Population-scale Long-Read Whole Genome Sequencing of UK Biobank Participants using Oxford Nanopore

Anne-Katrin Emde[1], Alex Dawson[2], Andreas Venizelos[2], Anthony Doran[2], Bjarni V. Halldórsson[4], Brynja Sigurpálsdóttir[4], Daniel Fry[5], Daníel Gudbjartsson[4], Doruk Beyter[4], Droplaug N. Magnúsdóttir[4], Frosti Jónsson[4], Gísli Másson[4], Graham Hall[2], Guillaume Holley[4], Hannes P. Eggertsson[4], Heather Jeffery[2], Jawahar Swaminathan[2], Julian Gehring[3], Lynn Ly[1], Mark Effingham[5], Melissa Lewins[5], Ólafur T. Magnússon[4], Páll Í. Ólason[4], Patrick Sulem[4], Philipp Rescheneder[3], Rachel Hipkin[2], Rocío Esteban[2], Sean McKenzie[1], Sissel Juul[1], Sverrir T. Sverrisson[4], Unnur Thorsteinsdóttir[4]

1. Oxford Nanopore Technologies Inc, New York, NY, USA 2. Oxford Nanopore Technologies plc, Oxford, United Kingdom 3. Oxford Nanopore Technologies GmbH, Munich, Germany
4. Amgen, deCODE genetics, Reykjavík, Iceland5. UK Biobank, Stockport, Greater Manchester, United Kingdom

Abstract

A large-scale effort is currently underway to sequence 50,000 UK Biobank participants on the Oxford Nanopore PromethION™ platform, allowing simultaneous genome-wide detection of genetic variation and DNA methylation. Here, we present the first pre-release data on the first ~5,000 high-coverage samples, including ~1,000 participants sampled at two different time points. Sequencing QC demonstrated highly consistent sequencing quality and robust sequencer performance. We discover thousands of novel genetic variants compared to short-read WGS, with a particularly high number of novel structural variants. Further, we create assemblies that resolve haplotype-specific sequence variation, even in complex genomic regions such as *LPA*, and investigate the epigenetic state genome-wide without capture or conversion bias. Genomes from individuals of non-European continental ancestries contribute novel variation at higher rates demonstrating that larger and more diverse cohorts will make crucial contributions to variant discovery and improve our understanding of the functional impact of global genetic variation, a prerequisite for equitable global health particularly in the era of precision medicine. Together with the extensive phenotypic data available for UK Biobank participants, this cohort enables comprehensive genetic and epigenetic association analyses.

Bringing Long Reads to the UK Biobank

The UK Biobank is a foundational biomedical database integrating genomic data with extensive phenotypic, clinical and longitudinal health information. A subset of 50,000 participants are receiving Nanopore long read whole genome sequencing that will allow to reach into difficult regions of the genome, profile methylation genome-wide, and comprehensively detect structural variation including tandem repeats at population-scale.

Sequencing is currently ongoing at deCODE genetics (Iceland), with an initial set of ~5,000 samples sequenced at 30x coverage, to be increased to a total of 25,000 samples at 30x, and the remaining 25,000 samples sequenced at 15x coverage. To date, this constitutes the largest cohort of whole human genomes sequenced with Oxford Nanopore R10 chemistry that will become publicly accessible.



50,000 participants

- Trios & duos
- Longitudinal samples
- Metabolic disorders
- Nervous system disorders
- Cancer diagnoses
- Healthy controls

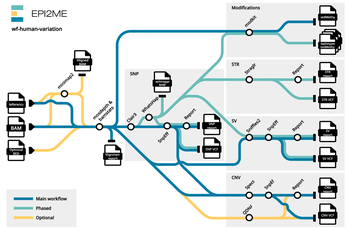
All data generated:

- PromethION™
- R10.4.1 flowcell
- LSK114 kit



Data availability:

- Data will become accessible to approved researchers through the UKB-RAP (2027)
- Output from automated analysis through EPI2ME wf-human-variation (v2.7.2) with added custom analysis
- Haplotagged CRAMs, VCFs, methylation, assemblies



