Human structural variation identified by Oxford Nanopore PromethION

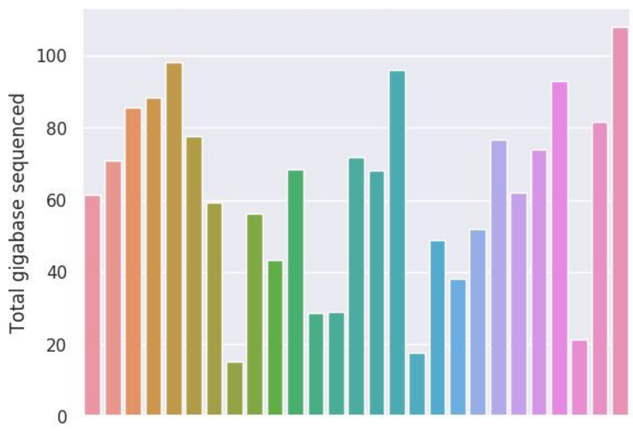
De Coster Wouter[a], De Roeck Arne[a], De Pooter Tim[b], D'Hert Svenn[b], De Rijk Peter[b], Strazisar Mojca[b], Sleegers Kristel[a] and Van Broeckhoven Christine[a]

[a] Neurodegenerative brain diseases group, Center for Molecular Neurology, VIB & University of Antwerp; [b] Neuromics Support Facility, Center for Molecular Neurology, VIB & University of Antwerp

Introduction

The majority of the structural variants in the genome, defined as changes in copy number or location of elements > 50 bp, remain hidden with currently dominant technologies. Long read sequencing has the advantage of a higher mappability, the ability to span breakpoints and align uniquely to repetitive sequences. For benchmarking and evaluation of tools we have sequenced the Yoruban reference genome NA19240, part of the HapMap and 1000 genomes project and well characterized using modern technologies[1], allowing independent validation of our findings. We reached 258 gigabase or 80x coverage and our data is publicly available on ENA (PRJEB26791).

Disclaimer: we have received consumables from ONT for this project and have been reimbursed for presenting at conferences.

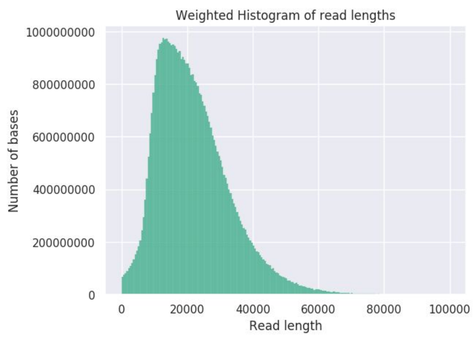


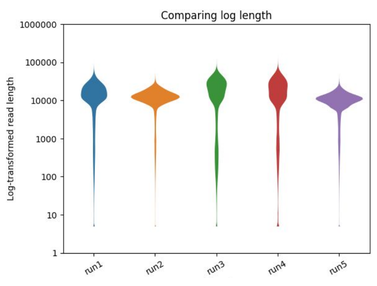
Methods

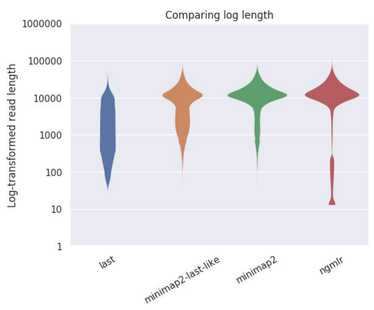
We have prepared libraries both using unsheared DNA or after shearing on the Megaruptor (Diagenode) to ~20kb. To remove small fragments from the library prior to end prep we use the BluePippin (Sage Science) in a High-Pass protocol, removing everything below a cut-off, which is based on the sizes of the DNA molecules as determined using the Fragment Analyzer (Agilent).

Structural variant sets were merged using SURVIVOR and compared and visualized using Python scripts. PromethION QC plots were generated using NanoPack[2].

Run	MegaRuptor Shearing	BluePippin Size selection	Yield [Gbase]
1	NA	> 10kb	59
2	20kb	> 8kb	69
3	NA	> 10kb	29
4	NA	> 10kb	29
5	20kb	> 6kb	72





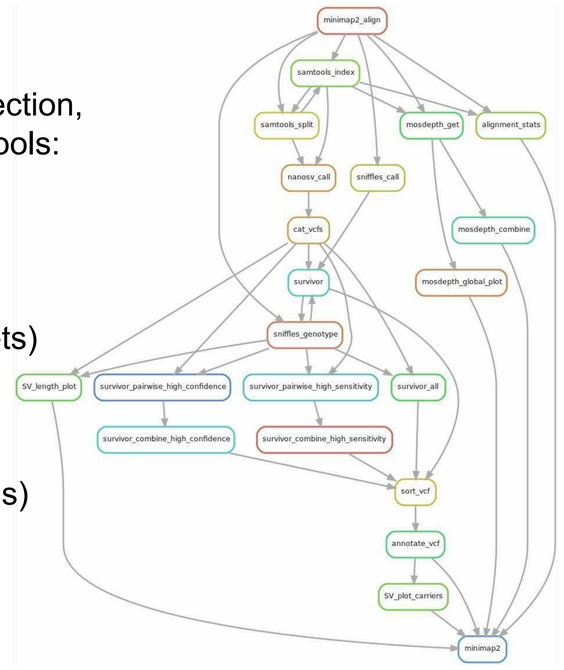


Structural variant detection workflow

We developed a Snakemake workflow for genome-wide detection, annotation and visualization of structural variants, integrating multiple tools:

