NanoVar: accurate characterization of patients’ genomic structural variants using low-depth nanopore sequencing

Cheng Yong Tham, Roberto Tirado-Magallanes, Yufen Goh, Melissa J. Fullwood, Bryan T.H. Koh, Wilson Wang, Chin Hin Ng, Wee Joo Chng, Alexandre Thiery, Daniel G. Tenen, Touati Benoukraf

Additional file 1
Figure S1: NanoVar SV characterization algorithm displayed as a decision tree. The algorithm consists of only conditional control statements coded in Python to analyse long-read alignment profiles for SV characterization.
Figure S2: SV simulation and read simulation information. (a) Kernel density plot of the distribution of simulated SV sizes across the three simulated datasets. (b) Read length and (c) indel percentage distribution of simulated long reads and whole-genome patient Nanopore sequencing reads.
Figure S3: Repetitive sequence analysis of SVs recalled in homozygous (Top) and heterozygous (bottom) simulation datasets by the different tools.
Figure S4: Recall benchmarking using NA12878 PacBio benchmark sample distributed by Parikh et al., 2016 (2676 DEL and 68 INS). The number of reads used had been downsized to 4x coverage (Top) and 8x coverage (Bottom).
Figure S5: Donut charts showing the distribution of SV classes characterized by NanoVar in Patient 1 and Patient 2. TLO: Translocation, TPO: Transposition.
Figure S6: Genome browser snapshots of Nanopore long reads and Illumina short reads at each SV breakpoint location for SVs in (a) Patient 1 and (b) Patient 2. The top one-third of each snapshot displays the long reads in blue rectangles, while the bottom two-thirds displays short reads in red and blue rectangles. The cumulative read coverages for long reads (blue) and short reads (grey) are displayed above all the respective reads. The Integrative Genomics Viewer (IGV, https://software.broadinstitute.org/software/igv/) was used for the visualization of reads.
Homozygous

| Tool       | Optimum threshold score | Recall Sim 1 | Recall Sim 2 | Recall Sim 3 | Recall Avg | Precision Sim 1 | Precision Sim 2 | Precision Sim 3 | Precision Avg | F1 score Sim 1 | F1 score Sim 2 | F1 score Sim 3 | F1 score Avg | Area under curve (AUC) Sim 1 | Area under curve (AUC) Sim 2 | Area under curve (AUC) Sim 3 | Area under curve (AUC) Avg |
|------------|-------------------------|--------------|--------------|--------------|------------|------------------|------------------|------------------|--------------|----------------|----------------|----------------|------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| NanoVar    | 1.0                     | 0.955        | 0.955        | 0.954        | 0.955      | 0.951           | 0.948           | 0.950           | 0.950        | 0.953          | 0.952          | 0.952          | 0.952     | 0.965                                        | 0.965                                        | 0.965                                        | 0.965                                        |
| NanoSV     | 0                       | 0.314        | 0.310        | 0.319        | 0.314      | 0.997           | 0.997           | 0.997           | 0.997        | 0.477          | 0.477          | 0.477          | 0.477     | 0.478                                        | 0.478                                        | 0.478                                        | 0.478                                        |
| Sniffles   | N/A                     | 0.832        | 0.836        | 0.832        | 0.833      | 0.897           | 0.895           | 0.896           | 0.896        | 0.863          | 0.864          | 0.864          | 0.864     | N/A                                           | N/A                                           | N/A                                           | N/A                                           |
| SVIM       | 0                       | 0.791        | 0.790        | 0.791        | 0.719      | 0.720           | 0.714           | 0.718           | 0.753        | 0.753          | 0.751          | 0.757          | 0.753     | N/A                                           | N/A                                           | N/A                                           | N/A                                           |
| Picky      | N/A                     | 0.588        | 0.583        | 0.583        | 0.584      | 0.023           | 0.023           | 0.023           | 0.023        | 0.044          | 0.044          | 0.044          | 0.044     | N/A                                           | N/A                                           | N/A                                           | N/A                                           |
| novoBreak  | 27.5                    | 0.616        | 0.630        | 0.606        | 0.618      | 0.965           | 0.949           | 0.963           | 0.959        | 0.752          | 0.757          | 0.746          | 0.752     | 0.654                                        | 0.666                                        | 0.649                                        | 0.650                                        |
| Delly      | N/A                     | 0.708        | 0.707        | 0.706        | 0.707      | 0.977           | 0.977           | 0.977           | 0.976        | 0.821          | 0.820          | 0.820          | 0.820     | N/A                                           | N/A                                           | N/A                                           | N/A                                           |

Heterozygous (Sim1)

| Tool       | Optimum threshold score | Recall 4x | Recall 8x | Recall 12x | Precision 4x | Precision 8x | Precision 12x | F1 score 4x | F1 score 8x | F1 score 12x | Area under curve (AUC) 4x | Area under curve (AUC) 8x | Area under curve (AUC) 12x | Area under curve (AUC) Avg |
|------------|-------------------------|-----------|-----------|------------|---------------|---------------|---------------|--------------|--------------|----------------|----------------------------|--------------------------|--------------------------|-----------------------------|
| NanoVar    | 1.0                     | 0.757     | 0.883     | 0.929      | 0.972         | 0.968         | 0.965         | 0.951        | 0.923        | 0.947         | 0.879                      | 0.980                    | 0.974                    |                             |
| NanoSV     | 0                       | 0.155     | 0.314     | 0.407      | 0.989         | 0.999         | 0.997         | 0.269        | 0.477        | 0.578         | 0.155                      | 0.313                    | 0.407                    |                             |
| Sniffles   | N/A                     | 0.579     | 0.832     | 0.894      | 0.988         | 0.848         | 0.792         | 0.704        | 0.840        | 0.840         | N/A                        | N/A                      | N/A                      |                             |
| SVIM       | 0                       | 0.641     | 0.793     | 0.848      | 0.674         | 0.571         | 0.483         | 0.657        | 0.663        | 0.616         | 0.586                      | 0.742                    | 0.804                    |                             |
| Picky      | N/A                     | 0.384     | 0.600     | 0.666      | 0.015         | 0.012         | 0.011         | 0.029        | 0.023        | 0.021         | N/A                        | N/A                      | N/A                      |                             |
| novoBreak  | 27.5                    | 0.585     | 0.977     | 0.732      | 0.625         |               |               |              |              |               |                            |                          |                          |                             |
| Delly      | N/A                     | 0.698     | 0.960     | 0.615      |               |               |               |              |              |               |                            |                          |                          |                             |

Table S1: Precision and recall values of Figure 2a (Top) and 2b (Bottom). Sim # refers to simulation dataset #, where # refers to the dataset number. Avg: Average.
**Table S2**: Precision and recall values for different SV classes characterized by different tools presented in Figure 2c (Top) and 2d (Bottom). Numbers in parentheses in the recall columns represent the number of true SV breakends recalled. SV class annotation accuracy is considered in this analysis. For BND total breakends, 10,000 breakends are points of transposition location, and 20,000 breakends are the left and right coordinates of genomic sequences that were inserted. DUP: tandem duplication, DEL: deletion, INS: insertion, BND: breakend, INV: inversion, TP: True positive, FP: False positive.

**Table S3**: Oxford Nanopore MinION sequencing details of Patient 1 and Patient 2.
### Table S4: Genomic coordinates of SVs characterized in Patient 1 and Patient 2.

| SV id | SV class | Chromosome | Coordinates          |
|-------|----------|------------|----------------------|
| 1-1   | Del      | 14         | 68528678-68528737    |
| 1-2   | Del      | 7          | 134665255-134665664  |
| 1-3   | Ins      | 18         | 40768187             |
| 1-4   | Ins      | 7          | 110844695            |
| 1-5   | Del      | 14         | 51911629-51912185    |
| 1-6   | Ins      | 14         | 53333449             |
| 1-7   | Del      | 4          | 79439473-79441569    |
| 1-8   | Ins      | 4          | 89221858             |
| 2-1   | Del      | 4          | 142108392-142108444  |
| 2-2   | Ins      | 4          | 137377851            |
| 2-3   | Del      | 5          | 79717888-79718187    |
| 2-4   | Dup      | 4          | 7834853-7835893      |
| 2-5   | Del      | 21         | 10475720-10476287    |
| 2-6   | Ins      | 10         | 35706138             |
| 2-7   | Dup      | 6          | 66974314-66974585    |
| 2-8   | Del      | 3          | 194894010-194894311  |

### Table S5: SV validation in AML patients and NanoVar recall status for PCR-discovered SVs. (a) Results of SV validation and shared-SV detectability by NanoVar in the respective patient samples. Y=Yes, N=No, n/a=Not applicable. (b) Summary results for Patient 1 and Patient 2.

#### (a) Patient 1 and Patient 2 validation results

| SV id | Validated | Shared-SV recalled | Patient 1 | Patient 2 |
|-------|-----------|--------------------|-----------|-----------|
| 1-1   | Y         | n/a                | n/a       | Y         |
| 1-2   | Y         | n/a                | n/a       | Y         |
| 1-3   | Y         | n/a                | n/a       | Y         |
| 1-4   | Y         | n/a                | n/a       | Y         |
| 1-5   | Y         | n/a                | n/a       | Not shared |
| 1-6   | Y         | n/a                | n/a       | Y         |
| 1-7   | Y         | n/a                | n/a       | Y         |
| 1-8   | Y         | n/a                | n/a       | Not shared |
| 2-1   | n/a       | Y                  | Y         | n/a       |
| 2-2   | n/a       | N                  | Y         | n/a       |
| 2-3   | n/a       | N                  | Y         | n/a       |
| 2-4   | n/a       | Y                  | Y         | n/a       |
| 2-5   | n/a       | Y                  | Y         | n/a       |
| 2-6   | n/a       | Y                  | Y         | n/a       |
| 2-7   | n/a       | Y                  | Y         | n/a       |
| 2-8   | n/a       | N                  | Y         | n/a       |

#### (b) Summary results

| Sample | SV validated | Shared-SV recalled |
|--------|--------------|--------------------|
|        | Individual   | Total              |
| Patient 1 | 8/8          | 16/16              |
| Patient 2 | 8/8          | 11/14              |

Y=Yes, N=No, n/a=Not applicable.
Table S6: Runtime and maximum memory usage consumed by the workflows of each tool using 24 threads for the SV characterization in Patient 1. For 3GS tools, 12 Gb of sequencing data was used, while for 2GS tools, 160 Gb of sequencing data was used. Data was collected using GNU Time. Please note that each tool workflow is inclusive of the read mapping step using their recommended aligner.

| Tool workflow* | CPU time (min) | Wall clock time (min) | Maximum resident set size (RAM in gigabytes) |
|----------------|----------------|-----------------------|---------------------------------------------|
| NanoVar        | 1188           | 196                   | 31.7                                        |
| Picky          | 16910          | 1194                  | 25.1                                        |
| Sniffles       | 13187          | 561                   | 18.9                                        |
| SVIM           | 12795          | 558                   | 18.9                                        |
| NanoSV         | 32076          | 5278                  | 21.5                                        |
| Delly          | 28281          | 3374                  | 33.2                                        |
| novoBreak      | 57157          | 3519                  | 45.0                                        |

*Comprise of sequence mapping and SV calling, using 24 threads

Table S7: GenBank accession number and name of viruses used for SV insertion simulation to mimic viral insertion events.

