structural variance detection using simulated long reads

We assessed the performance of 7 long-read SV detection algorithms through a realistic model of 24,600 SV events, as described in the Methods section "SV detection on simulated reads". Additionally, we made a comparison between PacBio and ONT technology and evaluated the impact of the read length and sequencing depth. For each simulated dataset, a separate score for each type of SV and for the four essential parameters that define SVs; namely position, length, type and genotype were calculated.

We performed a complete analysis on each of the 7 SV callers for a Nanopore and a PacBio Sequel II long reads and a HiFi reads dataset with a sequencing depth of 20x (Table 2). For each dataset, Picky had more than 19,000 false positives and false negatives, with an outlier of 46,502 false positives for the PacBio HiFi dataset. We therefore excluded Picky for any further analysis or graphical output. All the statistics of Picky for all three 20x coverage datasets can be examined in Additional file 2.

For a sequencing depth of 20x, cuteSV achieved the best overall performance for the more erroneous reads of Nanopore and PacBio CLR, while SVIM performs best for PacBio HiFi. cuteSV produced the lowest number of false positives independent from sequencing platform, median read length or coverage depth (Table 2 and Fig. 2). Although pbsv generally has a lower recall, it calls the position and length of the SVs more accurately

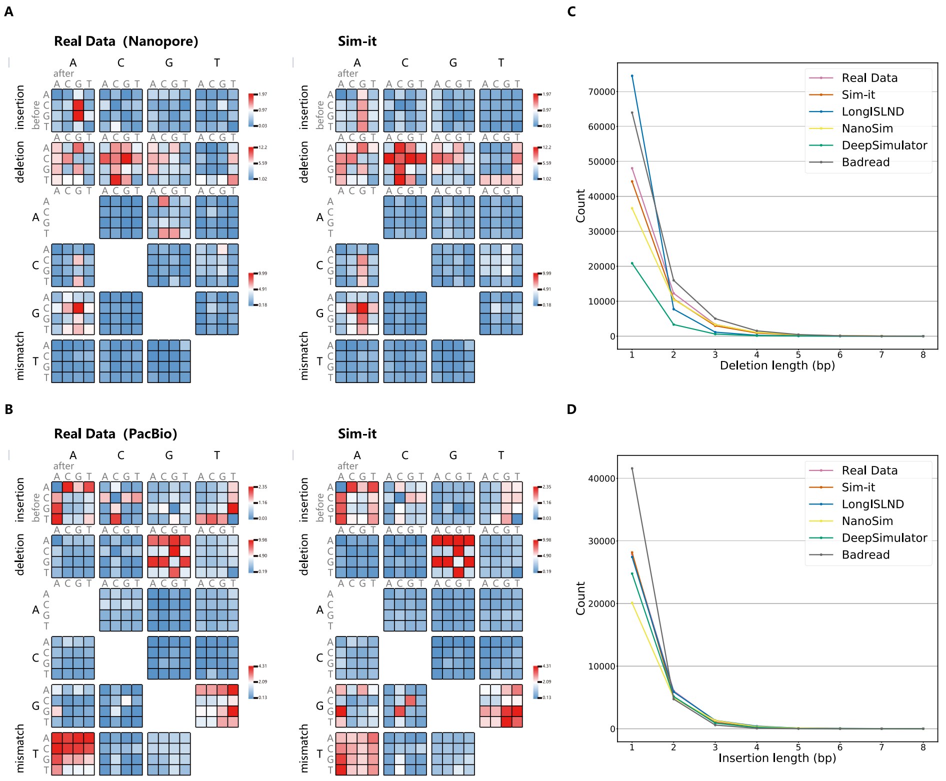


Fig. 1 Context-specific error patterns for mismatches and indels. **A** Context-specific error patterns for real data of Nanopore (9.4.1) and simulated data from Sim-it. **B** Context-specific error patterns for real data of PacBio Sequel II and simulated data from Sim-it. **C** Deletion lengths for real Nanopore data and the simulations of the benchmarked tools. **D** Insertion lengths for real Nanopore data and the simulations of the benchmarked tools

than any other tool, independent from the platform or coverage depth. Subsequently, this high accurateness results in a significant higher number of perfect matches compared to other tools. Perfect matches are SVs called with the correct type, genotype, exact length and position. For PacBio CLR and PacBio HiFi reads, pbsv manages to call respectively 46% and 58.2% of the detected SVs perfectly, which is quite impressive compared to the other tools. Only SVIM achieved a similar percentage for PacBio HiFi reads (56.3%), however not for PacBio CLR reads (7.6%). The highest recall is achieved by NanoSV and to a certain extend NanoVar (only for PacBio HiFi), however this is at the expense of a disproportional number of false positives.

The 24,600 SVs can be classified by 5 different types, namely deletions, insertions, duplications, inversions and complex substitutions. We calculated the recall and precision metrics for each type of SV; Table 3 shows the results for the Nanopore 20x dataset, data metrics for the PacBio 20x and PacBio HiFi 20x datasets reveal similar patterns and can be examined in Additional file 2. NanoSV only classifies insertions, other SVs are indicated as breakend (BND). None of the SV callers classify complex substitutions in their output, which explains the missing precision values for this type. These complex substitutions seem to be the most problematic, as their recall values are very low for each of the tools. Recall and precision values of inversions are also far below the average for each of the tools. The low precision value for duplications detected by NanoVar can be explained by the fact that a significant fraction of the insertions is typed as a duplication.

