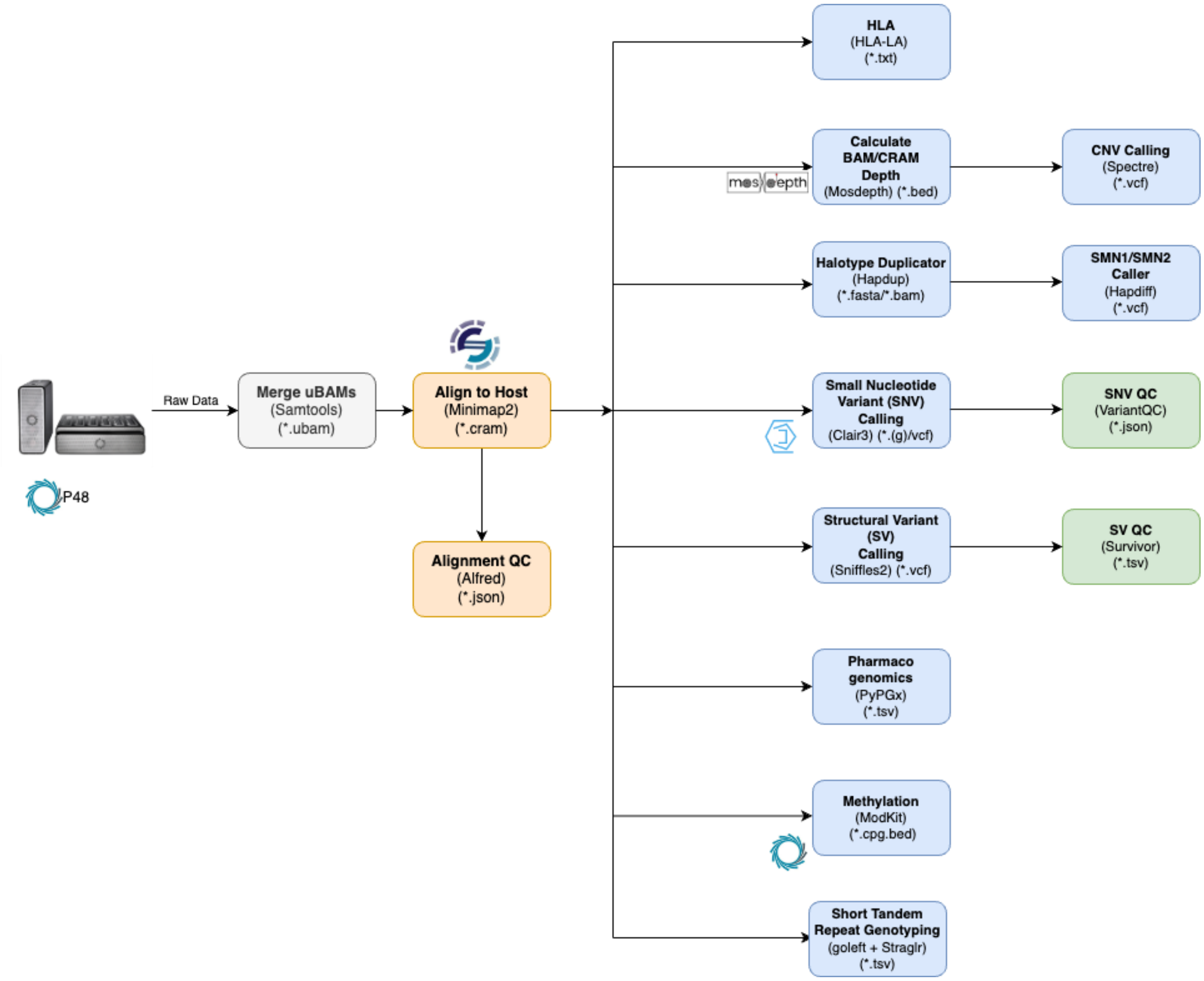
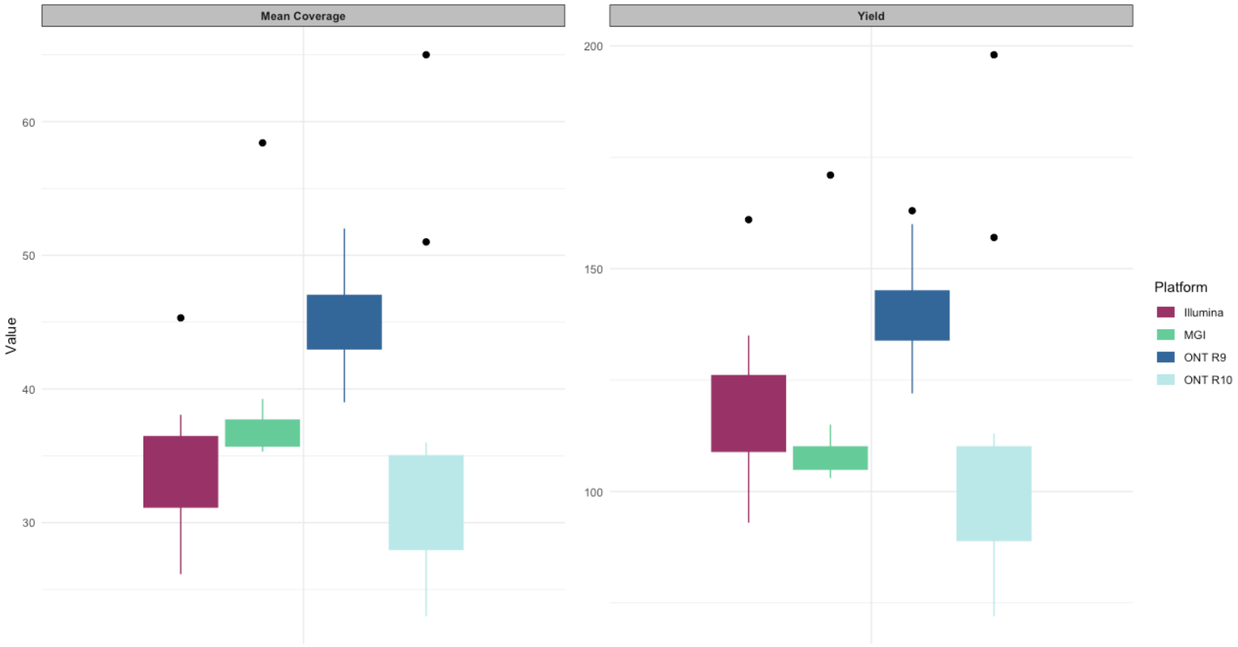
Supplemental material

SUPPLEMENTAL FIGURES

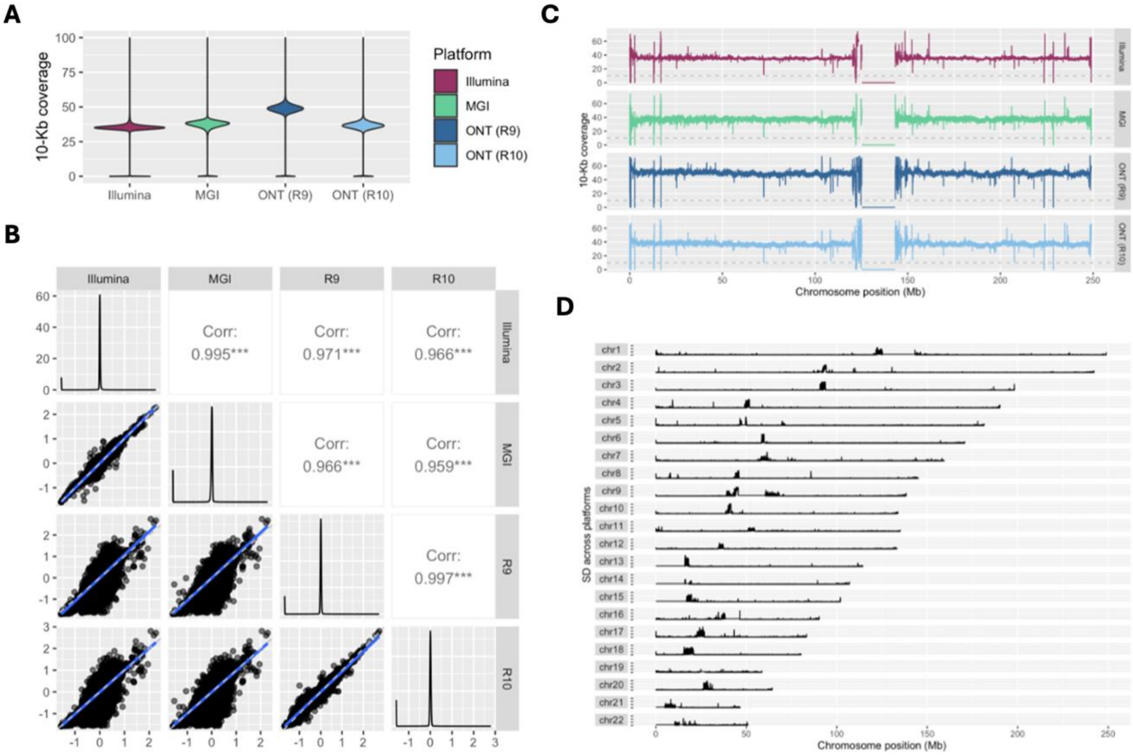
Supplemental Figure S1. ONT sequencing data analysis workflow.


P48
Raw Data
Merge uBAMs (Samtools) (*.ubam)
Align to Host (Minimap2) (*.cram)
Alignment QC (Alfred) (*.json)
HLA (HLA-LA) (*.txt)
Calculate BAM/CRAM Depth (Mosdepth) (*.bed)
mosdepth
CNV Calling (Spectre) (*.vcf)
Halotype Duplicator (Hapdup) (*.fasta/*.bam)
SMN1/SMN2 Caller (Hapdiff) (*.vcf)
Small Nucleotide Variant (SNV) Calling (Clair3) (*.(g)/vcf)
SNV QC (VariantQC) (*.json)
Structural Variant (SV) Calling (Sniffles2) (*.vcf)
SV QC (Survivor) (*.tsv)
Pharmaco genomics (PyPGx) (*.tsv)
Methylation (ModKit) (*.cpg.bed)
Short Tandem Repeat Genotyping (goleft + Straglr) (*.tsv)

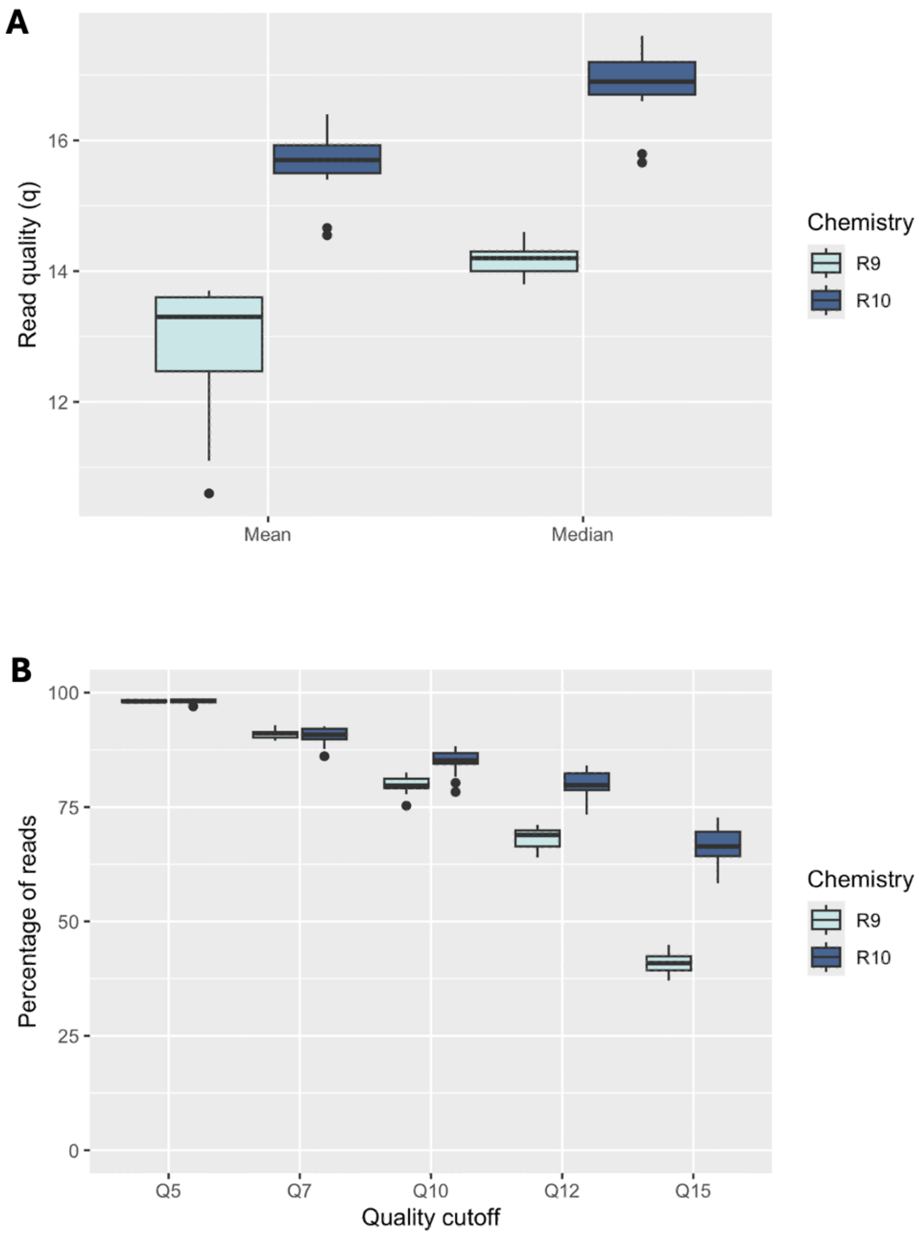
Supplemental Figure S2. Sequencing yield and mapping mean coverage distribution across samples and platforms.



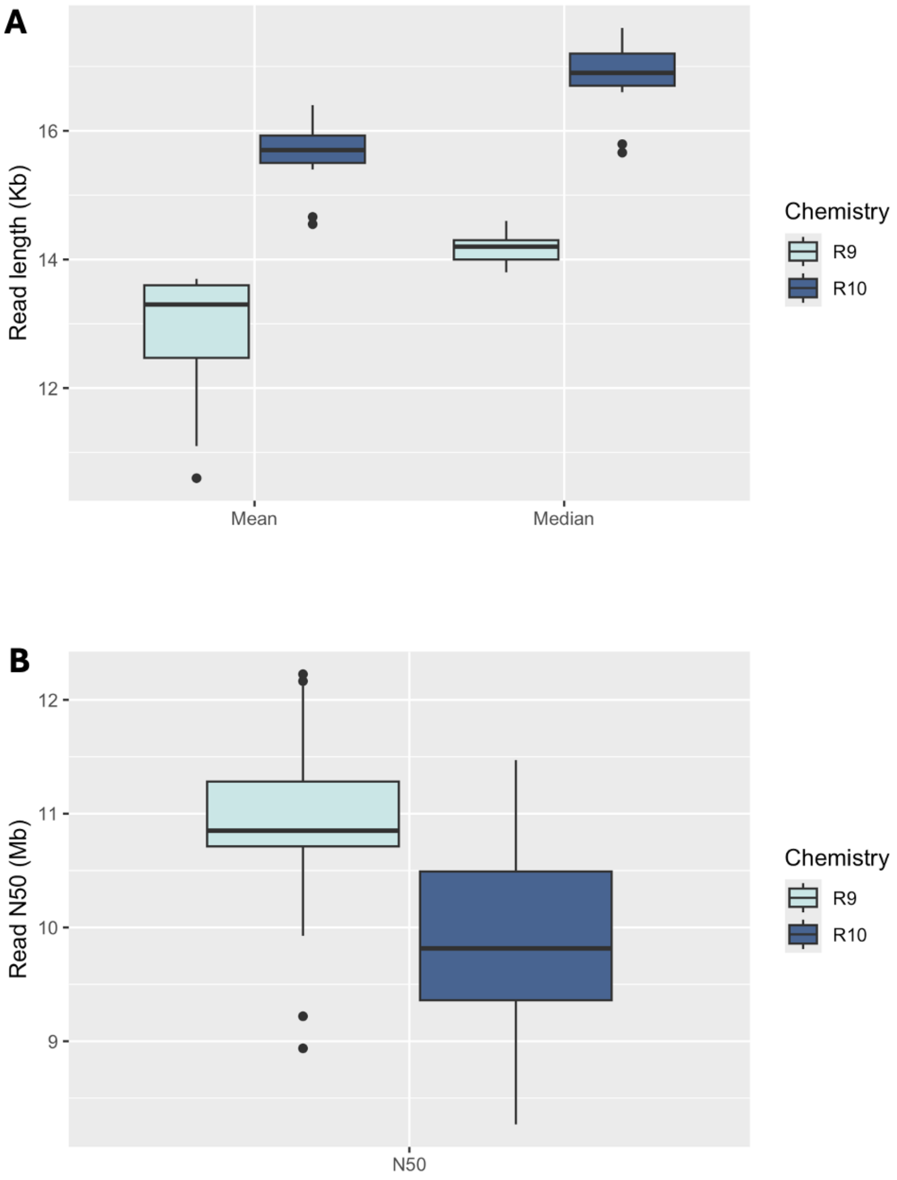
Supplemental Figure S3. Read quality distribution in ONT samples. **(A)** Coverage distribution across sequencing platforms. In each platform, we calculated mapping coverage in 1 kb windows of the genome, averaged across the 17 samples and we further averaged the resulting avalue across ten 1-kb windows to obtain the 10-kb coverage values. **(B)** all-to-all 10-kb coverage correlation. **(C)** 10-kb coverage along Chromosome 1 across all four platforms. **(D)** Genomic regions with highest difference (measured as the standard deviation) in 10-kb coverage values across platforms.



Supplemental Figure S4. Read quality distribution in ONT samples.

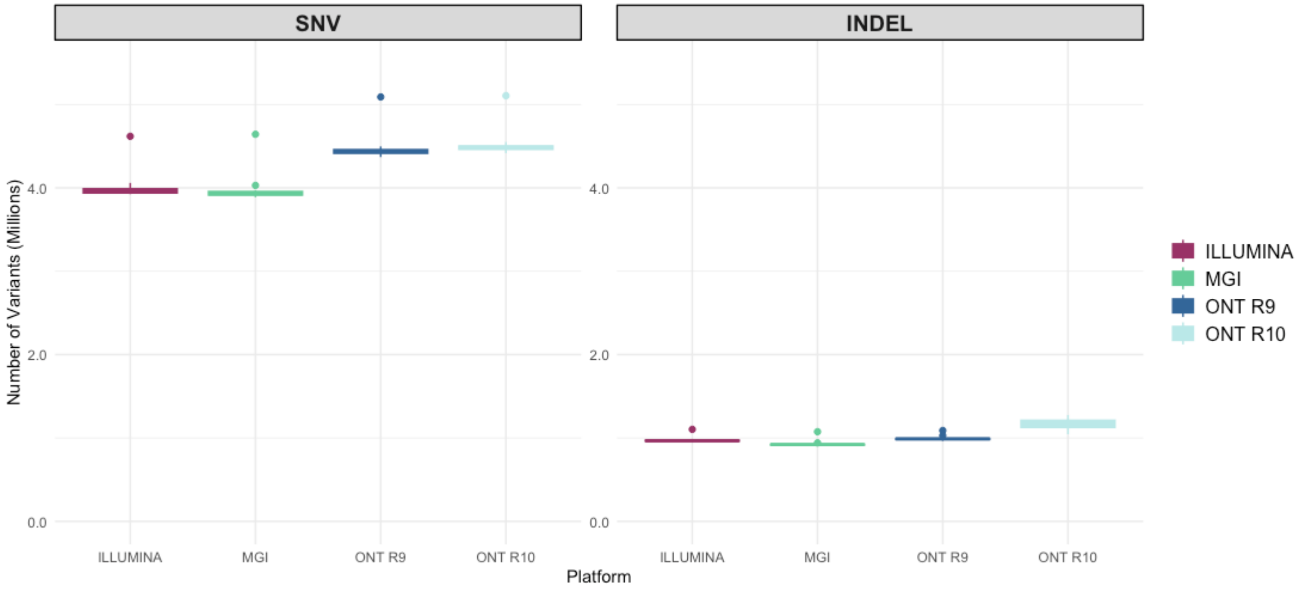


Supplemental Figure S5. Read length distribution in ONT samples.



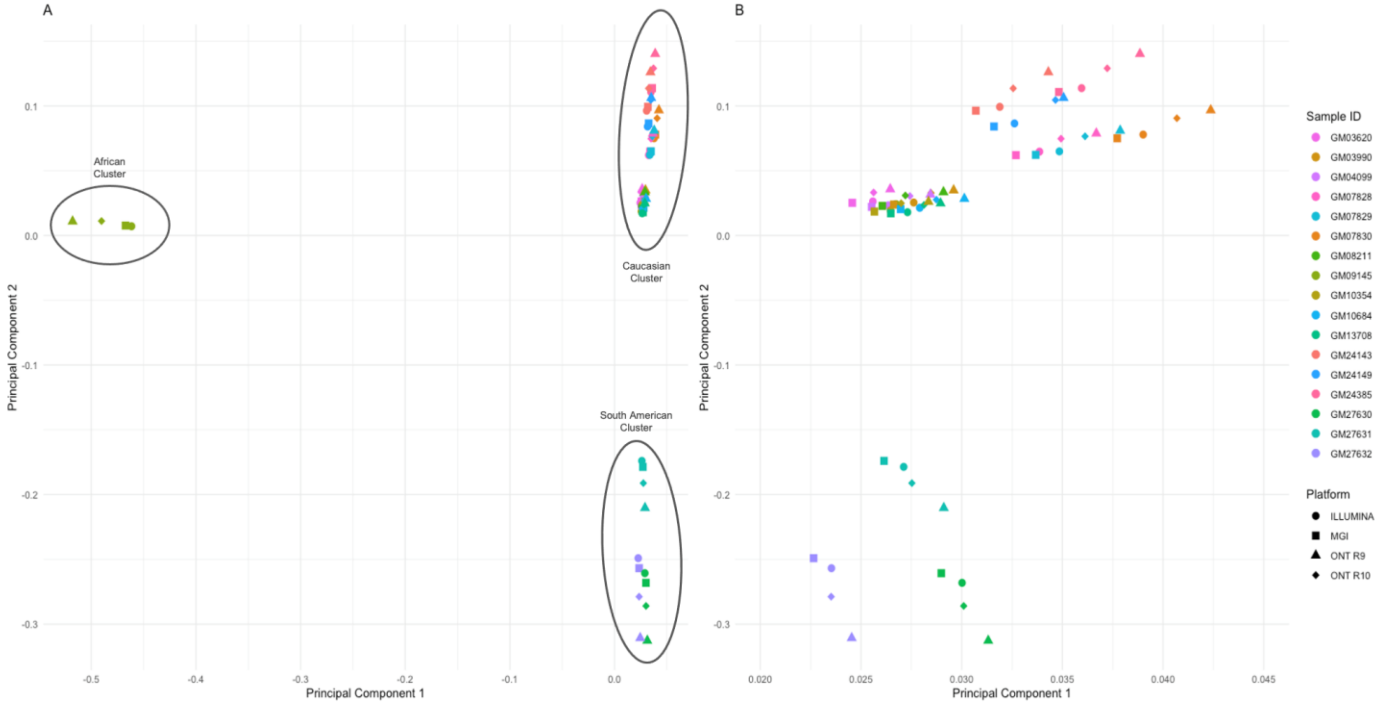
Supplemental Figure S6. Number of small variants called per sample across the four sequencing platforms.

The boxplot shows the distribution of SNVs and INDELs identified in Coriell samples sequenced using Illumina (4.0M SNVs, 1.0M INDELs), MGI (4.0M SNVs, 0.9M INDELs), ONT R9 (4.5M SNVs, 1.0M INDELs), and ONT R10 (4.5M SNVs, 1.2M INDELs).



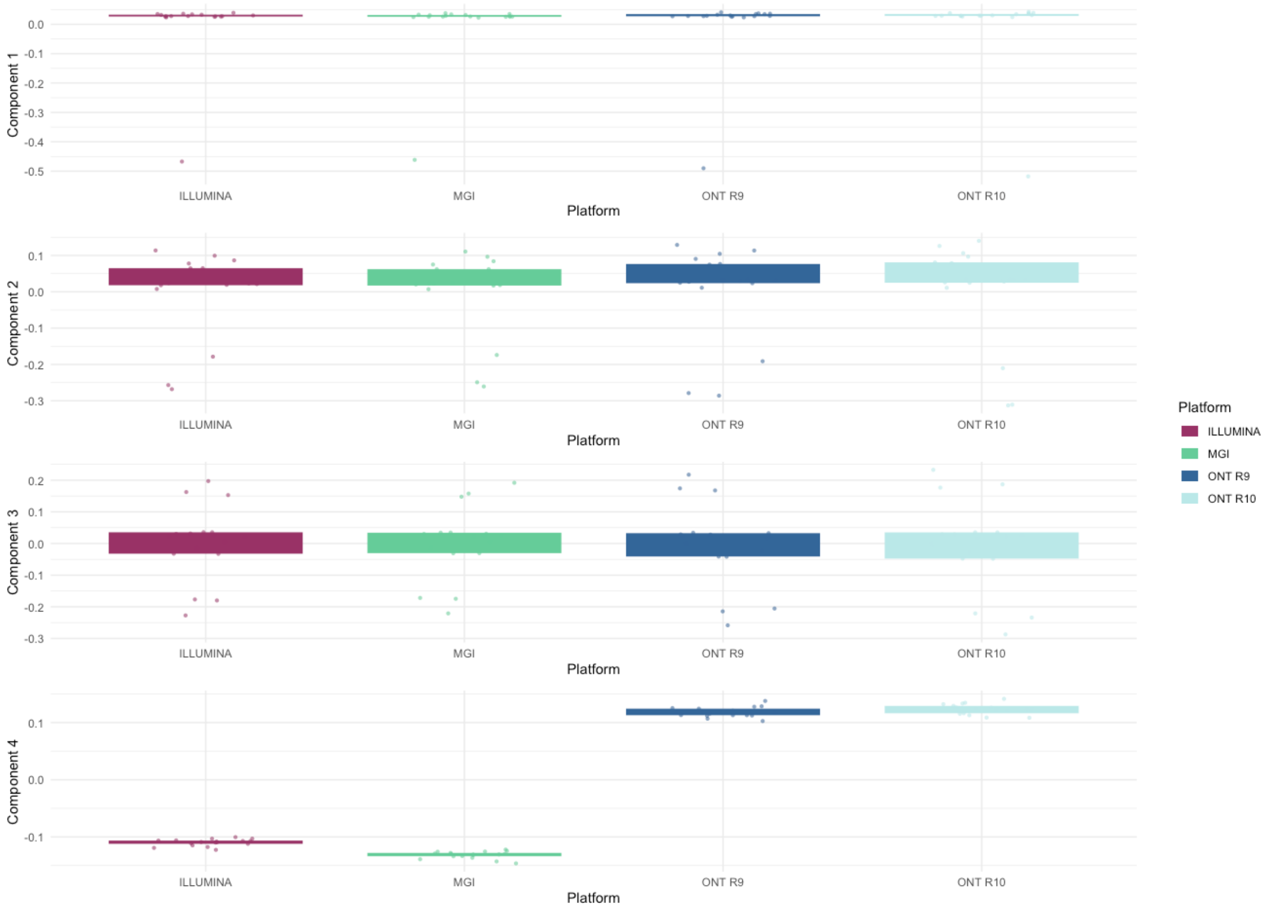
Supplemental Figure S7. Clustering of the 17 Coriell reference samples reflects ancestry and familiar relationship.

(A) Scatter plot of PC1 versus PC2 following PCA on the SNP/INDEL calls for the 17 Coriell reference samples. We manually labelled the main three ancestries (African, Caucasian and South American). **(B)** Same plot as in (A) but removing the samples of African ancestry, which makes more visible that samples from the same individual cluster together and the two families do it too. South American trio: GM27630 (son), GM27631 (father) and GM27632 (mother). Caucasian family: GM24143 (mother), GM24149 (father) and GM24385 (son).



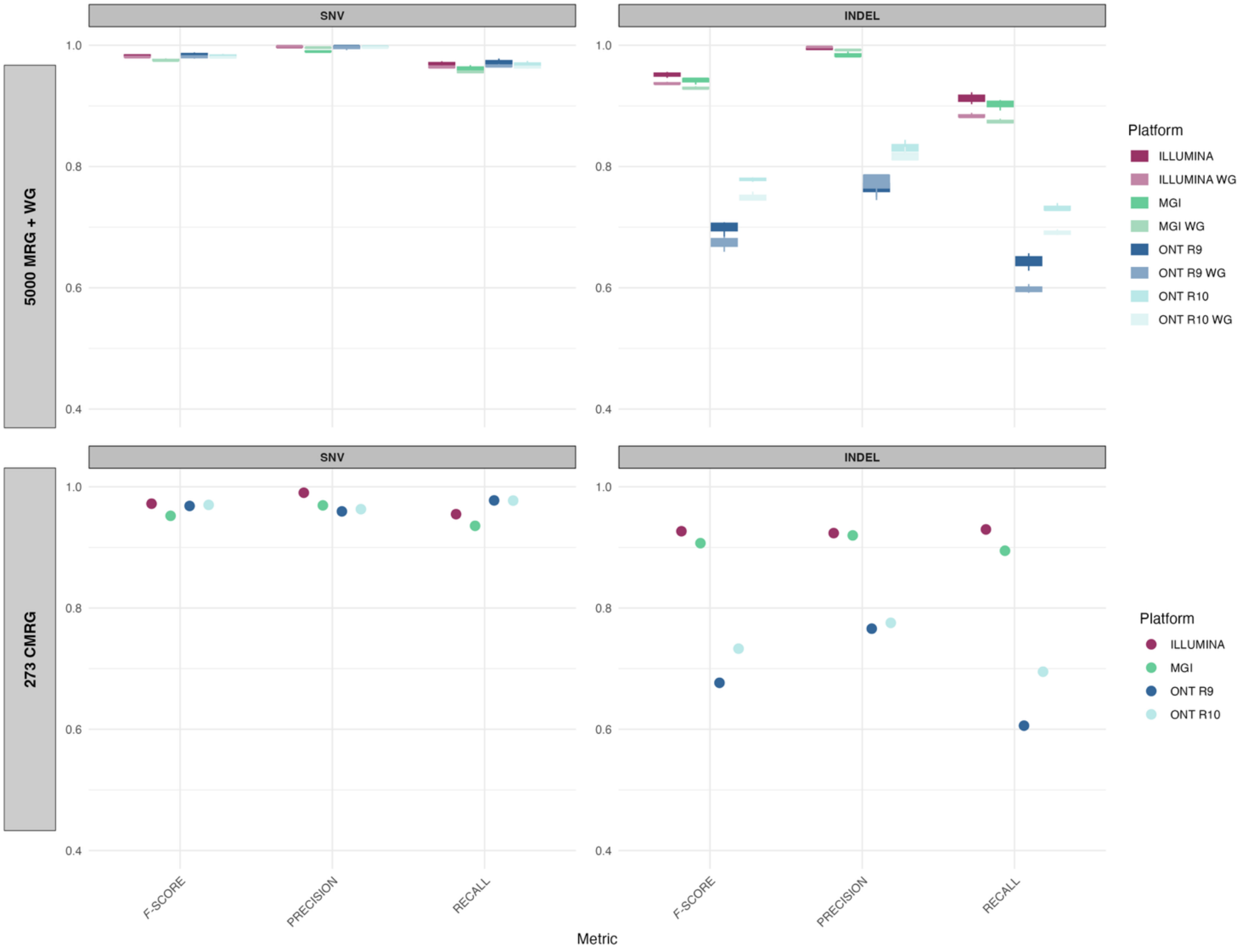
Supplemental Figure S8. PC4 captures differentiates sequencing technology.

Boxplots of the first four PCs from the PCA, grouped by sequencing platform (Illumina, MGI, ONT R9, ONT R10).



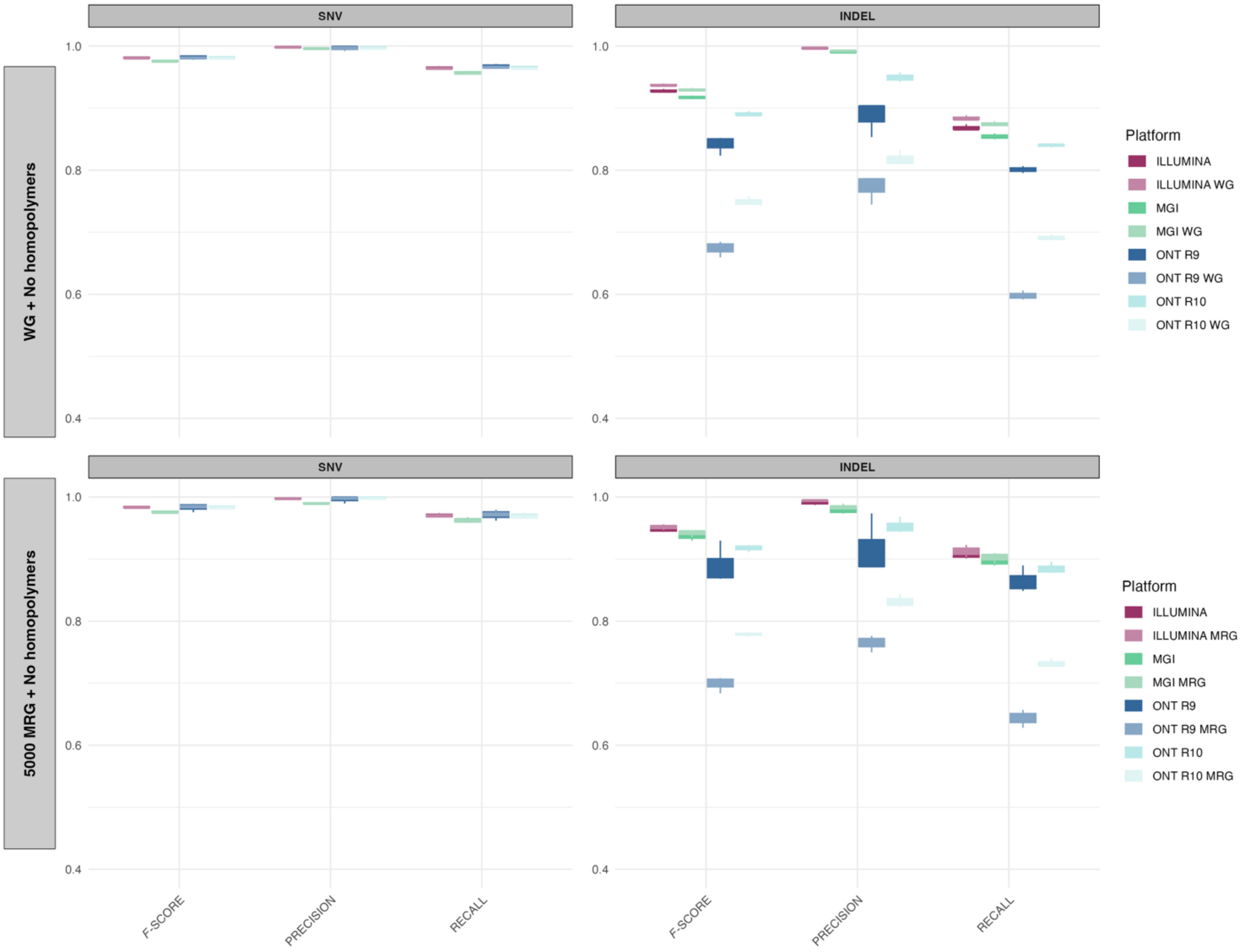
Supplemental Figure S9. SNV and INDEL variant calling performance in MRG and CMRG.

The top panel shows the performance of samples GM24143, GM24149, and GM24385 across 5,000 medically relevant genes (MRG) as well as at the whole-genome level. The bottom panel highlights the performance in a set of 273 challenging, clinically significant genes (CMRG) for sample GM24385.



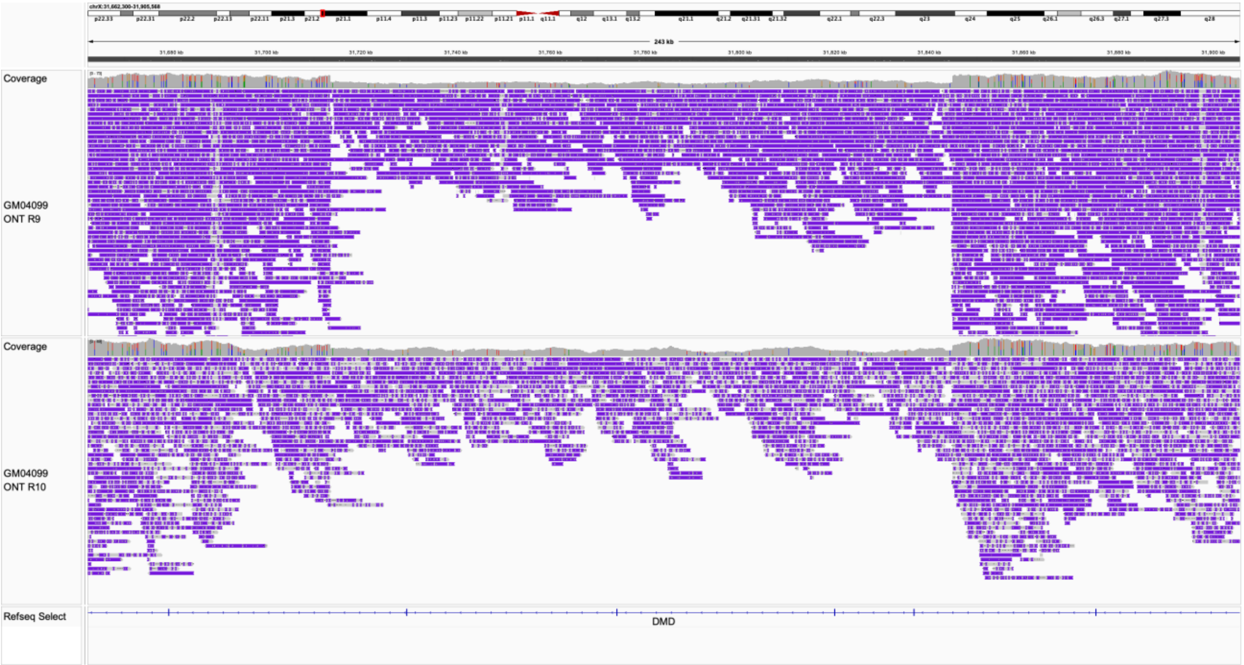
Supplemental Figure S10. Impact of homopolymer exclusion on SNV and INDEL calling performance across sequencing platforms.

The top panel shows *F*-score, precision, and recall distributions for samples GM24143, GM24149, and GM24385 across the whole genome; the bottom panel shows the same metrics restricted to the 5,000 medically relevant genes (MRG). Darker bars represent variant calling performance before homopolymer exclusion, while lighter bars reflect the results after excluding these regions.



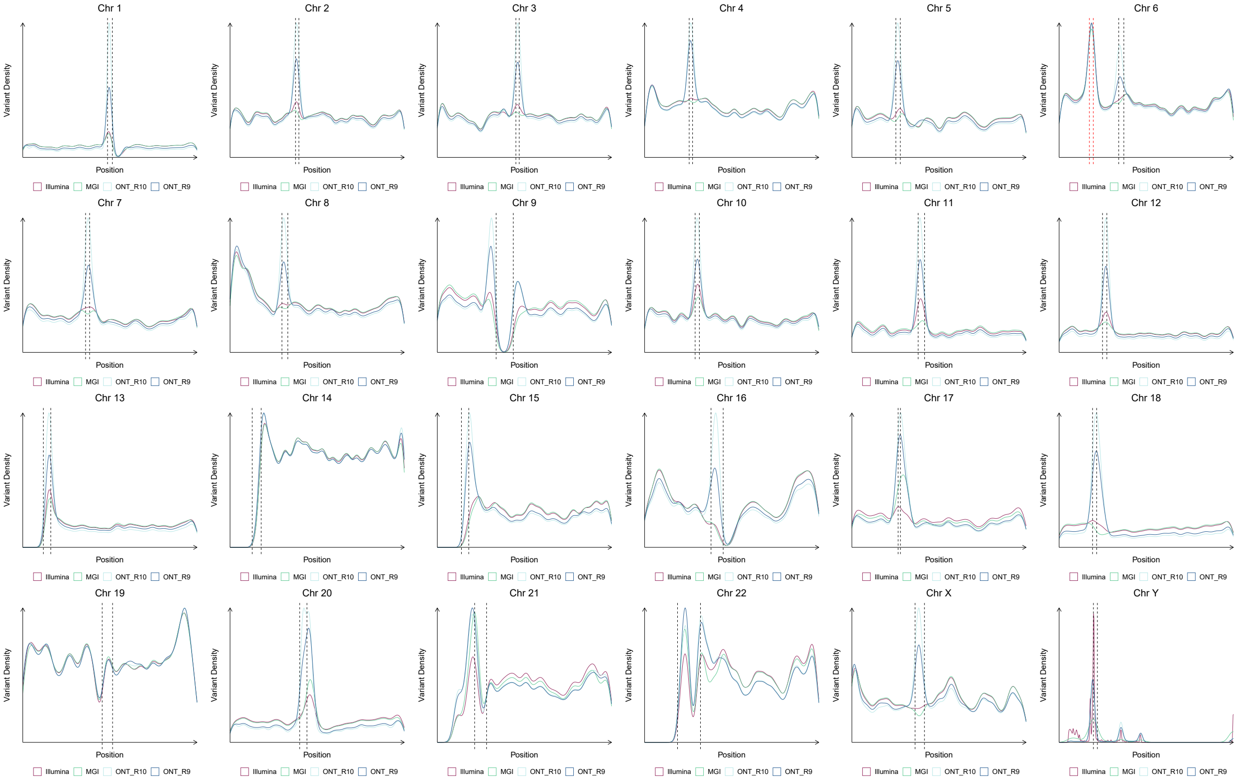
Supplemental Figure S11. IGV screenshots for Chromosome X in the region of the *DMD* gene for sample GM04099.

The top panel presents the ONT R9 results, while the bottom panel displays the ONT R10 results. In both cases, the decrease in coverage denoted by the smaller pile-up of aligned reads hints the presence of the deletion.



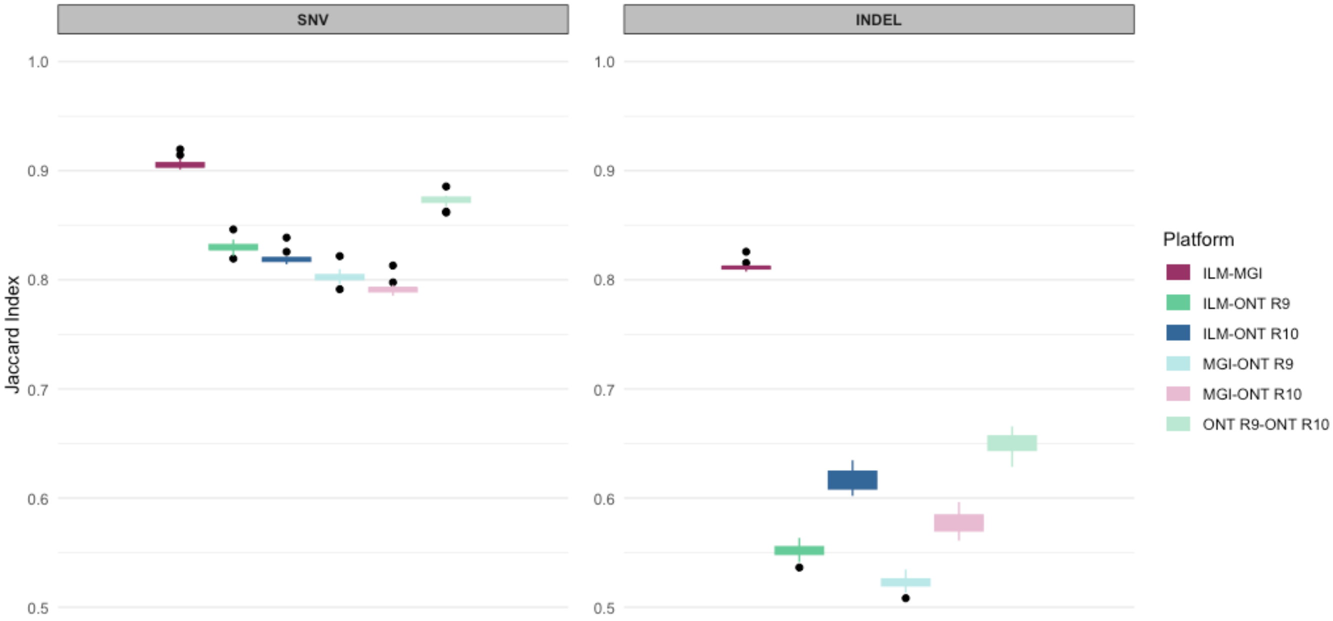
Supplemental Figure S12. SNV density chromosome profiles across platforms.

The dashed black lines mark the centromeres and, for Chromosome 6 only, the dashed red lines indicate the HLA region.



Supplemental Figure S13. SNV and INDEL genotype concordance between sequencing platforms.

The Jaccard index values were calculated to assess the concordance of SNV and INDEL detection across the different sequencing platforms. SNVs display consistently high concordance, indicating reliable detection capabilities across platforms. In contrast, INDELs show significantly lower concordance, particularly in the comparison between short and long reads platforms.



SUPPLEMENTAL NOTES

Supplemental Note S1. Sample Processing, DNA Extraction and Sequencing Protocols.

Coriell reference samples

Lymphoblastoid cell lines were obtained from Coriell Cell Repositories (Camden, NJ) and processed according to supplier instructions. We opted to purchase live cell lines and extract DNA in-house to preserve native DNA integrity, as shipping purified DNA may cause fragmentation detrimental to Oxford Nanopore sequencing. Upon receipt, cells in T-12.5 mL flasks were incubated overnight at 37 °C, then transferred to 50 mL centrifuge tubes and spun for 10 min at 100×g. Pellets were resuspended in complete RPMI 1640 medium with 15% FBS, seeded at approximately 1×10^6 cells per 25 mL flask containing 10 mL medium, and incubated at 37 °C with 5% CO_2. Once sufficient growth was achieved, cells were harvested and counted.

DNA extraction

DNA was extracted using the PureLink™ Genomic DNA Mini Kit (Thermo Fisher Scientific) following the manufacturer’s protocol. Briefly, cell pellets were obtained by centrifugation at 15,000 rpm for 5 minutes, resuspended in 200 µL of phosphate-buffered saline (PBS) solution and 20 µL each of proteinase K and RNase A. After adding 200 µL PureLink Genomic Lysis/Binding Buffer, samples were vortexed and incubated at 55°C for 10 minutes. Ethanol (200 µL of 96-100%) was then added, and the lysate (approximately, 640 µL) was transferred to a spin column for binding. The column was then washed twice with 500 µL of Wash buffer 1 and Wash buffer2 respectively, and finally DNA was eluted in 200 µL of elution buffer by centrifugation at maximum speed for 1 minute at room temperature.

Sequencing

Illumina

Whole genome sequencing (30×) library preparation was performed by using Illumina PCR free prep library kit. Following instructions from the manufacturer, gDNA input of 250 to 750ng was fragmented by Bead-linked transposome and ligation was done using IDT® for Illumina® UMI DNA/RNA UD Indexes Set A (96 Indexes, 96 Samples). All 24 samples were pooled based on index compatibility and sequenced using the NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles) on the NovaSeq 6000 System.

MGI

Library preparation was done using MGIEasy PCR-Free DNA Library Prep (96 RXN) kit. A total of 900 ng DNA in 48 µL was used. Preparation steps included fragmentation, size selection, end repair, adapter ligation, denaturation, circularization and exo-digestion. Double-size selection was performed for the samples with 0.6× and 0.2× DNA easy clean beads. Quality check for the single stranded circular libraries was performed using Qubit SS DNA kit. Library concentrations in the range of 0.6-3 ng/µL were considered qualified for DNA Nanoball (DNB) preparation. Samples with DNB concentrations between 8ng/ µL to 40ng/µL were pooled and loaded onto the DNBSEQ T10 flowcell and sequenced using DNBSEQ-T10RS DNB Sequencing Set (FCL PE100) (940-000078-00, MGI, Shenzhen, China). The recorded data was analysed using ZLIMS Elite v1.0.5.2 software with MEGABOLT_2 pipeline.

ONT

Library preparation was carried out using Ligation sequencing kit 110 and 114. A total of 1000 ng DNA in 50 µL was used for library preparation. Preparation steps included normalization, mechanical fragmentation using FastPrep, end repair, adapter ligation. Quality check for the double stranded libraries were performed using Qubit ds DNA kit. Libraries with 400ng/µL were loaded onto the PromethION flowcell and sequenced using PromethION 48. The recorded raw data was processed using MinKNOW software (Oxford Nanopore Technologies) with Dorado for real-time basecalling, utilizing the dna_r9.4.1_450bps_modbases_5mc_cg_hac_prom.cfg model for the R9 chemistry and the dna_r10.4.1_e8.2_400bps_5khz_modbases_5hmc_5mc_cg_hac_prom.cfg model for R10, to produce high-accuracy base calls from the raw signal data. Basecalling results were analyzed for downstream analyses as described in the next section.

Supplemental Note S2. Overcoming the challenges of ONT WGS data analysis

The initial step in the processing of sequencing data generated by ONT sequencers is the base-calling. This step consists of determining the nucleotide sequence (DNA or RNA) from the electric squiggle stored in the raw signal generated by the instrument. Because methylation modifications in the nucleotide sequence are reflected in the squiggle, such modifications can be captured during the base-calling. We refer to canonical base-calling to only determining the nucleotide sequence and concurrent base-calling to the simultaneous inference of the nucleotide sequence and its methylation modifications.

Two critical challenges in the processing of ONT data are the large size of the raw signal output and the demanding computing requirements to perform the base-calling. Historically, the aggregated size of raw data for a single ~30× human genome is ~700 Gigabytes (GB). Besides, extracting the nucleotide sequence and methylation marks from raw data requires GPU computing power and takes 24-36 hours (about one to one and half days) even with the most advanced hardware and software combination for a ~30× genome.

Over the last three years, we have been early adopters of hardware and software improvements to overcome such challenges. First, we replaced the Nvidia GV100 towers initially connected by default to ONT's PromethION 48 sequencer with the more powerful Nvidia A100. Second, we upgraded from Guppy version 4 to version 6 to enable concurrent base-calling and more recently to the newer Dorado base-caller to achieve a two-fold increase in base-calling performance relative to Guppy version 6 (https://aws.amazon.com/blogs/hpc/benchmarking-the-oxford-nanopore-technologies-basecallers-on-aws/). We observed that Guppy version 6 on Nvidia A100 achieved real-time base-calling not for a full PromethION 48 run (48 flow cells) – instead, it could catch up with 24 and 12 flow cells with canonical and concurrent base-calling, respectively.
Prior to these improvements, real-time concurrent base-calling was not feasible. We had to upload large raw signal files (~700 Gigabytes per ~30× genome) to our cloud to perform base-calling, increasing the run times and the cloud storage footprint. Besides, concurrent base-calling prior to Guppy version 6 was prohibitive in terms of runtime and associated computing cost.

By default, concurrent base-calling would generate multiple large intermediate files (e.g., sequence and alignment outputs) that add up to >1 TB for a ~30× human genome. However, in our current workflow, real-time base-calling on the A100 system directly generates unaligned BAM (uBAM) files that are used for downstream analysis. Since uBAM files contain both sequence and methylation calls, no additional intermediate sequence files are generated or retained. Indeed, even if more accurate base-callers are developed in the future, the cost of storing raw signal data is likely to exceed that of re-sequencing. We therefore recommend keeping only the uBAM files for downstream analysis.

We accelerated the mapping of the ONT sequencing reads to the reference genome by using Sentieon's Minimap2 version, which is approximately two times faster than the open-source version of the mapper. Indeed, that is one of the principal caveats we would see in using ONT's Human variation workflow, which uses the open-source Minimap2. Regardless of the mapper, we recommend storing the alignments in the CRAM format (instead of BAM). We found that, for ONT data, CRAM achieved on average ~40% compression relative to BAM, as well as data lossless and unaltered variant calls. In case of short nucleotide variant (SNV) calling, we have benchmarked the tool (https://github.com/HKU-BAL/Clair3) from version 1 to the current version 3 which showed significant improvements in terms of precision/recall as well the computational time improved with more newer releases. However, this SNV calling is a pain point in terms of the time taken to run a sample on gVCF mode (~ 20-24 hours for a ~30× genome) and is much faster with VCF mode (~5-6 hours). In the case of structural variant (SV) calling, we used Sniffles2 that produces structural variant calls within an hour and is highly recommended.