- ngmlr and minimap2 (Alignment)
- Sniffles, NanoSV and npInv (Structural variant calling)
- mosdepth (Read depth calculation)
- SURVIVOR (Combining structural variants sets)
- vcfanno (Annotation)
- cyvcf2 (Parsing vcf files in Python)
- seaborn and matplotlib (Plotting)
- samtools, bcftools and vcftools (File format specific manipulations)

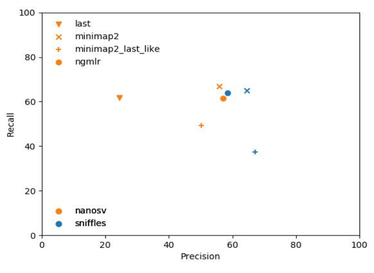
https://github.com/wdecoster/nano-snakemake

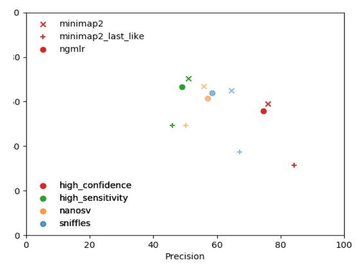


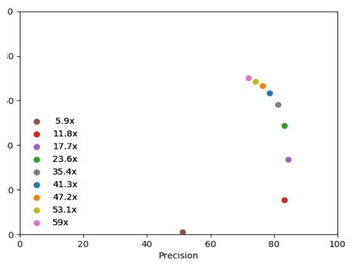
Results

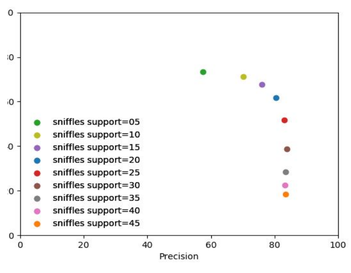
The PromethION sequencer generates, in our hands, consistently > 60 Gbase from a freshly extracted and sheared sample and up to 110 Gbase. Read lengths are comparable to the MinION, but since the latter system is already better understood the nucleotide level accuracy is currently slightly higher. We observe an inverse relationship between read length and yield. Alignment using minimap2 is the fastest, compared to LAST which is prohibitively slow for application in larger projects.

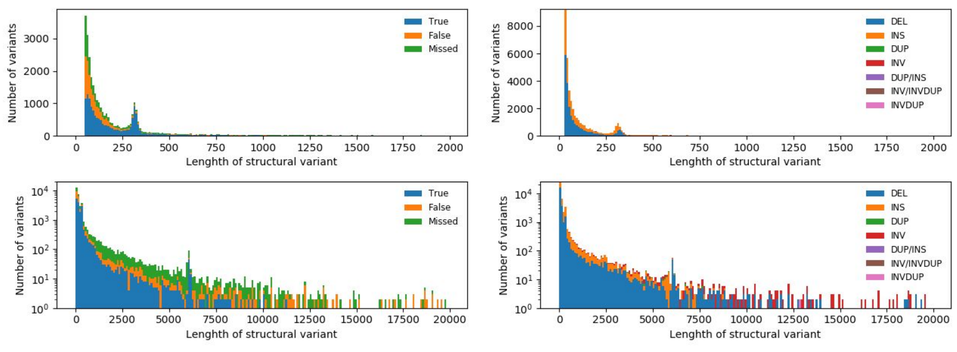
The number of structural variants identified ranged between 15821 and 123729, while our reference set contains 29436 variants. We evaluated combinations of aligners and structural variant callers and find that minimap2 followed by sniffles results in an optimal precision and recall. Another advantage is that those tools are the most computationally efficient both in terms of speed and memory usage. High confidence and high sensitivity call sets can be generated by taking respectively the union and intersection of variants identified by Sniffles and NanoSV. Accuracy of detection of inversions, with breakpoints often in highly repetitive regions, is generally poor, with best results obtained by npInv. By randomly downsampling the alignment we determined that relatively little is gained by increasing the coverage further above 40x. An optimal precision and recall is obtained by using a minimal number of supporting reads in Sniffles equal to ⅓ to ¼ of the genome coverage.











Results have been shared as a preprint:

De Coster et al. 2018. "Structural Variants Identified by Oxford Nanopore PromethION Sequencing of the Human Genome." *bioRxiv*. https://doi.org/10.1101/434118.

References

1 Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2017. "Multi-Platform Discovery Of Haplotype-Resolved Structural Variation In Human Genomes." bioRxiv.

2 De Coster, Wouter, Svenn D'Hert, Darrin T. Schultz, Marc Cruts, and Christine Van Broeckhoven. 2018. "NanoPack: Visualizing and Processing Long Read Sequencing Data." Bioinformatics

@wouter_decoster

wouter.decoster@uantwerpen.vib.be

https://github.com/wdecoster/

https://gigabaseorgigabyte.wordpress.com

bioRxiv THE PREPRINT SERVER FOR BIOLOGY