| GenBank accession no. | Virus name                  | GenBank accession no. | Virus name                  |
|-----------------------|-----------------------------|-----------------------|-----------------------------|
| NC_012959.1           | Human adenovirus 54         | NC_034618.1           | Human papillomavirus type 85 isolate 114B |
| NC_001460.1           | Human adenovirus A          | NC_004500.1           | Human papillomavirus type 92  |
| NC_011203.1           | Human adenovirus B1         | NC_005134.2           | Human papillomavirus type 96  |
| NC_011202.1           | Human adenovirus B2         | NC_001596.1           | Human papillomavirus type 9   |
| NC_001405.1           | Human adenovirus C          | NC_001401.2           | Adeno-associated virus 2      |
| NC_010956.1           | Human adenovirus D          | NC_001729.1           | Adeno-associated virus 3      |
| NC_003377.2           | Hepatitis B virus (strain ayw) | NC_006152.1           | Adeno-associated virus 5      |
| NC_001806.2           | Human herpesvirus 1 strain 17 | NC_018102.1           | MW polymavirus                |
| NC_001708.2           | Human herpesvirus 2 strain HG52 | NC_020106.1           | STL polymavirus strain MA138  |
| NC_001348.1           | Human herpesvirus 3         | NC_020890.1           | Human polymavirus 12 strain hu4103 |
| NC_007005.1           | Human herpesvirus 4         | NC_024118.1           | New Jersey polymavirus-2013 isolate NJ-PyV-2013 |
| NC_006273.2           | Human herpesvirus 5 strain Merlin | NC_001538.1           | BK polymavirus                |
| NC_001716.2           | Human herpesvirus 7         | NC_001699.1           | JC polymavirus                |
| NC_006333.1           | Human herpesvirus 8         | NC_009236.1           | KI polymavirus Stockholm 60   |
| NC_017944.1           | Human papillomavirus type 136 | NC_009539.1           | WU Polymavirus                |
| NC_017996.1           | Human papillomavirus type 140 | NC_010277.2           | Merkel cell polymavirus isolate R17b   |
| NC_021483.1           | Human papillomavirus type 154 isolate PV77 | NC_014406.1           | Human polymavirus 6          |
| NC_003378.1           | Human papillomavirus type 156 isolate GC01 | NC_014407.1           | Human polymavirus 7          |
| NC_001526.4           | Human papillomavirus type 16 | NC_014361.1           | Trichodysplasia spinulosa-associated polymavirus |
| NC_023981.1           | Human papillomavirus type 178 | NC_015150.1           | Human polymavirus 9          |
| NC_022095.1           | Human papillomavirus type 179 | NC_001069.1           | Simian virus 40               |
| NC_001357.1           | Human papillomavirus 18     | NC_022518.1           | Human endogenous retrovirus K113 |
| NC_001356.1           | Human papillomavirus 1      | NC_001802.1           | Human immunodeficiency virus 1 |
| NC_027528.1           | Human papillomavirus type 201 isolate HPV201 | NC_001722.1           | Human immunodeficiency virus 2 |
| NC_001352.1           | Human papillomavirus 2      | NC_001436.1           | Human T-lymphotropic virus 1  |
| NC_001591.1           | Human papillomavirus type 49 | NC_001488.1           | Human T-lymphotropic virus 2  |
| NC_001531.1           | Human papillomavirus type 5 | NC_001364.1           | Simian foamy virus           |
Table S8: Statistics of 3GS and 2GS reads and read mapping of real reads and simulated reads. All reads were mapped using Minimap2 aligner except for 2GS real reads which was aligned by BWA-MEM. Simulated reads were aligned to GRCh37 reference genome while real reads were aligned to GRCh38 reference genome. Statistics were calculated by SAMTools. The sequencing depth was estimated from the Lander/Waterman equation.

| Sample          | Number of reads | Read length (bp) | Total bases (x10^6) | Bases mapped (x10^6) | Percentage of bases mapped | Alignment error rate | Lengthwise genome coverage | Estimated sequencing depth |
|-----------------|-----------------|------------------|---------------------|----------------------|---------------------------|-----------------------|---------------------------|---------------------------|
| Simulation 1    | 2,080,000       | 6316             | 13,304              | 12,198               | 91.69%                    | 1.72E-01              | 90.55%                    | 3.95x                     |
| Simulation 2    | 2,080,000       | 6322             | 13,315              | 12,201               | 91.63%                    | 1.72E-01              | 90.57%                    | 3.95x                     |
| Simulation 3    | 2,080,000       | 6315             | 13,300              | 12,190               | 91.66%                    | 1.72E-01              | 90.58%                    | 3.95x                     |
| Patient 1       | 1,793,667       | 6271             | 12,434              | 11,308               | 90.94%                    | 1.60E-01              | 91.19%                    | 3.66x                     |
| Patient 2       | 1,939,432       | 6196             | 12,273              | 11,191               | 91.18%                    | 1.84E-01              | 91.01%                    | 3.62x                     |

Table S9: Primers used for SV validation in Patient 1 and Patient 2 shown in Figure 3, and nested primers used for PCR product reamplification for Sanger sequencing (SV 2-4).