Characterising structural variants with Oxford Nanopore

GETTING STARTED
SVs — encompassing insertions, deletions, duplications, inversions, and translocations — are a major source of genomic variation. In the human genome, they are responsible for ten times more variant bases than that of single nucleotide polymorphisms (SNPs) and are thirty times more likely to affect gene expression. Their analysis is of high importance for increasing our understanding of diseases including cancer, neurological disorders, and rare genetic diseases. In plants, SVs play a significant role in key traits, such as crop yield and climate resilience, making their identification important in agriculture and conservation efforts.

However, the size and genomic contexts of SVs have, until recently, limited their study. Sequencing via traditional short-read technology requires DNA to be fragmented to ~150–300 bp, meaning that SVs cannot typically be captured within a read. Accurate resolution of SVs from short-read data therefore represents a computational challenge; mis-mapping of short reads can even result in incorrectly called SVs.

The requirement for PCR in short-read sequencing means that SVs in regions that cannot be easily amplified — such as repetitive regions, in which they are frequently located — may be poorly represented in sequencing data, or missing entirely. Optical mapping provides an alternative method for locating SVs, but cannot provide sequence information and has limited resolution.

With nanopore sequencing, there is no upper read length limit: fragmentation is optional, enabling very long fragments of DNA to be sequenced intact. Reads routinely reach tens or hundreds of kilobases in length, with the current record spanning over 4 Mb. This means that many SVs can be sequenced end-to-end in single reads, allowing accurate characterisation and precise, nucleotide-level identification of breakpoint junctions. Larger SVs can be sequenced in far fewer, longer reads, with more overlap in between, greatly simplifying their assembly and reducing multimapping. Even highly complex SVs, such as those seen in chromothripsis, can be much more easily delineated.

The requirement for PCR in short-read sequencing means that SVs in regions that cannot be easily amplified — such as repetitive regions, in which they are frequently located — may be poorly represented in sequencing data, or missing entirely. Optical mapping provides an alternative method for locating SVs, but cannot provide sequence information and has limited resolution.

With nanopore sequencing, there is no upper read length limit: fragmentation is optional, enabling very long fragments of DNA to be sequenced intact. Reads routinely reach tens or hundreds of kilobases in length, with the current record spanning over 4 Mb. This means that many SVs can be sequenced end-to-end in single reads (Figure 1), allowing accurate characterisation and precise, nucleotide-level identification of breakpoint junctions. Larger SVs can be sequenced in far fewer, longer reads, with more overlap in between, greatly simplifying their assembly and reducing multimapping. Even highly complex SVs, such as those seen in chromothripsis, can be much more easily delineated.

**Figure 1: Integrative Genomics Viewer (IGV) visualisation of a phased heterozygous deletion**

This 1.8 kb deletion on chromosome 2 of the human genome reference sample HG002 was verified against the Genome in a Bottle SV truth set. Long nanopore sequencing reads were able to resolve the deletion, identify clear breakpoint junctions, and span the wildtype region end-to-end.

### References

Nanopore sequencing does not require amplification: native DNA can be sequenced directly, avoiding PCR bias and ensuring even coverage across the genome. This enables the analysis of SVs in any genomic context, including GC-rich and low-complexity areas. Epigenetic modifications remain intact, allowing SVs and their methylation state to be assessed simultaneously, without the need for any additional library preparation steps.

Comprehensive, genome-wide detection of SVs requires even, high depth of coverage. The PromethION™ platform provides the ultra-high throughput required for rapid, comprehensive detection of SVs at scale, even in very large genomes.

A single PromethION Flow Cell can generate sufficient data for >30x coverage of the human genome. This enables a whole-genome SV survey from a single sequencing run. With devices available for the utilisation of up to 24 or 48 individually-addressable flow cells — providing the capacity to generate terabases of sequencing data — the platform offers the flexibility to perform on-demand sequencing without any need to wait to batch samples. Sequencing can be scaled to meet the throughput and turnaround times needed: a single experiment can be performed across many flow cells or, where lower coverage is required or smaller targets are sequenced, many samples can be multiplexed on one flow cell.

The PromethION features powerful integrated compute, allowing real-time basecalling and analysis as a run progresses, without having to wait until the end of a sequencing run. Basecalling can begin as soon as sequencing starts, and a run can be stopped as soon as sufficient depth of coverage is achieved. The device is also available without capital cost, enabling access to PromethION sequencing with consumable costs only.

Read about high-throughput SV analysis in the human genome on the PromethION:
nanoporetech.com/resource-centre/workflow-structural-variation

Find out more about the PromethION platform:
nanoporetech.com/products/promethion
There are several options available for SV detection from nanopore sequencing data.

