Optimising accuracy and speed of structural variant calling using ONT WGS data: Application for analysis of tumour genomes

Timothy M. Freeman*, Adam Giess, Melanie Tanguy, Genomics England Research Consortium, Greg Elgar, and Alona Sosinsky

* Presenter, Genomics England Limited, UK

Abstract

Long-read sequencing with Oxford Nanopore Technology (ONT) offers potential advantages over the current Illumina short-read sequencing used for identifying clinically relevant variants in tumour genomes. A clinical pipeline incorporating ONT long reads may enable the identification of new variants in hard-to-map regions, as well as disentangling complex structural variants, while reducing false positive rates, sample prep complexity and sequencing runtime, to provide faster, more accurate genomic reports back to clinicians.

In order to realise these benefits, Genomics England is developing a clinical pipeline for analysing ONT-based whole genome sequencing for tumour and normal sample pairs. The pipeline outcomes include methylation profiling as well as detection of single-nucleotide variants, copy number variants (CNVs) and structural variants (SVs) with emphasis on clinically relevant cancer genes.

This poster focuses on our work optimising long-read SV calling with sarcoma, ovarian and haematological cancer cohorts. For each patient we have used two flow cells for the tumour sample and one flow cell for the germline. We have benchmarked five SV callers (cuteSV, delly, nanomonSV, sniffles, SAVANA) with respect to recall, specificity and runtime, developed a method of filtering out more than 90% of false positive SV calls, carried out testing to optimise command line parameters, and manually reviewed discrepancies in clinically relevant genes between ONT long reads, Illumina short reads and FISH test results for these patients. In addition, we provide a separate training data set of cancer samples within our Research Environment for collaboration with researchers and tool developers to improve existing algorithms for variant calling from ONT data and to further optimise our pipelines.

Results

Cohort and data for initial benchmarking

- 28 haematological (ALL and AML), 12 ovarian (HGSOC) & 24 sarcoma clinical cancer samples, with matched normal samples
- All tumours are sequenced on Illumina short-read HiSeq and on ONT PromethIONs. SVs in Illumina data set are identified by Manta and considered as a truth set for benchmarking
- In addition, haematological cancer samples have FISH test results for clinically relevant SVs

Methods

- Five SV callers for benchmarking (includes 2 separate versions of Sniffles)
- Running parameters are chosen to maximise SV recall without compromising specificity
- Potential discrepancies between Illumina short reads and ONT data are reviewed manually
- Endpoints: recall, specificity and runtime
- Feedback information on SV callers' limitations to developers to iteratively improve performance of updated versions

SV caller	Includes germline filtering?	Total runtime*, hrs (min/median/max)	Concordance with +ve FISH tests	Recall for SVs in clinically-relevant genes (95% CI)	Precision for SVs in clinically relevant genes
NanomonSV v0.4	YES	6/17/41	8/9	99 (97-99)	Low (<1 FP SV per sample)
Sniffles v1.0.11	NO germline SVs were subtracted from VCF	12/43/337	9/9	NA**	High (>10 FP SVs per sample)
Sniffles v2.2 B5	NO germline SVs were subtracted from VCF	12min/19min/56min (16 threads)	8/9	79 (72-86)	Moderate-high (<10 FP SVs per sample)
CuteSV v1.0.13	NO germline SVs were subtracted from VCF	1/4/5.5 (16 threads)	8/9	48 (40-57)	Moderate (>2 FP SVs per sample)
Delly v0.8.7	YES	23/38/45	8/9	NA**	Extremely high (>100 FP SVs per sample)
SAVANA v0.2.3	YES	36min/1h50min/3h (16 threads)	9/9	83 (75-91)	Low (<1 FP SV per sample)

*Samples were excluded from runtime analysis if more than 2 flow cells used for tumour sample or median read length below 5kb (8 samples). **Runtimes too long for SV caller benchmarking to progress to this stage of analysis.

Panel-of-normals filtering removes false positives

- Most long-read false positives are germline SVs that are not called in the germline sample due to lower coverage or are recurrent sequencing errors across samples
- Table of commonly occurring germline SVs (panel of normals) can be used to filter these FPs
- Sniffles v1 and v2 and cuteSV SVs called in 99 normal genomes, generating SV caller specific panels of normals (PONs)
- PON filtering removes more than 90% of FP SVs with 0.01 population frequency cut-off without filtering out any true clinically relevant SVs
- Specificity greatly improved without reducing recall