Table 2 Benchmark statistics on three simulated datasets of 24,600 SVs for 6 existing SV callers and combiSV (combiSV (6): all 6 tools combined)

		Sniffles	cuteSV	pbsv	NanoVar	NanoSV	SVIM	combiSV (6)	SURVIVOR (6)
Nanopore	Recall	80.1%	78.7%	74.3%	81.3%	84.0%	81.6%	81.5%	78.8%
	Precision	97.2%	98.4%	94.9%	87.1%	88.4%	92.6%	98.1%	97.9%
	F-score	**87.8%**	**87.4%**	**83.3%**	**84.1%**	**86.1%**	**86.8%**	**89.0%**	**87.3%**
	Perfect matches	11.4%	3.9%	34.2%	1.6%	0.5%	6.2%	31.4%	7.4%
	Position score	85.4%	77.9%	88.9%	77.2%	87.2%	80.8%	88.6%	80.5%
	Length score	87.9%	87.9%	92.0%	78.4%	87.7%	87.7%	91.9%	88.4%
	Type score	92.5%	94.5%	94.0%	88.2%	44.6%	94.0%	93.7%	93.9%
	Genotype score	58.7%	95.6%	94.1%	90.7%	94.7%	92.7%	95.4%	94.2%
	Total score	**63.3%**	**67.0%**	**64.0%**	**54.8%**	**56.3%**	**64.7%**	**73.1%**	**67.3%**
PacBio	Recall	78.8%	77.7%	74.1%	81.7%	84.1%	81.7%	81.3%	78.8%
	Precision	97.7%	98.6%	95.1%	87.2%	89.0%	93.3%	98.2%	98.0%
	F-score	**87.2%**	**86.9%**	**83.3%**	**84.4%**	**86.5%**	**87.1%**	**88.9%**	**87.3%**
	Perfect matches	12.6%	2.3%	46.0%	2.2%	4.2%	7.6%	41.9%	7.3%
	Position score	85.0%	77.1%	89.8%	77.5%	86.7%	81.2%	89.5%	82.0%
	Length score	89.7%	89.4%	92.9%	75.0%	90.5%	89.4%	93.0%	88.8%
	Type score	94.1%	94.5%	93.6%	87.7%	44.9%	94.1%	94.2%	89.6%
	Genotype score	58.5%	95.7%	94.2%	90.5%	94.7%	92.4%	95.0%	90.1%
	Total score	**63.1%**	**66.3%**	**64.5%**	**54.8%**	**57.6%**	**65.7%**	**73.5%**	**66.5%**
PacBio HiFi	Recall	71.1%	73.7%	73.4%	83.4%	80.9%	77.3%	77.5%	75.2%
	Precision	98.1%	98.8%	96.6%	87.3%	95.5%	96.4%	98.6%	97.9%
	F-score	**82.4%**	**84.4%**	**83.4%**	**85.3%**	**87.6%**	**85.8%**	**86.7%**	**85.1%**
	Perfect matches	25.2%	10.4%	58.2%	12.5%	28.9%	56.3%	56.2%	47.9%
	Position score	90.2%	82.1%	91.4%	84.6%	91.3%	91.0%	91.1%	90.3%
	Length score	92.8%	92.6%	93.5%	84.2%	93.1%	93.5%	94.2%	92.2%
	Type score	93.3%	93.9%	93.8%	88.7%	42.4%	93.7%	94.2%	94.1%
	Genotype score	47.2%	94.8%	92.8%	92.0%	96.1%	94.6%	95.2%	95.1%
	Total score	**57.4%**	**64.7%**	**65.4%**	**60.3%**	**63.2%**	**68.9%**	**71.0%**	**68.1%**
GIAB (Nanopore)	Recall	92.2%	93.6%	92.9%	89.5%	93.2%	93.5%	95.0%	93.5%
	Precision	92.3%	91.6%	89.9%	67.5%	62.3%	84.5%	92.4%	88.2%
	F-score	**92.2%**	**92.6%**	**91.4%**	**76.9%**	**74.7%**	**88.8%**	**93.7%**	**90.8%**
	Perfect matches	0.1%	1.4%	26.4%	0.7%	0.5%	2.3%	25.4%	1.7%
	Position score	67.2%	59.4%	73.5%	61.8%	73.0%	62.2%	72.6%	60.1%
	Length score	80.4%	78.8%	86.3%	50.1%	77.5%	79.6%	85.7%	81.0%
	Type score	97.8%	97.7%	93.2%	61.3%	51.0%	96.5%	98.6%	98.5%
	Genotype score	38.4%	96.7%	90.5%	82.3%	89.4%	85.5%	94.2%	86.9%
	Total score	**57.0%**	**64.9%**	**67.0%**	**13.6%**	**11.5%**	**55.0%**	**72.6%**	**59.8%**

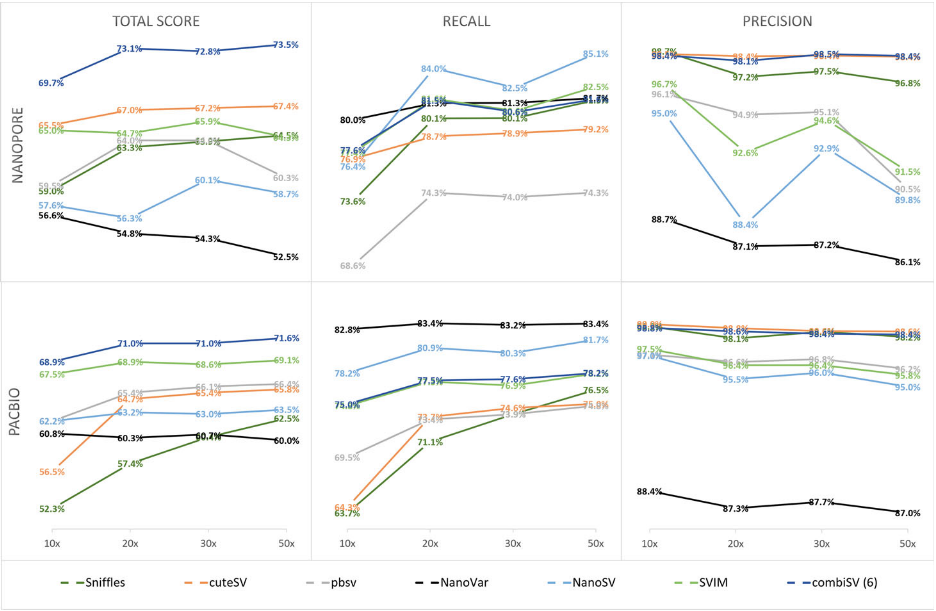


Fig. 2 Structural variance detection stats for a series of Nanopore and PacBio HiFi datasets with increasing sequencing depths