**EPI2METM** is a cloud-based data analysis platform, providing easily accessible, real-time, end-to-end analysis workflows — without the need for prior bioinformatics experience. The **EPI2ME FASTQ SV caller** enables rapid detection of deletions, insertions, and duplications in a whole human genome dataset. Data is first aligned to the human reference genome GRCh38 via LRA6, then SVs called using cuteSV7. The workflow outputs a searchable list of variants and their genomic locations; the results are downloaded in a Variant Call Format (VCF) file for sharing and further analysis.

For those with experience in use of the command line, an end-to-end pipeline, **pipeline-structural-variation**, is available. Here, LRA and cuteSV are wrapped into a Snakemake workflow, providing further flexibility and the option to call SVs in the human genome and other organisms. A tutorial is now available for this pipeline within **EPI2ME Labs**: an innovative, customisable bioinformatics solution designed to help you in developing your skills and confidence in nanopore sequencing data analysis. There are also several community-developed tools available for detection of SVs, with new software and updates in continual development.

### SV calling in the human genome

The performance of pipeline-structural-variation was benchmarked against the preliminary SV truth set produced by the Genome in a Bottle Consortium (GiAB) for the well-characterised human genome HG002 (NA24385)5. Sequencing was performed on R9 flow cells; data was then basecalled with Guppy 5 using the high accuracy (HAC) model. Sequencing data which passed MinKNOW™ QC filtering, representing from 10x to 50x depth of coverage of the genome, was assessed for precision and recall in SV detection (**Table 1**). The results show that strong SV calling metrics can be achieved across a range of depths of coverage; for best results, we recommend sequencing to a depth of coverage of 30x or more, though this can be reduced to 15x for a more light-touch whole-genome survey.

An HG002 nanopore open dataset, providing both raw signal .fast5 files and basecalled .fastq files, is available. This can be used to develop familiarity with analysis tools for alignment and SV calling. The dataset can be accessed here: labs.epi2me.io/gm24385_2021.05/

<table>
<thead>
<tr>
<th>Depth</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x</td>
<td>91.3</td>
<td>94.5</td>
<td>92.9</td>
</tr>
<tr>
<td>20x</td>
<td>96.3</td>
<td>95.4</td>
<td>95.8</td>
</tr>
<tr>
<td>30x</td>
<td>97.2</td>
<td>95.2</td>
<td>96.2</td>
</tr>
<tr>
<td>40x</td>
<td>97.8</td>
<td>95.3</td>
<td>96.5</td>
</tr>
<tr>
<td>50x</td>
<td>98.0</td>
<td>95.3</td>
<td>96.5</td>
</tr>
</tbody>
</table>

**Table 1.** Precision, recall, and F1 metrics for SV detection in human genome HG002 against the GiAB preliminary SV truth set, across a range of sequencing depths of coverage.
Which approach do I choose?

START HERE

What is your experimental aim?

Whole-genome SV detection

Which of these features are most important for your application?

Generating ultra-long reads e.g. to span very large SVs or repeat expansions

High depth of coverage e.g. for SV detection in highly heterogeneous samples

Enriching megabase/ chromosome-scale regions

Routine sequencing of regions spanning 10s-100s kb

Targeted SV detection

Which best describes your targeted sequencing needs?

Targeted method: adaptive sampling
Performed during sequencing; targeted library prep not required*

Are your target regions amenable to PCR?

NO

YES

Ultra-Long DNA Sequencing Kit

Preparation time ~200 minutes†

Input recommendation 6 million tissue culture cells or 1–2 ml blood

Fragmentation Transposase-based

Read length R9.4.1 flow cells: 50–100+ kb N50, reads up to 4+ Mb

BUY NOW: store.nanoporetech.com/ultra-long-dna-sequencing-kit.html

Ligation Sequencing Kit

Preparation time ~60 minutes

Input recommendation 1000 ng dsDNA

Fragmentation Optional; recommended for inputs of 100-500 ng

Read length Equal to fragment length

BUY NOW: store.nanoporetech.com/ligation-sequencing-kit110.html

Cas9 Sequencing Kit

Preparation time 110 minutes

Input recommendation 5 μg dsDNA

Fragmentation Not required

Read length Equal to fragment length

BUY NOW: store.nanoporetech.com/cas9-sequencing-kit.html

PCR Sequencing Kit

Preparation time 60 minutes + PCR

Input recommendation <100 ng gDNA

Fragmentation Optional

Read length Equal to fragment length post-PCR

BUY NOW: store.nanoporetech.com/pcr-sequencing-kit.html

*Read more about planning an adaptive sampling experiment in the Nanopore Documentation section of the Community: community.nanoporetech.com/docs
†Excludes an overnight elution step (≥ 12h)
Careful extraction of HMW DNA for nanopore sequencing ensures that SVs can be preserved within single, long fragments for sequencing. Pipetting gently using wide-bore tips and avoiding vortexing also help to preserve long fragments. To find and compare protocols that are optimised for fragment length for your sample type, visit the ‘Prepare’ Documentation section of the Nanopore Community.