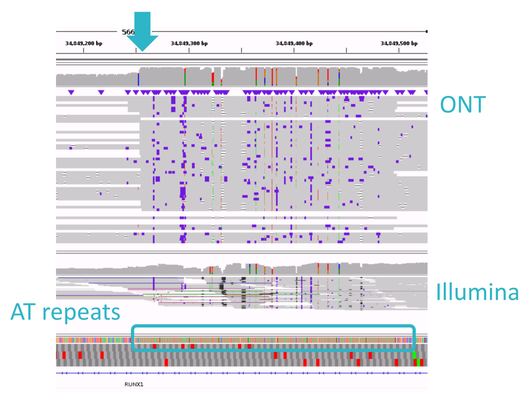
Examples

Examples of gain in the accuracy of SV calling over Illumina short reads (viewed in IGV):

Clinically relevant RUNX1-RUNX1T1 fusion in haematological cancer case,

Detectable with ONT long reads, missed by Illumina short reads

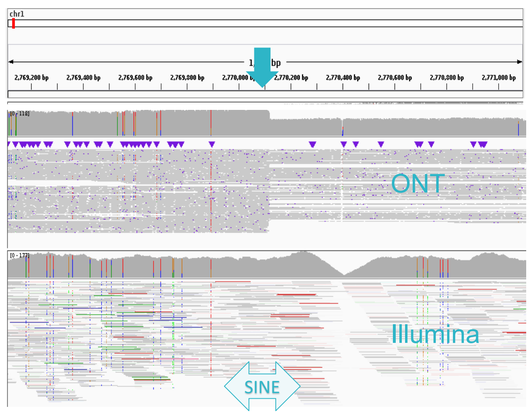
- chr8:92066805-chr21:34849248
- This fusion SV is called in long-read data with SAVANA, sniffles v1&2 and cuteSV
- Breakpoint on chr21 overlaps with simple 243bp $(AT)_n$ repeat region, which is too large for short reads to span across and map uniquely



Predicted in-frame TTC34-CAMTA1 inversion in sarcoma case

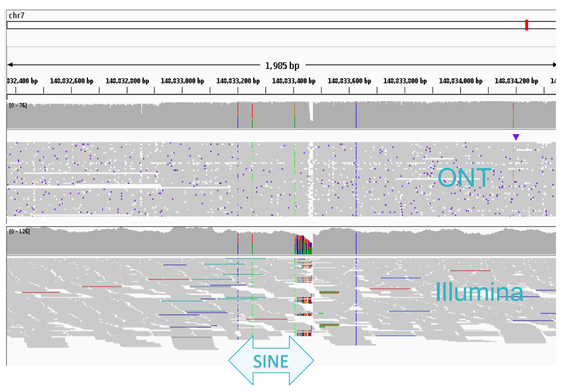
Detectable with ONT long reads, missed by Illumina short reads

- chr1:2770109-6975173
- This inversion SV is called with SAVANA and sniffles v2 and is predicted to create a productive fusion between the TTC34 and CAMTA1 genes
- Short reads do not map uniquely at the CAMTA1 breakpoint due to the overlap with SINE element that resembles multiple genomic regions



False-positive EZH2 inversion called by Illumina short reads

- chr7:145613082-148833369
- Short reads call false positive inversion due to reads misalignment with 279bp SINE repeat
- There are no ONT long reads supporting this inversion



Summary

Out of the SV callers tested, we were able to eliminate CuteSV from further consideration due to poor recall rates of clinically relevant SVs (especially inversions). Sniffles v1 and Delly were eliminated due to excessively long runtimes, which would be unsuitable for potential clinical use. This left SAVANA, NanomonSV and Sniffles v2, which had the highest recall values for clinically relevant SVs, of 83/99/79% respectively. NanomonSV displayed the best recall across cohorts, but was too slow for implementing into a clinical pipeline. SAVANA provided good recall, good runtimes and low false positive rates, while Sniffles v2 had the fastest runtimes. Runtimes could be further optimised with parallelisation for SAVANA and Sniffles v2, and specificity could be further improved with panel-of-normals filtering, SAVANA was the best overall performing candidate at the time of writing and the closest to achieving the requirements for implementation within a clinical pipeline.

SV calling algorithms using ONT data were able to identify clinically-relevant SVs missed by Illumina short reads at repetitive genomic regions where aligners could not uniquely map short reads overlapping breakpoints. There was a low sample size of clinically-relevant deletions, insertions and duplications for some cancer types, so future work will need to look at larger cohorts to confirm strong performance for these SV types. We have additional ONT-sequenced cancer cohorts available for further benchmarking. Sequencing data for >100 additional cancer patients with a diverse range of cancer types is shared in our Research Environment, so that external researchers and developers can use this resource to test and improve their tools. See https://www.genomicsengland.co.uk/research/academic for details of gaining access. We are carrying out similar work for benchmarking CNV, SNV and methylation calling tools on ONT cancer data and we keenly invite developers of these tools and others to use our Research Environment to continuously test and improve these tools.

More information: genomicsengland.co.uk
timothy.freeman@genomicsengland.co.uk

#nanoporeconf