Table 3 Precision and recall statistics for each type of SV from the Nanopore 20x dataset. (combiSV (6): all 6 tools combined)

		Sniffles	cuteSV	pbsv	NanoVar	NanoSV	SVIM	combiSV (6)	SURVIVOR (6)
Precision	Deletions	92.3%	97.3%	97.5%	88.5%	-	94.3%	95.3%	97.0%
	Insertions	88.8%	90.1%	87.2%	81.2%	84.8%	83.8%	89.2%	89.1%
	Duplications	67.0%	73.7%	51.5%	22.8%	-	65.4%	76.3%	72.2%
	Inversions	54.0%	39.4%	77.2%	63.1%	-	77.8%	66.1%	37.8%
	Complex substitutions	-	-	-	-	-	-		
Recall	Deletions	91.4%	93.9%	92.6%	92.6%	95.3%	95.8%	95.6%	91.6%
	Insertions	85.5%	84.1%	73.4%	88.9%	88.5%	87.7%	87.1%	85.6%
	Duplications	89.6%	83.1%	87.6%	93.2%	94.6%	93.6%	89.2%	91.6%
	Inversions	67.6%	50.6%	53.5%	63.5%	65.9%	50.0%	60.6%	66.5%
	Complex substitutions	18.0%	6.4%	5.7%	9.7%	23.8%	8.2%	10.5%	7.0%

To investigate the influence of increased sequencing coverage, we simulated 4 different datasets with sequencing depths of 10x, 20x, 30x, and 50x for both Nanopore and PacBio HiFi (Fig. 2). The general trends for increased sequencing depth are an increased recall and decreased precision, although depending on the tool, they can be disproportional to each other. NanoVar was designed to work on low sequencing depths and therefore does not display much gain in recall, yet a significant reduction in precision. Sniffles benefits the most from additional coverage with increasing recall together with almost no loss of precision. pbsv has a stable performance across all coverages, except for Nanopore 50x, which exhibits a steep decrease of precision. The big drop in precision for NanoSV and SVIM at 20x and 50x coverage of Nanopore are caused by the additional filtering step we implemented for the "minimal read support" (3 for 10x and 20x, 5 for 30x and 50x), which is necessary to keep a good balance between recall and precision. This pattern is to some degree visible for all tools, with an exception to cuteSV, which has stable precision values across all coverages.

Besides sequencing depth, it is often believed that increasing sequencing lengths can improve assemblies and variance detection. We compared the SV detection metrics for five datasets of Nanopore with median read lengths of 15, 25, 40, 75, and 100 kbp. We observed an increase in recall with increasing read lengths for all tools except NanoVar, with the most pronounced improvement from median lengths of 15k to 25k. pbsv shows a modest rise in recall of 2% between 15k and 100k lengths, while Sniffles, cuteSV, SVIM, NanoSV and combiSV show an increase between 6 and 8%. Both pbsv and NanoVar show a significant drop in precision for read lengths of 75 and 100 kbp. As pbsv is specifically designed for shorter PacBio reads, it could be an explanation for this drop in precision. NanoVar is the only tool that has no benefit from longer read lengths, as we observed a drop in both recall and precision for median read lengths of 75 and 100 kbp (Fig. 3). All metrics of this comparison can be found in Additional file 2.

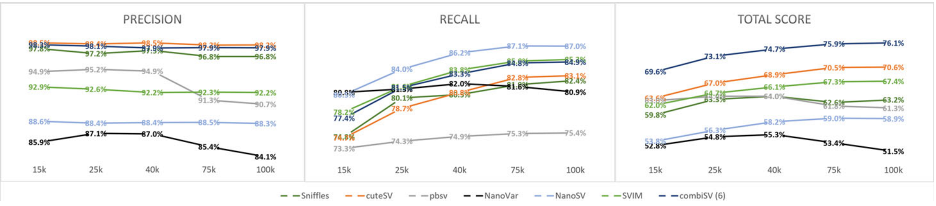


Fig. 3 Structural variance detection stats for a series of Nanopore datasets with increasing sequencing lengths

Structural variance detection using real datasets

We based our simulated datasets on a SV call set of NA19240 (nstd152), which was obtained through an elaborated SV study that combined a wide range of sequencing data [3]. To compare our simulation to the original genome, we performed the same benchmark on a public available PacBio CLR dataset of that study. Recall and precision values of the real dataset were significantly lower, with an average of respectively 65% and 50%. An even more striking difference were the recall percentages of around 60% for complex substitutions, while these values ranged between 1 and 20% for the simulated datasets, independent from sequencing platform or sequencing depth. While the overall lower recall and precision values were to be expected due to inaccuracies of the real SV call set, we found the large rise in recall for complex substitutions questionable. We therefore examined several alignments of SVs that were typed as complex substitutions. We found that most of these complex substitutions are in fact insertions or deletions, which would explain the high recall values. Most of the complex substitutions in nstd152 were determined by merging of experiments (optical mapping, sequence alignment and de novo assembly) and not associated to just one method. It is possible that conflicting findings between methods were thought to be caused by complex substitutions as they consist of both a deletion and an insertion. We added some concrete examples with screenshots of alignments and BLAST results of individual reads in Additional file 1: Figs. S7-S38 as evidence of these findings.

