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

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 NanoPack2.

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

- ngmvr and minimap2 (Alignment)
- Sniffles, NanoSV and npInv (Structural variant calling)
- mosdepth (Read depth calculation)
- SURVIVOR (Combining structural variants sets)
- vcfanno (Annotation)
- cvcf2 (Parsing vcf files in Python)
- seaborn and matplotlib (Plotting)
- samtools, bsoftools and vcf tools (File format specific manipulations)

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 minmap2 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 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 minmap2 is the fastest, compared to LAST which is prohibitively slow for application in larger projects.

The results have been shared as a preprint:

References