**How do I design my protocol?**

The interactive Protocol Builder allows you to create your own end-to-end protocol for SV analysis and other applications. Advice starts at the beginning with DNA extraction, then goes through library preparation, sequencing, and finally data analysis.

Design your bespoke protocol with the Protocol Builder:
community.nanoporetech.com/knowledge/protocol_builder

**How can I extract high molecular weight (HMW) DNA?**

Careful extraction of HMW DNA for nanopore sequencing ensures that SVs can be preserved within single, long fragments for sequencing. Pipetting gently using wide-bore tips and avoiding vortexing also help to preserve long fragments. To find and compare protocols that are optimised for fragment length for your sample type, visit the ‘Prepare’ Documentation section of the Nanopore Community.

Find DNA extraction protocols:
community.nanoporetech.com/docs/prepare/extraction_protocols
SAMPLE PREP
Should I perform fragmentation or size selection?

Fragmentation is optional for nanopore sequencing. When aiming to capture an SV spanning tens or even hundreds of kilobases in a single read, you may wish to avoid any fragmentation, to help preserve long and ultra-long molecules; this also provides a simpler workflow. Although fragmentation increases sample preparation time, we have observed that combining size selection with light sample shearing can improve read length N50. For genome-wide SV calling, we recommend aiming for a read length N50 of 25-35 kb. Discarding shorter DNA molecules via size selection can help ensure more sequencing time is devoted to producing long reads, to better span large SVs.

LIBRARY PREP
How can I perform target enrichment without compromising read length?

With nanopore sequencing, it is also possible to perform targeted sequencing without PCR, allowing enrichment of large SVs in regions that cannot be amplified. With Cas9-based targeted sequencing, fragments spanning 10s-100s kb can be enriched in under two hours, without the need for PCR. The method is simple, streamlined, and enables enrichment of panels of targets. Meanwhile, adaptive sampling is an innovative technique allowing targeted enrichment or depletion during sequencing – no additional library prep needed. The method, which requires bioinformatics experience, is ideal for enriching very large SVs reaching the megabase scale. Furthermore, with either method, base modifications are preserved and can be characterised from the sequence data.

Read our end-to-end, best practice workflow on SV calling:
nanoporetech.com/resource-centre/workflow-structural-variation

Find out more about methods of targeted sequencing:
nanoporetech.com/targeted-sequencing-guide
From sample to answer

LIBRARY PREP

How do I sequence SVs in multiplex?

Multiple samples can be sequenced in a single sequencing run via the addition of a different barcode sequence to each sample during library prep; samples are then pooled together and prepared for sequencing as a single library. The Native Barcoding expansion packs use a ligation-based method of barcode attachment, enabling PCR-free multiplexed sequencing and preservation of methylation. If you plan to target SVs of interest via PCR, barcoding can be incorporated using the PCR Barcoding Kit and a four-primer PCR method. This is achieved using custom ‘inner’ primers, targeting your regions of interest, which are tailed with a specific sequence enabling subsequent amplification with barcoded nanopore primers. This allows amplification of barcoded target sequences.

Find out more about sample barcoding methods:
store.nanoporetech.com/sample-prep.html

SEQUENCING

Which device should I use?

Nanopore sequencing technology is highly scalable, with a range of devices suited to different requirements. For portable, real-time sequencing at the sample source, the MinION™ Mk1B can be plugged into a laptop, whilst the MinION Mk1C integrates compute and a high-resolution touchscreen, providing an all-in-one sequencing solution. Both are ideal for genome-wide SV analysis in microbes and other small genomes, and targeted SV detection.

Scaling up, the benchtop GridION™ provides up to five times the sequencing capacity of the MinION and MinION Mk1C, enabling great flexibility. For rapid, sensitive SV detection across large genomes, the PromethION P24 and P48 facilitate high-depth sequencing at scale.

Compare nanopore sequencing devices in detail:
nanoporetech.com/products/comparison
How can I check the quality of my library before generating my dataset?

The Flongle™ is a ‘flow cell dongle’: an adapter which converts MinION and GridION devices for use with small, cost-effective Flongle Flow Cells. Sequencing a small amount of library on a Flongle Flow Cell provides a useful QC step, indicating read length and throughput prior to performing your final sequencing run. Flongle can also be used for SV experiments that require smaller datasets, such as targeted sequencing of a single region.

Read more about Flongle:
nanoporetech.com/products/flongle

Which workflow should I use to analyse my data?

A variety of SV analysis solutions are available for all levels of bioinformatics expertise. The cloud-based EPI2ME FASTQ SV Caller for Human Analysis enables real-time identification of SVs present in human genomes; it is ideal for those who are new to analysis or wish to simplify their workflow. Available on GitHub, pipeline-structural-variation requires use of the command line; it is suitable for both whole-genome and targeted data derived from either human or non-human samples.

This pipeline is integrated into a step-by-step notebook tutorial available via EPI2ME Labs. A nextflow-based workflow is also available on EPI2ME Labs; this is an ideal option for those wishing to perform high-throughput analyses and reporting.

Find out more about analysing data with EPI2ME, EPI2ME Labs, and bioinformatics pipelines:
nanoporetech.com/analyse

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pipeline-structural-variation</th>
<th>EPI2ME Labs tutorial</th>
<th>EPI2ME Labs workflow (wf-human-sv)</th>
<th>EPI2ME workflow (FASTQ SV caller for human analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>* * * * *</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Human</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Non-human</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Whole-genome</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Targeted</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Variants detected</td>
<td>INS, DEL</td>
<td>INS, DEL</td>
<td>INS, DEL</td>
<td>INS, DEL, DUP</td>
</tr>
</tbody>
</table>
How can I assess methylation in SVs?

Amplification-free nanopore sequencing allows you to detect SVs and their methylation state from the same dataset, without any special library preparation steps. There are several analysis tools for methylation detection, which are available on GitHub. We recommend the following:

**Megalodon:** a high-performance tool, in rapid development, that can be used to perform basecalling, modification detection, reference alignment, and variant calling; this approach is especially recommended where samples have been sequenced to a lower depth of coverage.

**f5c/nanopolish:** given a set of aligned nanopore reads and raw signal, f5c can be used for modification detection and can optionally utilise NVIDIA graphics cards for acceleration. f5c is a good choice for when samples have been sequenced to a higher depth of coverage.

**Taiyaki:** this research software enables you to train basecalling models to detect your modifications of interest.

Find out more about modified base detection using nanopore sequencing: nanoporetech.com/applications/investigation/epigenetics

Find Oxford Nanopore analysis tools on GitHub: github.com/nanoporetech
Case study 1: Population-scale analysis of structural variants in the human genome with long nanopore sequencing reads

Beyter et al.\textsuperscript{11} used high-throughput nanopore sequencing for population-scale analysis of SVs in the human genome. In their study at deCODE genetics, Iceland, they performed whole-genome native DNA sequencing of 3,622 individuals, selected from a database of DNA sequence variation in the Icelandic population. With a read length N50 of 19.9 kb, the sequencing data revealed a median of 22,636 autosomal SVs per individual, spanning a median total length of 10.02 Mb. The authors noted that, in previous large-scale studies using short-read technology, only ~2,000-11,000 SVs had been identified per genome.

A high-confidence SV set was then generated, genotyped, and imputed into data from 166,281 Icelanders. Analysis revealed peaks in read length corresponding to SINE, SVA, and LINE elements, with more SVs observed in telomeric regions. SVs hypothesised to affect protein function were found to be disproportionately rare. Comparison of the SVs against phenotype-associated GWAS variants revealed 5,238 SVs to be correlated with 11,194 markers. Association of the SVs with the phenotypic data available at deCODE genetics found that carriers of a 14,154 bp deletion overlapping the first exon of \textit{PCSK9}, a target of cholesterol-lowering drugs, had lower levels of LDL cholesterol. The team highlighted how ‘a better understanding of the biochemical causes and consequences of SVs will be essential to understand human evolution and disease’.

Case study 2: Revealing extensive structural variation in a major crop plant genome

In their study of the economically important crop \textit{Brassica napus} (canola), Chawla et al. stressed the significance of plant genome SVs in trait variation\textsuperscript{12}. Characterisation of this variation could aid in plant breeding for crop improvement; however, SV detection via short-read sequencing has proved difficult due to the complex, repetitive nature of such plant genomes. To characterise SVs in the allopolyploid \textit{B. napus} genome, the team generated long nanopore sequencing reads for four diverse plant lines, encompassing genotypes with different flowering times from around the world. By enriching for ultra-high molecular weight DNA through extraction and sample preparation, they obtained read length N50s of 10.8-28.9 kb in sequencing; this enabled the detection of complex SVs spanning up to 28 kb. They found that ‘up to 10% of all genes were affected by small- to mid-scale SV events’ spanning 30-30,000 bp; nearly half of these were between 100-1,000 bp, making them difficult to detect via short-read sequencing. The data allowed for thorough analysis of SVs and their functional effects, including variants associated with flowering time and disease resistance. Writing about the study, Harmeet Singh Chawla explained how their results suggested that the use of long nanopore sequencing reads to characterise complex plant genomes ‘might lead to the discovery of previously ‘hidden’ functional gene variation, with major implications for trait regulation and crop improvement.’\textsuperscript{13}