The large discrepancy between our simulated dataset and the real dataset are an indication that the SV call set of NA19240 (nstd152) has not been called accurately. The fact that the real NA19240 dataset has much lower precision and recall values than the real GIAB SV call set makes it unlikely that the discrepancy between the real and simulated SV call sets of NA19240 is caused by an inaccurate simulation. As additional validation of our simulation accuracy, we also simulated the GIAB SV callset and called SVs with Sniffles, SVIM and cuteSV. Both the recall and precision values of the simulated and real data were within a 2-5% range from each other.

Improved SV calling with combiSV

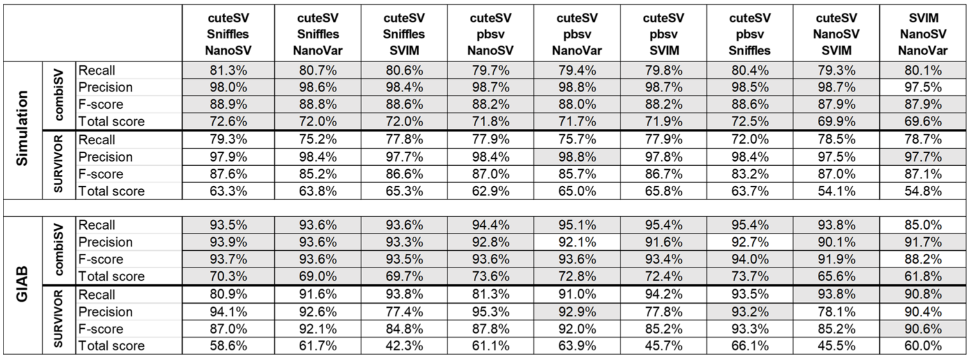
This benchmark revealed the strengths and weaknesses of each SV calling tool for long read sequencing. With this performance data we were able to develop a tool (combiSV) that can combine the outputs of cuteSV, pbsv, Sniffles, NanoVar, NanoSV, and SVIM into a superior SV call set, with Sniffles, pbsv or cuteSV as mandatory input (it can run without, but not recommended). The VCF outputs of each tool serve as input and the minimum count of supported reads for the variance allele has to be given. The complete wall time is under 1 minute and less than 1 GB of virtual memory is required. By combining the strengths of each of the 6 SV callers, we were able to eliminate distinct weaknesses and improve overall performance (Table 2). The most significant improvements were the ratio of total matches versus false positives and the accurate definement of the SV parameters. The added value of combiSV can also be seen by the sequence depth analysis (Fig. 2), where combiSV has consistently the best overall performance and does not show any significant drops in recall or precision for any of the sequencing depths. The improved performance of combiSV is less pronounced by the precision and recall values of the individual SV types, which can be explained by the fact that the performance gain was mostly limited for deletions and insertions. Most importantly, combiSV also showed significant improvement for the real GIAB dataset, as it combines the highest recall and precision from all tools, together with the accuracy from pbsv. This high recall is also achieved without NanoSV, as combiSV(3) only combines pbsv, sniffles and cuteSV. The combination of all 6 callers reduced the recall and precision slightly, which is probably caused by the high number of false positives of NanoSV and NanoVar. Therefore, it is not necessary to include the output of all 6 SV callers to run combiSV, although it is advised to add two additional callers besides cuteSV, pbsv or Sniffles. Despite the fact that combiSV takes less than one minute to run, total runtime will increase because multiple SV callers are being used. To have an idea how this will affect the total runtime, we performed a system requirements benchmark for each SV caller (Additional file 1: Table S5).

Currently, there are no similar tools as combiSV available for long sequencing reads. Therefore, we limited our comparison of combiSV to SURVIVOR, a tool that combines VCF files based on overlap. When combining the output of the 6 SV callers, combiSV produced a higher *F*-score and Total score for each of the 4 datasets in Table 2. Combining 6 tools requires additional effort and computational time, we therefore produced an additional benchmark where we tested 9 combinations of 3 SV callers for combiSV and SURVIVOR on the simulated Nanopore (20x) and the GIAB datasets (Table 4 and Additional file 1: Table S6). For the simulated dataset, combiSV achieved a higher F-score and Total score for each combination. For the GIAB dataset, combiSV had a higher Total score for all combinations and a higher *F*-score for 7 out of 8 combinations. The only combination with a higher F-score for SURVIVOR was SVIM, NanoSV and NanoVar. With a recall of 85%, combiSV performs significantly less on this combination, as other combinations have recalls above 93.5%. This is because combiSV is designed to have at least cuteSV, Sniffles or pbsv in the combination, while SVIM, NanoSV, and NanoVar can be added as support.

Discussion

We developed a realistic simulated model to benchmark existing structural variation detection tools for long-read sequencing. This was accomplished with Sim-it, a newly developed tool for the simulation of structural variation and long sequencing reads. Although there are several tools available that can simulate structural variation or long sequencing reads, a benchmark study to assess the accuracy of these simulators was

Table 4 Comparison between combiSV and SURVIVOR for 9 combinations of three SV callers on a simulated Nanopore dataset of 20x and the GIAB reference dataset (Nanopore). The highest scores between combiSV and SURVIVOR are indicated in gray



