# Evaluation of Oxford Nanopore whole genome methylation sequencing for phenotype inference and association testing

Sean McKenzie[1], Alex Dawson[2], Heather Jeffery[2], Anne-Katrine Emde[1], Rachel Martin[2], Phill James[2], Philipp Rescheneder[2], Sissel Juul[1]

[1]Oxford Nanopore Technologies, Inc. New York, NY
[2]Oxford Nanopore Technologies, plc. Oxford, UK

## Abstract

DNA methylation at CpG sites is an important biomarker of biological state, including age,[1,2] sex,[3] cell composition,[4] and diverse health-relevant phenotypes.[5] Because whole genome short read methods like whole genome bisulfite sequencing (WGBS) or enzymatic methyl-seq (EM-seq) remain relatively costly and workflow-intensive, large human methylation association studies have relied heavily on arrays and reduced representation bisulfite sequencing (RRBS)[6–8]. However, EPIC arrays target fewer than 5% of human CpGs,[7] while RRBS typically recovers between 5 and 15% of CpGs,[8] limiting discovery to a minority of CpG sites. Oxford Nanopore sequencing preserves native DNA modifications and enables direct CpG methylation inference alongside primary sequence,[8,16] allowing methylation to be linked with nearby variants, regional co-methylation patterns, and haplotype-informative SNPs[11,12]. A lack of chemical or enzymatic conversion simplifies wet lab workflows and increases the final usable fraction of sequencing data relative to standard short read methods[13]. These features make nanopore sequencing a compelling candidate for whole-genome methylation analysis and phenotype inference, motivating evaluation across common use cases.

Here, we evaluate Oxford Nanopore whole genome methylation sequencing (WGMS) for phenotype inference and association testing using 22 healthy human whole-blood samples with matched EPIC array and WGMS data. We examine three common applications: methylation-based chronological age prediction ("epigenetic clocks"), whole-blood cell-type deconvolution, and a proof-of-principle epigenome-wide association study (EWAS).

## Experimental design

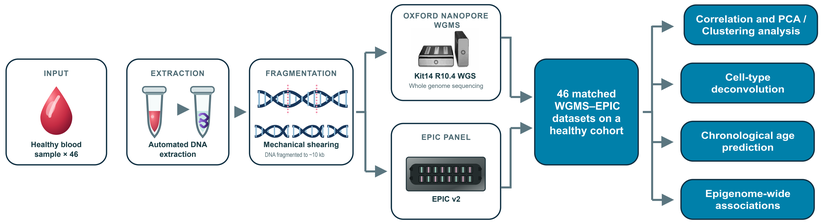



*Figure 1*: Flowchart showing the wetlab and drylab experiments undertaken to compare Oxford Nanopore WGMS and EPIC array methylation.

Blood samples were obtained from 46 healthy donors through Cambridge Bioscience, across a range of ages (18-65). DNA was extracted from the donor blood by chemagic BBS (blood, buffy coat and saliva) DNA kit H96 with DNA extractions occurring on the day of blood delivery. Shearing to ~10 kb was performed with the genogrinder (speed 1750 spm at ~30ng/ul, total per well volume of 120 ul). Whole genome sequencing with LSK114 was performed in-house with one sample per flow cell, run on a PromethION24 for 72 hours with no washes. Live basecalling was performed during sequencing, using the v5.2.0 hac basecalling model with 5mC and 5hmC detection in CG contexts and aligning to the hg38 human reference. In parallel, samples were sent to an external vendor to assay with the EPIC v2 array, with QC and initial data processing provided by the vendor. For all analyses, WGMS 5mC and 5hmC values were combined to ensure comparability to arrays, and data were downsampled to 30X coverage unless otherwise indicated. Methylation values for each sample on each platform were plotted and compared to assess large-scale concordance, and then data from both platforms was analyzed according to platform-specific best practices for cell-type deconvolution, age/trait prediction (using pre-trained models as well as *ab initio* models with leave-one-out cross validation), and Epigenome-wide association studies (EWAS).

## Concordance in methylation frequencies and multidimensional clustering

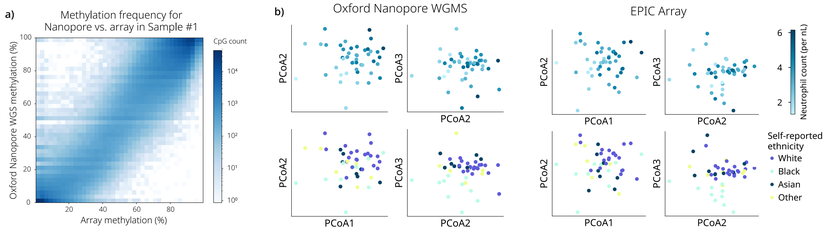



*Figure 2*: a) Correlation between Oxford Nanopore WGS-derived methylation proportions and EPIC array-derived beta values. b) Principal coordinate projection of samples from WGS and EPIC values.

To investigate the high-level patterns in the data and comparability of Oxford Nanopore and EPIC array derived methylation values, we calculated the correlation of methylation proportions derived from the Oxford Nanopore WGMS data (down-sampled to 30X coverage) and the beta values derived from the EPIC arrays at sites common to both technologies. Across 46 matched samples, ONT and EPIC methylation estimates were highly correlated (Figure 2a); Pearson's R range between 0.97-0.98, with a mean of 0.977, while the mean absolute divergence was 7.6% across all shared sites across all samples. Principal coordinate analysis based on all sites (not just common sites) showed extremely concordant high-level patterns in both data types, with a first principal coordinate axis dominated by cell type composition (predominantly neutrophil proportion) and the next two highly correlated with self-reported ethnicity (Figure 2b).

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

## Chronological age prediction

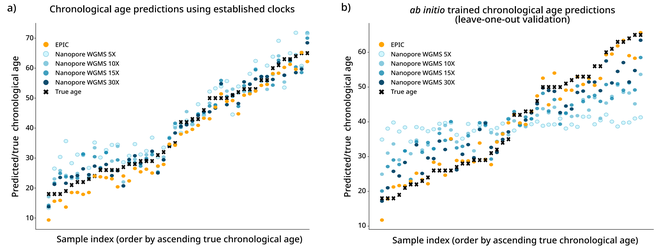



*Figure 3*: Methylation-derived age estimates using a) existing models trained on large methylation cohorts (validated jointly using all samples) and b) *ab initio* models trained using the inference data type (validated with leave-one-out cross validation)

One important use of DNA methylation data is predicting traits in a sample, either using existing models or by training models *ab initio*. Chronological age models (so called "epigenetic clocks") are particularly common given the robust signal of aging in DNA methylation data and the general interest in the epigenetic basis of aging. We applied a range of models from Zhang et al.[14] to Oxford Nanopore WGMS data downsampled to various depths as well as to our matched EPIC array data (Figure 3a; Table 1). Following intercept calibration, we observed comparable performance for the best models for both 30X WGMS and EPIC array data, although different models were found to be optimal for each technology (the BLUP model for WGMS, the Elastic Net model for EPIC array). Downsampled WGMS data showed excellent performance down to 5X coverage. We also trained models ab initio with all the data types and assessed by leave-one-out cross-validation, given the small sample size (Figure 3b; Table 1). We were able to train robust models for all datasets except for the 5X coverage WGMS dataset, although performance dropped off much faster with decreasing depth than in our inference-only test, highlighting the limitations of site selection and model fitting with such a high feature-to-sample ratio.

| Dataset | Best pre-trained clock | | Best *ab initio* trained clock | | | |
|---|---|---|---|---|---|---|
| | MAE | R | Method | CpGs | MAE | R |
| **5X WGMS** | 5.26 | 0.944 | ridge | 2,000 | 13.1 | 0.599 |
| **10X WGMS** | 3.18 | 0.969 | ridge | 200 | 8.86 | 0.920 |
| **15X WGMS** | 3.03 | 0.968 | elastic net | 200 | 7.23 | 0.915 |
| **30X WGMS** | 2.57 | 0.979 | ridge | 500 | 4.54 | 0.954 |
| **EPIC** | 2.42 | 0.999 | elastic net | 2,000 | **2.42** | 0.980 |

*Table 1*: Mean average error (MAE) and Pearson's R for methylation-derived age estimates using existing pre-trained models and *ab initio* models, as well as model information for the best performing *ab initio* models.

## Cell-type deconvolution

| Cell type | WGMS R | EPIC R |
|---|---|---|
| **Lymphocytes** | 0.983 | 0.977 |
| **Monocytes** | 0.795 | 0.901 |
| **Neutrophils** | 0.972 | 0.974 |

*Table 2*: Pearson's R for predicted vs. CBC-derived cell type proportions using Oxford Nanopore WGMS or EPIC array data.

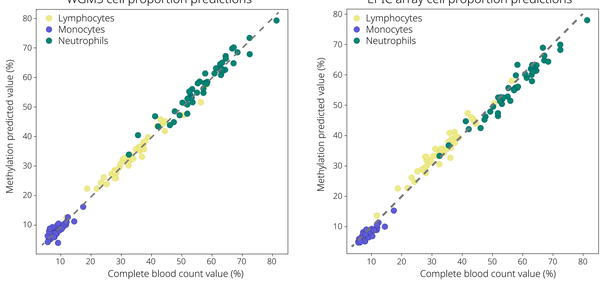
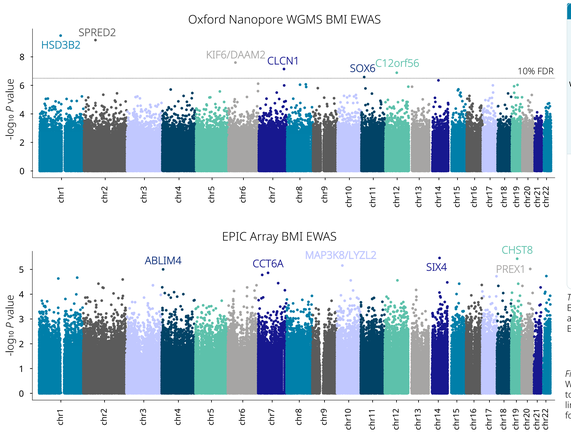



*Figure 4*: Blood cell proportion estimates from a) WGMS and b) EPIC array methylation values vs CBC-derived values. Points are colored by cell type, and a dotted line indicates 1:1 parity.

Peripheral blood samples contain DNA extracted from a variety of cells from various blood cell lineages, with each lineage containing a distinct methylation profile regulating cell-specific gene expression. Per sample cell-type proportion estimation ("cell-type deconvolution") is routinely used to preserve real biological signal during downstream analyses. To compare the performance of Oxford Nanopore WGMS data to traditional EPIC array methods, we followed best practices for cell type deconvolution for each datatype and compared proportion estimates to those derived from a complete blood count (CBC) from sample provider. Methylation deconvolution and CBC resolve different sets of cell types, so estimates were merged to most granular shared categories: lymphocytes, monocytes, and neutrophils. Both methylation technologies provided estimates with high concordance to CBC values (Figure 4; Table 2). WGMS gave more accurate estimates for Lymphocytes (Pearson's R 0.983 vs. 0.977), while EPIC arrays gave more accurate estimates for Monocytes (0.901 vs. 0.795).

## Epigenome-wide association study




| Tech | Site ID | P value | Gene | Most relevant known phenotype |
|---|---|---|---|---|
| WGMS | chr1:119422219-119423430 | 3.25E-10 | HSD3B2 | Clinical congenital adrenal hyperplasia |
| | chr2:65478559-65478565 | 6.69E-10 | SPRED2 | EWAS BMI |
| | chr6:39754594-39755547 | 2.52E-08 | KIF6/DAAM2 | GWAS BMI |
| | chr7:143351147-143351296 | 7.17E-08 | CLCN1 | GWAS myotonic syndrome |
| | chr12:64385927-64386419 | 1.30E-07 | C12orf56 | GWAS body height |
| | chr11:15650617-15650770 | 2.61E-07 | SOX6 | GWAS BMI |
| | chr14:50892497-50893193 | 4.43E-07 | ABHD12B | GWAS peripheral vascular disease |
| | chr6:169078571-169079241 | 7.63E-07 | THBS2 | GWAS BMI |
| | chr8:106081377-106082265 | 8.50E-07 | ZFPM2 | GWAS BMI |
| | chr8:73071815-73073753 | 8.98E-07 | SBSPON | GWAS pulse pressure |
| EPIC | cg14985973_TC21 | 0.000003 | SIX4 | GWAS body height |
| | cg07777720_BC21 | 0.000004 | CHST8 | EWAS BMI |
| | cg14864161_BC21 | 0.000007 | MAP3K8/LYZL2 | GWAS body height |
| | cg19056467_BC21 | 0.00001 | PREX1 | EWAS BMI |
| | cg20536272_TC21 | 0.00001 | ABLIM2 | GWAS alcohol drinking |
| | cg19098690_TC21 | 0.000014 | CCT6A | GWAS body height |
| | cg02391394_BC21 | 0.000017 | IL6 | GWAS pulse pressure |
| | cg00733379_BC21 | 0.000019 | TTC28 | GWAS lean body mass |
| | cg14196211_TC11 | 0.000019 | MRPL12 | Clinar combined oxidative phosphorylation deficiency 45 |
| | cg19297398_BC21 | 0.000021 | GPATCH2 | GWAS BMI |

*Table 3*: Top ten BMI EWAS hits for Oxford Nanopore WGMS and EPIC array data, annotated with the closest protein coding gene and the most relevant known phenotype for that gene from the EWAS Atlas and OpenTargets databases.

*Figure 5*: Manhattan plots for BMI EWAS hits in Oxford Nanopore WGMS and EPIC array data, with the closest gene labeled for the top six hits in each technology. A line representing the FDR=10% line is shown for the Nanopore plot (no hits met this threshold for the EPIC array analysis)

Epigenome-wide association studies (EWAS) can reveal how DNA methylation regulates and responds to physiological change and environmental exposure, and can help identify or refine drug targets. For example, EWAS of phenotypes associated with metabolic disorder, including BMI, may be helpful for designing lifestyle and pharmeceutical interventions for this disorder and its comorbidities. We conducted EWAS of BMI using both WGMS and EPIC array data, controling for age, sex, cell type composition, and self-reported ethnicity. In order to increase methylation frequency estimation precision and improve detection power, we identified correlated tracts of neighboring CpGs in the WGMS data and aggregated methylation counts within these tracts for our WGMS EWAS. After FDR correction, we identified six significant associations in the WGMS data and none in the EPIC array data (Figure 5). However, because the disparate data types necessitated very different statistical models, direct comparison of *P* values is likely inappropriate. We therefore looked at the top ten hits for each platform and assessed biological credibility of the hits by comparison to the EWAS Atlas and OpenTargets database (Table 3). We found that one WGMS hit matched a previous BMI EWAS hit, while four matched BMI GWAS hits. Meanwhile, two EPIC array hits matched previous BMI EWAS hits and one more matched a BMI GWAS hit.

## Conclusion

Oxford Nanopore WGMS produced methylation estimates that were highly concordant with EPIC v2 arrays at shared CpGs (mean Pearson's R = 0.977), and both platforms captured the same dominant population-level structure, driven by cell composition and ancestry. Across the three common applications, nanopore data performed comparably to arrays. Pre-trained epigenetic clocks achieved similar accuracy after intercept calibration and remained robust down to 5X coverage. *Ab initio* models were more sensitive to sequencing depth, reflecting the challenge of feature selection when features vastly outnumber samples. Methylation-based cell-type deconvolution from both platforms agreed closely with complete blood counts, with each platform modestly outperforming the other for different lineages. In a proof-of-principle BMI EWAS, aggregating co-methylated CpG tracts enabled detection of FDR-significant associations in the nanopore data. Top hits from both platforms overlapped previously reported BMI EWAS and GWAS loci, supporting their biological credibility. Together, these results suggest that nanopore WGMS can serve as a direct replacement for array-based methylation analyses while extending coverage to the large majority of CpGs that arrays miss, and it adds sequence and haplotype context. Larger cohorts will be needed to fully realize the discovery potential of genome-wide, long-read methylation data.

## References

1. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
2. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49:359–367.
3. McCarthy NS, et al. Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. *BMC Genomics*. 2014;15:981.
4. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13:86.
5. Wahl S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*. 2017;541:81–86.
6. Gu H, et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc*. 2011;6:468–481.
7. Pidsley R, et al. Critical evaluation of the Illumina Methylation Epic BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*. 2016;17:208.
8. Simpson JT, et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14:407–410.
9. Rand AC, et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods*. 2017;14:411–413.
10. Liu Y, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol*. 2021;22:295.
11. Gigante S, et al. Using long-read sequencing to detect imprinted DNA methylation. *Nucleic Acids Res*. 2019;47:e46.
12. Akbari V, et al. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. *Genome Biol*. 2021;22:68.
13. Oliva N, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol*. 2018;19:33.
14. Zhang Q, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological aging. *Genome Med*. 2019;11:54.