References

1. Sudlow, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS medicine (2015).
2. UK Biobank Whole-Genome Sequencing Consortium. "Whole-genome sequencing of 490,640 UK Biobank participants." Nature (2025).
3. Wadsworth, et al. "Dark and camouflaged genomic regions remain challenging in CHM13." Scientific Reports (2026).
4. Holley, Eggertsson et al. "An Icelandic pangenome reference." Nature (2026).
5. Schachtl-Riess et al. "Frequent LPA KIV-2 variants lower lipoprotein (a) concentrations and protect against coronary artery disease." Journal of the American College of Cardiology (2021).
6. Sigurpalsdottir et al. "Nanopore sequencing identifies parent-of-origin specific age-associated methylation changes at imprinted loci in the human genome." Nature Communications (2026).
7. Loyfer, et al. "A DNA methylation atlas of normal human cell types." Nature 613.7943 (2023).
8. Horvath, Steve. "DNA methylation age of human tissues and cell types." Genome biology (2013).
9. Levine, et al. "An epigenetic biomarker of aging for lifespan and healthspan." Aging (2018).
10. Zhang, et al. "Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing." Genome medicine (2019)
11. Bai et al., "Genome-wide associations of structural variants with human traits through imputation from long-read assemblies." Nature Genetics (2026).
12. Chirmade et al. "GWAS Svatalog: a visualization tool to aid fine-mapping of GWAS loci with structural variations." Heredity (2026).
13. Gustafson et al. "High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation." Genome research (2024).
14. Schloissnig et al. "Structural variation in 1,019 diverse humans based on long-read sequencing." Nature (2025).
15. Beyter, et al. "Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits." Nature genetics (2021).
16. https://github.com/epi2me-labs/wf-human-variation
17. https://github.com/chhylp123/hifiasm
18. https://github.com/lh3/seqtk
19. https://github.com/marbl/merqury
20. https://github.com/nloyfer/wgbs_tools
21. https://github.com/nloyfer/UXM_deconv
22. https://github.com/fritzsedlazeck/sniffles
23. github.com/ACEnglish/truvari
24. github.com/ACEnglish/kanpig
25. https://github.com/HKU-BAL/clair3
26. https://github.com/dnanexus-rnd/GLnexus
27. https://github.com/fakedrtom/SVAFotate

Consistent High Quality across Thousands of Sequenced Genomes

An important aspect in large-scale studies is the ability to deliver consistent data quality over multiple batches sequenced over long periods of time. We observe highly consistent sequencing quality, achieved over more than one year of sequencing, with key metrics meeting or exceeding expectations (Fig. 1).
Coverage targets of 15x and 30x have been met, with mean coverages of 16.8x and 31.3x, respectively. Read length N50 and N10 values average approximately 17 kb and 30 kb, respectively, with mean per-read base QV ≥20 (Fig. 1a). In the 30x coverage samples, 85% of the genome is covered by at least one read longer than 30 kb. Coverage remains robust across traditionally difficult regions of the genome, including those challenging for short-read sequencing (Fig. 1d). Physical phasing in high-coverage samples encompasses approximately 90% of SNVs and 80% of CpGs, with 75% of all genes and 65% of coding genes fully phased and an average phase block N50 of 350 kb (Fig. 1b+c).

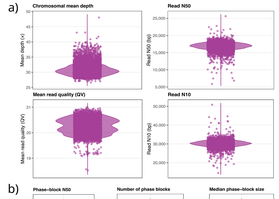
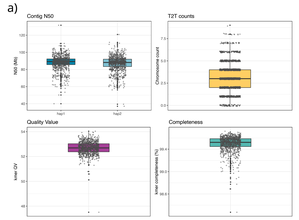




Figure 1: Sequencing QC across the first ~5,000 high coverage samples. a) Key sequencing data QC metrics: average mapped read coverage, read length N50s, average read quality and read length N10 (min read length covering >=10% of sequenced bases), b) Phase block length statistics based on Clair3 variants and subsequent whatshap phasing, c) Percentage of phased SNVs, CpGs, and full length genes (protein coding genes shown separately), c) Heatmap of breadth of coverage in different region subsets at different coverage thresholds (median across samples, mean across regions)

High Quality Assemblies for 970 Individual Human Genomes



A benefit of long reads is the ability to assemble highly contiguous whole human genomes. We generated assemblies with hifiasm (v0.25) using only the ONT data on the longitudinal samples (combined >60x coverage). In the resulting dual assemblies we observe high contiguity with contig N50 around 87Mb and a median of three chromosomes assembled telomere-to-telomere (Fig. 2a). Kmer analysis based on merqury further shows high completeness with coverage of 99.5% of the genome and high base accuracy with QV>52.

Apart from allowing for pangenome-based analyses, haplotype-resolved assemblies enable direct resolution of structurally complex genes, such as *LPA*. The *LPA* gene encodes apolipoprotein(a), a key component of lipoprotein(a) [Lp(a)], whose elevated circulating levels are an established causal risk factor for cardiovascular disease. From the assemblies, we are able to identify allele-specific KIV-2 repeat sequences and investigate repeat structure as well as smaller sequence variants (Fig. 2b). For example, we call known splice modifier variants 4733G>A (observing 22.7% carrier frequency) and 4925G>A (19.1%). We replicate the known inverse correlation of repeat copy number with measured Lp(a) levels as well as Lp(a) lowering effect of splice modifier variants, particularly when both 4733G>A and 4925G>A are present in the same individual.

Figure 2: Assembly quality and LPA repeat structure resolution. a) Key assembly QC metrics. b) Schematic example of allele-specific sequence-resolved *LPA* KIV-2 units. c) Total KIV-2 repeat copy number vs. measured Lp(a) conditioning on the presence of splice modifier variants: both = presence of both 4733G>A and 4925G>A detected; one = either 4733G>A or 4925G>A allele detected; wt = neither variant detected (wildtype).