			cuteSV Sniffles NanoSV	cuteSV Sniffles NanoVar	cuteSV Sniffles SVIM	cuteSV pbsv NanoSV	cuteSV pbsv NanoVar	cuteSV pbsv SVIM	cuteSV pbsv Sniffles	cuteSV NanoSV SVIM	SVIM NanoSV NanoVar
Simulation	combiSV	Recall	81.3%	80.7%	80.6%	79.7%	79.4%	79.8%	80.4%	79.3%	80.1%
		Precision	98.0%	98.6%	98.4%	98.7%	98.8%	98.7%	98.5%	98.7%	97.5%
		F-score	88.9%	88.8%	88.6%	88.2%	88.0%	88.2%	88.6%	87.9%	87.9%
		Total score	72.6%	72.0%	72.0%	71.8%	71.7%	71.9%	72.5%	69.9%	69.6%
	SURVIVOR	Recall	79.3%	75.2%	77.8%	77.9%	75.7%	77.9%	72.0%	78.5%	78.7%
		Precision	97.9%	98.4%	97.7%	98.4%	98.8%	97.8%	98.4%	97.5%	97.7%
		F-score	87.6%	85.2%	86.6%	87.0%	85.7%	86.7%	83.2%	87.0%	87.1%
		Total score	63.3%	63.8%	65.3%	62.9%	65.0%	65.8%	63.7%	54.1%	54.8%
GIAB	combiSV	Recall	93.5%	93.6%	93.6%	94.4%	95.1%	95.4%	95.4%	93.8%	85.0%
		Precision	93.9%	93.6%	93.3%	92.8%	92.1%	91.6%	92.7%	90.1%	91.7%
		F-score	93.7%	93.6%	93.5%	93.6%	93.6%	93.4%	94.0%	91.9%	88.2%
		Total score	70.3%	69.0%	69.7%	73.6%	72.8%	72.4%	73.7%	65.6%	61.8%
	SURVIVOR	Recall	80.9%	91.6%	93.8%	81.3%	91.0%	94.2%	93.5%	93.8%	90.8%
		Precision	94.1%	92.6%	77.4%	95.3%	92.9%	77.8%	93.2%	78.1%	90.4%
		F-score	87.0%	92.1%	84.8%	87.8%	92.0%	85.2%	93.3%	85.2%	90.6%
		Total score	58.6%	61.7%	42.3%	61.1%	63.9%	45.7%	66.1%	45.5%	60.0%

needed. Besides Sim-it, the combination of Varsim and LongISLND (despite the aberration for the length of deletions) could also have been used for this benchmark study. We simulated in total 5 PacBio and 8 Nanopore whole genome sequencing datasets of GRCh38 with coverages ranging between 10x and 50x and lengths between 15 and 100 kbp. With these simulations, we assessed the performance of 7 SV callers and the influence of increasing sequencing depths and read lengths.

For most datasets, cuteSV, or SVIM produced the best overall performance with a good balance between recall and precision. cuteSV has the highest precision across all datasets, yet performs significantly less for PacBio HiFi datasets with a coverage below 30x. pbsv defines the SVs the most accurate across all datasets and since it is designed for PacBio, it performs the best on this type of data. NanoSV and NanoVar have high recall numbers, however at the cost of a disproportional high false positive rate (to a lesser extent for PacBio HiFi data). We found similar patterns for the high-fidelity SV call set of GIAB, although with some distinct differences. The real GIAB dataset had for each of the tools a higher recall and a lower precision compared to the simulations. The higher recall can be explained by the fact that the GIAB call set only contains insertions and deletions in non-complex regions, which are easier to call than other types of SVs or SVs in complex regions. Similar lower precision values from a simulation of the SV call set of GIAB suggests that these lower values are sample-specific and not caused by inaccurate simulations. The high precision values for the simulations of the 24,600 SVs could be misleading, as the SV callset of the Chaisson et al. study possibly contains a significant number of false positives that we simulated as true SVs.

Recall and precision values are the preferred metrics to measure SV detection accuracy. Called SVs are scored as 0 or 1, based on a reference SV set. However, there is no consensus about how accurate the position, length or SV type has to be called to be matched with the reference set. Therefore, we chose a tolerant matching algorithm and included a total score that integrates the accuracy of the call, which is not integrated in the *F*-score. The downside of our total score is that some SV callers that do not call SV types or genotypes are too heavily punished. In addition, we added separate accuracy scores for position, length, type and genotype.

It is often assumed that higher sequencing depths and longer read lengths will improve assembly and variance calling outcomes. Yet in our benchmark, increasing sequencing depths does not guarantee improved structural variation calling. Although there was still a modest rise in recall numbers for sequencing depths above 30x, we did observe a disproportional rise in false positives above 30x. This rise in false positives was not observed for increasing sequencing lengths, while we observed an increase in recall for longer read lengths across all methods.

Finally, we looked at precision and recall rates for each type of SV. Each tool showed the best performance for deletions and insertions, which are the majority of SVs in a human genome. More problematic SVs are inversions and complex substitutions, wherefore recall rates are respectively between 50–68% and 5–25%. As complex substitutions are not defined by any of the tools, it seems likely that these algorithms are not designed to detect this type of SV. New SV callers or updates of existing ones could make significant improvements in this direction. Although the SV study we used as blueprint [3] detected around 3000 complex substitutions per individual, we discovered that most of these complex substitutions were insertions or deletions. The actual prevalence of this type of structural variation is therefore possibly not accurate and requires further studies in order to map the complete structural variation profile in the human genome.

This study shows that a simulated model can be beneficial to gain a better understanding in the performance of structural variation detection tools. The development of combiSV was solely based on simulated datasets, but the recall and precision values from combiSV for the GIAB dataset shows that the statistics from the simulated data is transferable to real datasets. It is crucial that the simulations are as accurate as possible. Currently, Sim-it does not simulate small indels and SNPs, although they can influence the detection of small SVs and will therefore be included in the next update.

Conclusions

This extensive benchmark unveiled the strengths and weaknesses of each SV detection algorithm and provided the blueprint for the integration of multiple algorithms in a new SV detection pipeline, namely combiSV. This Perl script can combine the VCF outputs from cuteSV, Sniffles, pbsv, NanoVar, NanoSV, and SVIM into a superior call set that has the high precision of cuteSV, the accuracy of pbsv and the high recall of SVIM. combiSV also achieves a higher recall, precision and accuracy compared to SURVIVOR, an existing algorithm to generate a consensus VCF. The added value of combiSV on simulated data was supported by the real dataset of GIAB, where the gains were even more outspoken, which demonstrates the strengths of an accurate simulated model for the development of new bioinformatic tools.

Methods

Sim-it

We developed a new structural variation and long-read sequencing simulator, called Sim-it. The structural variation module outputs fasta files of each haplotype, plus an additional one that combines all SVs in one sequence. A set list of SVs can be combined with additional random generated SVs as input. The long-read sequencing module outputs sequencing reads based on a given error profile and 4 metrics (coverage, median length, length range and accuracy). We provide error profiles for Nanopore, PacBio RS II, Sequel II, and Sequel HiFi reads. Additional error profiles can be generated with a custom script. Both simulation modules (SV and long reads) can be used separately or simultaneously, starting from a sequence file as input. We also provide plots with the length distributions for the simulated sequencing reads and structural variations (insertions, deletions and inversions). Sim-it was written in Perl and does not require any further dependencies. Sim-it is open source and can be downloaded at https://github.com/ndierckx/Sim-it, where a more complete manual can be found.

Benchmark of structural variation simulators

We compared Sim-it (v1.2) with RSVSim (v1.24.0) [18], SVEngine (v1.0.0) [19], VISOR (v1.1) [20], and VarSim (v0.8.4) [15] for computing resource consumption and available features. Runtime performance was measured using the Unix time command and Snakemake (v5.7.0) [21] benchmark function on the custom VCF of 24,600 SVs. We did not evaluate the performance of SCNVSim [10] and SURVIVOR [22] because they do not accept a custom VCF file. All scripts were executed on a Xeon E7-4820 with 512 GB of memory.

Benchmark of the long-read simulators

We compared Sim-it (v1.0) with the long-read simulators PBSIM (v1.0.4) [23], Badread (v0.1.5) [24], PaSS [25], LongISLND (v0.9.5) [26], DeepSimulator (v1.5) [27], Simlord (v1.0.3) [28], SURVIVOR (v1.0.7) [22], and NanoSim (v2.6.0) [29] for computing resource consumption and error frequency within context-specific patterns for mismatches and indels using real data of Nanopore and PacBio sequencing. Runtime performance was measured using the Unix time command and Snakemake (v5.7.0) benchmark function on the 15x sequencing coverage simulation with chromosome 1 of GRCh38. Context-specific error patterns were analyzed by a custom perl script with alignment 30x simulated read to 60 Kbp sequence. All scripts were executed on a Xeon E7-4820 with 512GB of memory. More details on the error profiles used for each simulation can be found in Additional file 1: Section 2.

Train customized error profiles for Sim-it

Error profiles were trained by a customized script, which aligns each read individually to the assembled reference sequence with BLAST [30]. For each kmer of 3 bp, the error rates of substitutions, insertions and deletions of the middle nucleotide were determined, along with the length patterns of deletions and insertions. The *E. coli* K12 substrain MG1655 dataset of PacBio Sequel II and PacBio RS II was downloaded from the github website of Pacific Biosciences. Using the above two datasets, we trained the error profile of PacBio Sequel II and PacBio RS II. We also downloaded the GIAB HG002 dataset of PacBio Sequel II HiFi reads powered by CCS. To reduce the computational time, we trained the error profile of PacBio Sequel II HiFi reads based on chromosome 1 of GRCh38. The Nanopore error profile is based on sequencing reads for chromosome 1 of GRCh38 from the publicly available GIAB HG002 dataset GM24385.

SV detection on simulated reads

We used the simulated data from Sim-it to validate 6 structural variant callers, namely Sniffles (v1.0.11) [1], SVIM (v1.3.1) [31], NanoSV (v1.2.4) [32], Picky (v0.2.a) [33], NanoVar (v1.3.8) [34], cuteSV (v1.0.10) [35], and pbsv (v2.3.0). A list of 24,600 SVs, derived from sample NA19240 of dbVAR nstd152 [3], was used to simulate Nanopore, PacBio CLR reads and PacBio HiFi reads for GRCh38 at a sequencing depth of 20x. This set of SVs consists out of 10,469 insertions, 10,031 deletions, 857 duplications, 170 inversions and 3073 complex substitutions. We also simulated 20x normal read using GRCh38 with not structural variants at all. Besides for pbsv, we aligned the simulated reads to GRCh38 using Minimap2 (v2.17-r941) [36]. The alignment for pbsv was performed using pbmm2 (v1.3.0) with default parameters. The exact parameters that were used for the alignments and SV callers can be found in Additional file 1: Section 1. Besides the six SV callers, we also included SURVIVOR to the benchmark. This tool combines VCF files by merging overlapping SVs.

Furthermore, we simulated additional Nanopore and PacBio HiFi reads for GRCh38 at sequencing depths of 10x, 30x and 50x to study the influence of increasing sequencing depths for SV calling. Each of the Nanopore simulations had a median read length of 25 kbp, we also included four additional simulations of 15 kbp, 40 kbp, 75 kbp, and 100 kbp with a sequencing depth of 20x. PacBio long reads have a median length of 25,000 bp and the PacBio HiFi reads a median length of 15,000 bp. An additional filtering step was added for each VCF output; we only retained variances that obtained a PASS for the FILTER value, that have a length of 50 bp or more and wherefore at least 3 (for sequencing depths 10x and 20x) or 5 (for sequencing depths 30x and 50x) reads support the variance. This additional filtering step significantly improved the output for each tool compared to the raw VCF output.

Benchmark metrics were calculated by comparing the VCF output of each SV caller against the simulated reference set of 24,600 SVs. For each detected SV, we looked for possible matches in the reference set within a 1600-bp range of the detected position. When the length of the SV was determined, we tolerated an error margin of 35% for SVs longer than 300 bp and no error margin for shorter SVs. If these two conditions were met, a detected SV was matched to the SV of the reference set, independent from the type or genotype that was called. Based on these paired SVs, recall, precision and the F-score (2*((precision*recall)/(precision+recall))) were calculated. As there are multiple metrics that define the performance of an SV detection algorithm, we adopted an overall score that combines each of the metrics. For each detected SV, a maximal score of 1 was possible: 0.4 for the correct position, 0.2 for the correct length, 0.2 for the correct type of SV, and 0.2 for the correct genotype. The scores for length and position proportionally decreased with difference compared to the reference set. Finally, the number of false positives were subtracted from the total score and eventually expressed as a percentage of the maximum possible score (Table 2).

SV detection on real datasets

The Genome in a Bottle (GIAB) Consortium recently developed a high-quality SV call set for the son (HG002/NA24385) of a broadly consented and available Ashkenazi Jewish trio from the Personal Genome Project. We performed a benchmark on the latest most conserved BED file (HG002_*SVs*_Tier1_v0.6.2.bed) for this sample, which contains 5260 insertions and 4138 deletions. The public available ultralong Nanopore reads (GM24385) with an average sequencing depth of 45x were used for this benchmark. This GIAB dataset was also simulated to estimate the impact of simulated versus real reads on recall and precision values. Furthermore, we compared SV detection metrics of a public available PacBio dataset of NA19240 [3] with an average sequencing depth of 37x against the results of our simulated datasets.

combiSV

With the results of the SV detection benchmark, we developed a script to combine the results of cuteSV, pbsv, Sniffles, NanoVar, NanoSV and SVIM. The output VCF files of each of the 6 tools serve as input, from which the files of cuteSV or Sniffles are obligatory to run combiSV. For each SV detection tool, we examined the connections between the false positive rates and the accuracy of the stats (position, SV type and genotype) to each type of SV and genotype. When multiple callers detect the same SV, for each stat one SV caller will be prioritized based on the statistical analysis of the simulated benchmark. pbsv wil for example be prioritized for position and length stats, while cuteSV for the genotype. To further improve recall percentages, specific types of SVs that exhibited low false positive rates in the benchmark will be included in the final SV set, e.g., homozygous SVs from SVIM or heterozygous insertions and deletions from Sniffles. The number of callers needed to confirm an SV is set for 2 when 2 to 5 callers are combined and 3 when

6 callers are combined. This number can be adjusted manually, with also an additional option to exclude the calls that were only supported by one caller. The minimal coverage of the alternative allele is set to 3 as default value, but can be adjusted for datasets with high sequencing depths. The script was written in Perl and does not require any further dependencies. combiSV is open source and can be downloaded at https://github.com/ndierckx/combiSV.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02551-4.

Additional file 1: A PDF file with additional figures and tables. The file contains the following subsections: "Parameters for the structural variation callers", "Long read simulators benchmark", "Error profiles of long read simulators" and "Complex substitutions in NA19240".

Additional file 2: An Excel spreadsheet with 14 tables that contain additional metrics for the SV detection benchmark.

Additional file 3: Review history.

Peer review information

Andrew Cosgrove and Barbara Cheifet were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 3.

Authors' contributions

ND conceived, designed and scripted Sim-it and combiSV. ND analyzed the data for the SV caller benchmark. TL ran the simulations and existing SV calling tools for the benchmark. TL and ND designed and executed the SV and long-read simulator benchmark. ND wrote the manuscript. JRV and ZX provided guidance and reviewed the manuscript. The authors read and approved the final manuscript.

Funding

This project was supported by FWO TBM project T003819N to J.R.V. We thank the Center for Precision Medicine at Sun Yat-sen University for providing the high performance computers. This project was supported by National Key R&D Program of China (2019YFA0904401 to Z.X.)

Availability of data and materials

The list of simulated SVs and both Sim-it and combiSV can be found on github: https://github.com/ndierckx/ [37, 38] under the Apache License 2.0 and on Zenodo [39, 40] under the Creative Commons Attribution 4.0 International License. The E. coli K12 substrain MG1655 dataset of PacBio Sequel II was downloaded from https://github.com/PacificBiosciences/DevNet/wiki/Microbial-Multiplexing-Data-Set---48-plex:-PacBio-Sequel-II-System,-Chemistry-v2.0,-SMRT-Link-v8.0-Analysis. The E. coli K12 substrain MG1655 dataset of PacBio RS II was downloaded from https://github.com/PacificBiosciences/DevNet/wiki/E.-coli-Bacterial-Assembly. The GIAB HG002 dataset of PacBio Sequel II HiFi reads powered by CCS was downloaded from https://github.com/PacificBiosciences/DevNet/wiki/Sequel-II-System-Data-Release:-HG002-SV-and-SNVs-(HiFi-Reads-powered-by-CCS). The publicly available GIAB HG002 dataset GM24385 was downloaded from https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/. The structural variation simulation of the 24,600 SVs was derived from sample NA19240 of dbVAR nstd152: https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd152/. The GIAB BED file which contains 5,260 insertions and 4,138 deletions: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/. The PacBio dataset of NA19240: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/working/20160905_smithm_pacbio_aligns/NA19240_bwamem_GRCh38DH_YRI_20160905_pacbio.bam.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 21 November 2020 Accepted: 22 November 2021
Published online: 15 December 2021

References

1. Sedlazeck F, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz M. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8. https://doi.org/10.1038/s41592-018-0001-7.

2. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14(5):305–14. https://doi.org/10.1093/bfgp/elv014.
3. Chaisson M, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784. https://doi.org/10.1038/s41467-018-08148-z.
4. Sudmant P, et al. An integrated map of structural variation in 2,504 Human genomes. Nature. 2015;526(7571):75–81. https://doi.org/10.1038/nature15394.
5. Chen S, Krusche P, Dolzhenko E, Sherman R, Petrovski R, Schlesinger F, Kirsche M, Bentley D, Schatz M, Sedlazeck F, Eberle M. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20(1):291. https://doi.org/10.1186/s13059-019-1909-7.
6. Wenger A, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62. https://doi.org/10.1038/s41587-019-0217-9.
7. Jain M, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45. https://doi.org/10.1038/nbt.4060.
8. Ardui S, Ameur A, Vermeesch J, Hestand M. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46(5):2159–68. https://doi.org/10.1093/nar/gky066.
9. Jain M, Olsen H, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community [published correction appears in Genome Biol. 2016 Dec 13;17 (1):256]. Genome Biol. 2016;17(1):239. https://doi.org/10.1186/s13059-016-1103-060.
10. Brown C, Clarke J. Nanopore development at Oxford Nanopore. Nat Biotechnol. 2016;34(8):810–1. https://doi.org/10.1038/nbt.3622.
11. Audano P, Sulovari A, Graves-Lindsay T, Cantsilieris S, Sorensen M, Welch A, Dougherty M, Nelson B, Shah A, Dutcher S, Warren W, Magrini V, McGrath S, Li Y, Wilson R, Eichler E. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176(3):663–675.e19. https://doi.org/10.1016/j.cell.2018.12.019.
12. Zook J, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;10. https://doi.org/10.1038/s41587-020-0538-8.
13. Qin M, Liu B, Conroy J, Morrison C, Hu Q, Cheng Y, Murakami M, Odunsi A, Johnson C, Wei L, Liu S, Wang J. SCNVSim: somatic copy number variation and structure variation simulator. BMC Bioinforma. 2015;16(1):66. https://doi.org/10.1186/s12859-015-0502-7.
14. Mu J, Mohiyuddin M, Li J, Bani Asadi N, Gerstein M, Abyzov A, Wong W, Lam H. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics. 2015;31(9):1469–71. https://doi.org/10.1093/bioinformatics/btu828.
15. Hermetz K, Newman S, Conneely K, Martin C, Ballif B, Shaffer L, Cody J, Rudd M. Large inverted duplications in the human genome form via a fold-back mechanism. PLoS Genet. 2014;10:e1004139. https://doi.org/10.1371/journal.pgen.1004139.
16. Williams T, Kelley C. Gnuplot 4.5: an interactive plotting program. 2011. http://gnuplot.info. Accessed 3 Oct 2020.
17. Danecek P, Bonfield J, Liddle J, Marshall J, Ohan V, Pollard M, Whitwham A, Keane T, McCarthy S, Davies R, Li H. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008. https://doi.org/10.1093/gigascience/giab008.
18. Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics. 2013;29(13):1679–81. https://doi.org/10.1093/bioinformatics/btt198.
19. Xia L, Ai D, Lee H, Andor N, Li C, Zhang N, Ji H. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. GigaScience. 2018;7(7):giy081. https://doi.org/10.1093/gigascience/giy081.
20. Bolognini D, Sanders A, Korbel J, Magi A, Benes V, Rausch T. VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. Bioinformatics. 2019;36(4):1267–1269. https://doi.org/10.1093/bioinformatics/btz719.
21. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018;34(20):3600. https://doi.org/10.1093/bioinformatics/bty350.
22. Jeffares D, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck F. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8(14061):1–11.
23. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator–toward accurate genome assembly. Bioinformatics. 2013;29(1):119–21. https://doi.org/10.1093/bioinformatics/bts649.
24. Wick R. Badread: simulation of error-prone long reads. J Open Source Softw. 2019;4(36):1316. https://doi.org/10.21105/joss.01316.
25. Zhang W, Jia B, Wei C. PaSS: a sequencing simulator for PacBio sequencing. BMC Bioinforma. 2019;20(1):352. https://doi.org/10.1186/s12859-019-2901-7.
26. Lau B, Mohiyuddin M, Mu J, Fang L, Bani Asadi N, Dallett C, Lam H. LongISLND: in silico sequencing of lengthy and noisy datatypes. Bioinformatics. 2016;32(24):3829–32. https://doi.org/10.1093/bioinformatics/btw602.
27. Li Y, Han R, Bi C, Li M, Wang S, Gao X. DeepSimulator: a deep simulator for Nanopore sequencing. Bioinformatics. 2018;34(17):2899–908. https://doi.org/10.1093/bioinformatics/bty223.
28. Stöcker B, Köster J, Rahmann S. SimLoRD: Simulation of Long Read Data. Bioinformatics. 2016;32(17):2704–2706. https://doi.org/10.1093/bioinformatics/btw286.
29. Yang C, Chu J, Warren R, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. GigaScience. 2017;36(4):1–6. https://doi.org/10.1093/gigascience/gix010.
30. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
31. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. 2019;35(17):2907–15. https://doi.org/10.1093/bioinformatics/btz041.
32. Cretu Stancu, M. vanR, Renkens I, Nieboer M, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski M, Marschall T, de Ridder J, Kloosterman W. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Commun. 2017;8(1):1326. https://doi.org/10.1038/s41467-017-01343-4.

33. Gong L, Wong C, Cheng W, Tjong H, Menghi F, Ngan C, Liu E, Wei C. Picky comprehensively detects high-resolution structural variants in nanopore long reads. Nat Methods. 2018;15(6):455–60. https://doi.org/10.1038/s41592-018-0002-6.
34. Tham C, Tirado-Magallanes R, Goh Y, Fullwood M, Koh B, Wang W, Ng C, Chng W, Thiery A, Tenen D, Benoukraf T. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. 2020;21(1):56. https://doi.org/10.1186/s13059-020-01968-7.
35. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21:189. https://doi.org/10.1186/s13059-020-02107-y.
36. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
37. Dierckxsens N. Sim-it: A structural variance and Nanopore/PacBio sequencing reads simulator. Github. 2021. https://github.com/ndierckx/Sim-it. Accessed 9 Nov 2021.
38. Dierckxsens N. combiSV. Github. 2021. https://github.com/ndierckx/combiSV. Accessed 9 Nov 2021.
39. Dierckxsens N. Sim-it: A structural variance and Nanopore/PacBio sequencing reads simulator. Zenodo. 2021. https://doi.org/10.5281/zenodo.5707600.
40. Dierckxsens N. combiSV. Zenodo. 2021. https://doi.org/10.5281/zenodo.5707574.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.