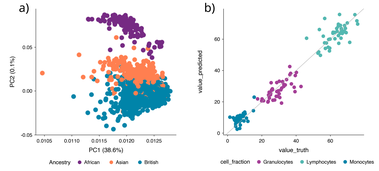
Comprehensive Genome-Wide Methylation Profiling

Oxford Nanopore sequencing preserves native DNA modifications and enables direct CpG methylation inference alongside the primary genomic sequence, without the need for additional steps in the wet lab workflow such as enzymatic conversion. It allows for methylation to be linked to nearby variants, which can reveal regional or haplotype-specific patterns, as well as permits unbiased genome-wide study of epigenetics, such as epigenome-wide association studies (eWAS) or detection of allele-specific methylation QTLs.

c)

Model	# CpGs	medAE 15x	medAE 30x
Horvath	353	9.33	8.19
SkinHorvath	391	5.28	4.08
BLUP	319,604	4.87	3.24
EN	514	4.30	3.27
BLUP_ONT	319,604	**3.23**	**2.30**
deCODE	1,373	3.33	2.53

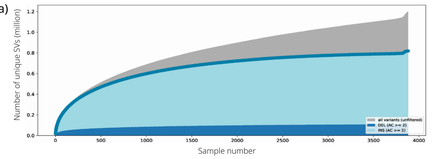
Figure 3: Methylation metrics and predictor performance. a) PCA of methylation matrix colored by genetic ancestry, b) Measured cell type fractions vs computationally deconvoluted cell types, c) Table: Methylation-based age prediction performance across various models (median absolute error vs recorded age [years])



Using subcohorts containing both high and low coverage samples we observe that a PCA based on methylation (Fig. 3a) reveals ancestry-specific patterns in PC2 (PC1 is dominated by cell types). Deriving cell type composition computationally with UXM_deconv, tested on a small subcohort of 30x samples, tracks well with measured cell type fractions (Fig. 3b). Further, methylation clocks can accurately predict age at both 30x and 15x coverage (Fig. 3c).

Structural Variant Discovery across 3,881 Long Read Genomes

Long-read sequencing enables more complete resolution of structural variation and has uncovered both previously uncharacterized functional variation and candidate causal SVs underlying established GWAS loci.
We joint genotyped SVs in the first almost 4k participants sequenced at 30x using the STAK pipeline (Sniffles2 + truvari collapse + kanpig). This resulted in a total of 1.2m SVs across the autosomes, a large proportion of which are insertion alleles (Fig. 4a). PCA captures the known population structure of this cohort (Fig. 4b).
In total, 867k SVs are non-singletons seen in at least two individuals (84k outside tandem repeat regions). On average, we detect about 21k SVs per participant, with 20.9k SVs per British and 24.1k per African individual. Especially African ancestry individuals (rightmost samples in Fig. 4a) contribute higher number of novel SVs. Out of the 867k singletons, 442k (28k outside TRs) are not contained in the UKB short read SV call set based on about half a million participants, indicating the immense potential to make novel discoveries from SVs in this cohort.



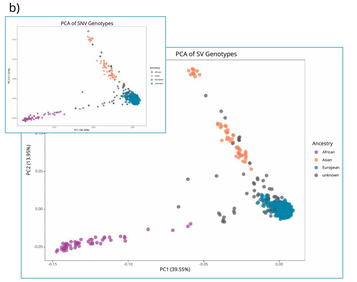


Figure 4: Cohort-level Structural Variant Calling. a) Total number SVs per sample added, broken down by singletons (grey) and SVs seen at least twice (AC>=2) further broken down into DEL and INS. b) PCA based on SV genotypes mirrors that of SNV genotypes (Clair3 + GLnexus) reflecting known continental ancestries in the UK Biobank.

Conclusion

The UK Biobank long-read WGS cohort will expand catalogues of genetic variation beyond what short-read sequencing can resolve. Access through the UKB-RAP, alongside clinical and phenotypic data, will enable discovery of novel associations and deeper insight into functional mechanisms, particularly through the combination of genetics with epigenetics and through sequence-resolved, allele-specific investigation of complex loci such as *LPA*. Initial QC and exploratory analyses highlight the consistent high quality of the dataset and its potential for novel discoveries in both genetics and epigenetics. Its scale and complexity further offer opportunities for continued method development, including pangenome-based approaches that better incorporate genomic diversity. Extending population-scale long-read whole genome sequencing to include more global genetic diversity will be essential to broaden variant discovery, advance biological understanding, and support more equitable genomics research.

Oxford Nanopore Technologies, the Wheel icon, MinION are registered trademark of Oxford Nanopore Technologies plc in various countries. All other brands and names contained are the property of their respective owners. © 2026 Oxford Nanopore Technologies plc. All rights reserved.
Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition